ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

PROGRAMACIÓN LINEAL MULTICRITERIO Y MINIMIZACIÓN DEL ERROR EN PROBLEMAS DE CLASIFICACIÓN: APLICACIÓN EN LA CALIFICACIÓN CREDITICIA

PROYECTO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

EDWIN AUGUSTO PUCUJI JÁCOME
winedp@hotmail.com

DIRECTORA: Dra. SANDRA ELIZABETH GUTIÉRREZ POMBOSA
sandra.gutierrez@epn.edu.ec

Quito, DICIEMBRE 2015

DECLARACIÓN

Yo, Edwin Augusto Pucuji Jácome, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

_____

Edwin Augusto Pucuji Jácome

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Edwin Augusto Pucuji Jácome, bajo mi supervisión.

_____

Dra. Sandra Elizabeth Gutiérrez Pombosa

DIRECTORA

DEDICATORIA

A mis padres, Augusto y Leticia, por su amor, trabajo, sacrificios y por su comprensión, son los mejores padres.

AGRADECIMIENTOS

Primero que nada, quiero agradecerte a ti Dios por la vida, por protegerme en todo momento y por haberme dado la oportunidad para llegar tan lejos, por brindarme el privilegio de conocer personas maravillosas de las que aprendí. Gracias además por poder contar con mis seres queridos y con su cariño.

A mis padres, por su comprensión, motivación y apoyo que me han brindado para alcanzar todas y cada una de mis metas así como me impulsan a alcanzar mis sueños. Por haberme brindado la oportunidad de estudiar esta carrera, por su esfuerzo, dedicación y entera confianza.

A mi papá, gracias por tu apoyo, tu seriedad, firmeza y humildad, han sido la orientación y luz en mi camino, pauta para poder realizarme en mis estudios y en mi vida. Gracias por enseñarme a cultivar la tierra porque así me has enseñado a cultivar mis metas.

A mi mamá, por enseñarme a confiar en mí mismo, a tener fe en Dios, por apoyarme y enseñarme a luchar por mis sueños. Gracias por que siempre me has levantado los ánimos tanto en los momentos difíciles de mi vida estudiantil como personal. Gracias por la paciencia y las palabras sabias que siempre tienes para mis enojos, tristezas y momentos felices.

Agradezco a la Dra. Sandra Gutiérrez Pombosa por su confianza, ayuda e interés al brindarme la oportunidad de desarrollar este proyecto, por las sugerencias recibidas, el seguimiento y la supervisión continúa del mismo. Así mismo agradezco a todos los profesores de la Facultad de Ciencias por enseñarme un mundo nuevo de conocimientos.

Un reconocimiento especial a la Srta. Msc. Patricia Álvarez Yaulema, por haberse interesado en mi trabajo y el apoyo recibido para la realización de este Proyecto. También me gustaría expresar mi más profundo agradecimiento a todas aquellas personas que con su ayuda han colaborado en la realización del presente trabajo, en especial a mis compañeros de trabajo por apoyarme en todo momento.

Un agradecimiento muy especial a Mayra, mi hermana, por apoyarme siempre, por todos sus consejos y por darme a mi sobrino Mikael, la alegría de la casa; a mi cuñado Marco por sus consejos y apoyo incondicional; a toda mi familia, en especial a mi tía Virginia por su ayuda incondicional.

A Cristina, por ser una mujer muy especial en mi vida, en realidad he disfrutado momentos felices durante el desarrollo de este proyecto, gracias por tus consejos en todo momento, por los regaños que bien merecidos me das cuando me equivoco y por compartir conmigo esos momentos de tristezas y alegrías.

Índice general

Índice de figuras

Índice de tablas

RESUMEN

Los modelos de medición de riesgo se han convertido en herramientas fundamentales en la administración de las instituciones financieras. En el presente trabajo se estudia en detalle varias metodologías para tratar el problema de clasificación y minimizar su error de clasificación, además para la práctica se enfoca en el problema de predecir el riesgo que los solicitantes de crédito representan para una institución microfinanciera. El documento presenta las principales características de los modelos de clasificación basados en la rama de la Inteligencia Artificial y la Investigación de Operaciones donde se han desarrollado modelos como las máquinas de soporte vectorial y la programación lineal multicriterio y de restricción múltiple, respectivamente, cuya fundamentación parte del mismo análisis discriminante, método tradicional que permite identificar las características que diferencian a dos o más grupos, y crear un discriminador capaz de distinguir con la mayor precisión posible a los miembros de uno u otro grupo. Las Máquinas de Soporte Vectorial constituyen estructuras de aprendizaje automático que han demostrado un excelente papel en aplicaciones de clasificación, se fundamentan en transformar el espacio de entrada en otra dimensión superior en el que el problema puede ser resuelto mediante un hiperplano óptimo por medio de una función núcleo. Uno de los objetivos principales de la modelización matemática es minimizar el error entre las simulaciones y sus respectivos fenómenos de la vida real, persiguiendo este ideal se busca estudiar el error que se comete en las metodologías de clasificación para ello se proponen algoritmos basados en programación lineal multicriterio y de restricción múltiple, los mismos que para enfrentar tal problema persiguen dos objetivos simultáneamente, el primer objetivo es el de maximizar las distancias mínimas entre las observaciones y el hiperplano crítico, y el segundo objetivo es el de separar las observaciones minimizando la suma de las desviaciones entre las observaciones. La información utilizada para el entrenamiento y validación de las metodologías propuestas, está constituida por un conjunto de variables socio-demográficas y del buró de crédito correspondientes a 16.860 clientes del portafolio de microcrédito de una institución microfinanciera. Una vez seleccionadas las variables significativas a través de la técnica de árboles de decisión se obtienen las reglas de clasificación y son utilizadas en el entrenamiento y la validación de los modelos de clasificación. Como resultado se observa una eficiencia del 87.88 % clasificados correctamente.

Palabras claves: Problema de Clasificación, Análisis discriminante Lineal, Máquinas de Soporte Vectorial, Programación Lineal Multicriterio, Comportamiento de los usuarios de Tarjetas de Crédito.

ABSTRACT

The risk measurement models have become essential tools in the management of financial institutions. In this work is studied in detail various methodologies to address the problem of classification and minimize misclassification, in addition to practice it focuses on the problem of predicting the risk that applicants for credit represent a microfinance institution. The paper presents the main features of classification models based on the branch of Artificial Intelligence and Operations Research where models have been developed such as support vector machines and the multi-criteria linear programming and multiple restriction, respectively, whose foundation part discriminant analysis of the same traditional method to identify the characteristics that differentiate two or more groups, and create a discriminator able to distinguish with the greatest precision possible for members of one group or another. Support Vector Machines constitute machine learning structures that have demonstrated an excellent role in sorting applications, are based on transforming the input space into another higher dimension in which the problem can be solved by an optimal hyperplane by means of a function core. One of the main objectives of mathematical modeling is to minimize the error between the simulations and their phenomena of real life, pursuing this ideal is to study the error made in the methodologies of classification for that algorithms based on linear programming is proposed multicriteria and multiple restriction, the same as to face such a problem two objectives simultaneously, the first objective is to maximize the minimum distance between observations and critical hyperplane, and the second objective is to separate the observations by minimizing the sum of the deviations between observations. The information used for training and validation of the proposed methodologies, consists of a set of socio-demographic variables and the corresponding credit bureau to 16.860 customers microcredit portfolio of a microfinance institution. After selecting the significant variables through the technique of decision trees classification rules they are obtained and used in training and validation of the classification models. As a result an efficiency of 87.88 % correctly classified is observed.

Keywords: Classification Problem, Linear Discriminant Analysis, Support Vector Machines, Multicriteria Linear Programming, Credit Cardholders'Behavior.

Capítulo 1

INTRODUCCIÓN

1.1 Antecedentes

La característica fundamental del sistema financiero es su alto grado de regulación, la regulación bancaria tiene como finalidad la búsqueda del buen funcionamiento del sistema, la misma que se ha interesado por la solvencia de las entidades financieras, desarrollando un número importante de normativas que tratan de salvaguardar este objetivo.

La aprobación de créditos es la actividad principal del sistema bancario, es aquella que mejor la define y a la que designa la mayor parte de sus esfuerzos, la que origina la mayor parte de su rentabilidad y los mayores riesgos. En este ámbito el riesgo se define como la posibilidad de que se produzca un hecho generador de pérdidas que afecten el valor económico de las entidades bancarias, por lo tanto, la creencia basada en no asumir aquellas operaciones que no ofrecen garantías adecuadas. Por otra parte el negocio bancario supone, la gestión de riesgos con el propósito de obtener una rentabilidad. Una entidad financiera es una máquina de gestión de riesgos en busca de rentabilidad. De todos los riesgos a los que está expuesta una entidad financiera, el riesgo de crédito es el principal.

El Riesgo de crédito es la posibilidad de pérdida debido al incumplimiento del prestatario o la contra parte en operaciones directas, indirectas o de derivados que conllevan el no pago, el pago parcial o la falta de oportunidad en el pago de las obligaciones pactadas. La administración de riesgos es el proceso mediante el cual las instituciones del sistema financiero identifican, miden, controlan/mitigan y monitorean los riesgos inherentes al negocio, con el objeto de definir el perfil de riesgo, el grado de exposición que la institución está dispuesta a asumir en el desarrollo del negocio y los mecanismos de cobertura, para proteger los recursos propios y de terceros que se encuentran bajo su control y administración.

1.2 Definición del problema

La capacidad de almacenamiento de información digital en el campo financiero se ha duplicado, esto ha provocado la aparición de las denominadas *fosas de datos*: datos que son almacenados y descansan en paz, sin que nadie los reclame o los recuerde [Carrizosa y Martín, 2005], lo cual ha desarrollado una disciplina conocida como *Minería de Datos*, cuyo propósito es explorar grandes volúmenes de datos para extraer información potencialmente útil. El desarrollo tecnológico-científico en los últimos años permite el uso de herramientas matemáticas de diversos campos, como la Investigación de Operaciones, la Estadística o la Inteligencia Artificial, que son aplicadas en múltiples sectores, donde se generan bases de datos de gran tamaño, en ocasiones muy desestructuradas y con ruido (valores perdidos, valores erróneos o valores atípicos) [Apte, 2003],[Hand et al., 2001],[Hastie, 2001].

Una de las tareas más importantes en el campo de la Minería de Datos es la Clasificación, la misma que consiste en encontrar un conjunto de modelos que describen y distinguen los grupos (clases) de datos. Los algoritmos de clasificación utilizan los datos de una muestra de entrenamiento (datos pre-clasificados) para inferir o construir funciones o modelos que asignen futuras observaciones en una de las clases predefinidas. La función de decisión se utiliza entonces para predecir la clase a la que corresponderán observaciones futuras.

Para evaluar los expedientes de los solicitantes de crédito, en los diferentes segmentos como por ejemplo: créditos de consumo, comercial, vivienda y en especial los microcréditos se han usado distintos métodos de clasificación, analizando el historial crediticio de los clientes para estimar el comportamiento de nuevos clientes y otorgar una calificación crediticia o probabilidad de incumplimiento, acorde con las características del solicitante de crédito. De manera similar a lo que sucede en el ámbito financiero, sucede también en el campo de la medicina y el industrial, donde se han aplicado modelos de clasificación para determinar si un paciente está enfermo o sano, o para determinar si los productos son defectuosos o no, respectivamente.

Herramientas de las ramas de la Estadística y la Inteligencia Artificial se han desarrollado para tratar los problemas de clasificación, métodos como los modelos de regresión logística, técnicas discriminantes, árboles de decisión y redes bayesianas se han empleado exitosamente. Sin embargo la optimizació matemática, de manera especial el campo de la Investigación de Operaciones mediante la utilización de la Programación Lineal Multicriterio también hoy en día contribuye exitosamente a tratar los problemas de Clasificación, con un aporte muy importante que permite la corrección del error que se

incurre al ajustar problemas de clasificación, un error de clasificación se produce cuando se asigna una observación perteneciente a la clase y_1 a la clase y_2 o viceversa.

De lo anteriormente expuesto: el problema a estudiar en este proyecto es, analizar las técnicas tradicionales y modernas para ajustar problemas de clasificación y como éstas ayudan a minimizar el número de errores de clasificación. Para varias aplicaciones, el problema es mucho más complejo que simplemente minimizar el número de errores de clasificación, como por ejemplo, en un problema de diagnóstico médico: Observar que, si un paciente que no tiene cáncer se diagnostica incorrectamente como tiene cáncer (Error Tipo I), las consecuencias pueden ser un poco de malestar del paciente además de la necesidad de realizar más investigaciones. A la inversa, si un paciente con cáncer se diagnostica lo más saludable (Error Tipo II), el resultado puede ser la muerte prematura debido a la falta de tratamiento. Así, las consecuencias de estos dos tipos de error pueden ser dramáticamente diferentes. En el ejemplo, claramente sería mejor que se considere hacer menos errores de la segunda clase, incluso si esto fuese a costa de cometer más errores de la primera clase.

Se puede formalizar estas cuestiones a través de la introducción de una *función de pérdida (loss)*, también llamada *función de costos (costs)*, que es una medida única, global de pérdida incurrida en tomar cualquiera de las decisiones o acciones disponibles. Uno de los conceptos fundamentales es la noción del error o ***loss*** para medir el acuerdo entre la predicción **f(x)** y la salida ***y*** deseada. Una función de pérdida: **L** : $\mathbb{Y}\times\mathbb{Y} \rightarrow \mathbb{R}^{+}$ es introducida para evaluar el error. La elección de la función de pérdida **L(f(x),y)** depende del problema de clasificación que se resuelva.

El problema que se plantea entonces es minimizar la pérdida total incurrida. Se supone que, para un nuevo valor de ***x***, la clase verdadera es C_k y que se asigna ***x*** a la clase C_j (donde j puede o no puede ser igual a k). Entonces se incurre en algún nivel de pérdida que se denota por L_{kj}, es decir, el elemento kj de una matriz de pérdida. En el ejemplo mencionado anteriormente, la matriz de pérdida en particular dice que si se toma la decisión correcta, no hay pérdida incurrida; por el contrario existe una pérdida mínima, si un paciente sano se diagnostica como enfermo de cáncer, mientras que si se diagnostica a un paciente que tiene cáncer como lo más saludable, existe una pérdida mucho mayor.

La solución óptima es la que minimiza la función de pérdida. Sin embargo, esta función depende de la distribución de probabilidad de la muestra **P(x,y)**, que es desconocida, surge entonces la necesidad de la investigación de modelos matemáticos que minimicen la función de perdida de la mejor manera posible, es por esta razón que en el presente

proyecto se estudian los principales modelos desarrollados para tratar el problema de la corrección del error en problemas de clasificación, tomando ventaja de los algoritmos de clasificación basados en optimización, por ejemplo: la programación lineal (LP), ésta es una herramienta útil para discriminar el análisis de un problema dado grupos adecuados (por ejemplo, “bueno” y “malo”) [Freed and Glover, 1981]. Adicionalmente la programación lineal multicriterio (MCLP) ha mejorado la clasificación de las clases a través de la corrección del error, minimizando la suma de las desviaciones entre las observaciones y maximizando las distancias mínimas de las observaciones al valor crítico, simultáneamente [Shi, 2010],[Kou et al., 2003],[Shi et al., 2002].

Además se implementa un modelo de programación lineal multicriterio y de restricción múltiple (MC2LP) basado en efectos de la corrección de errores. En éste modelo, se suman dos hiperplanos más para detectar cuidadosamente las observaciones mal clasificadas. En consecuencia, una discusión sutil es involucrarse con respecto a la relación entre los dos tipos de errores y las desviaciones. De hecho, en la estadística, la reducción del error del tipo I hará que sea cada vez mayor el error del tipo II, y viceversa. Por lo tanto, es importante centrarse y tratar minuciosamente los distintos tipos de error, [Wang and Shi, 2013], [Wang and Shi, 2012].

La minería de datos para la toma de decisiones de la gestión de la cartera de crédito clasifica el comportamiento de los diferentes titulares (clientes) del crédito en términos de su pago a las entidades de crédito, tales como bancos y firmas de préstamos hipotecarios. Las instituciones financieras tienen diferentes variables para describir el comportamiento de los titulares del crédito, algunas categorías de las variables son saldos, abonos y adelantos en efectivo. Algunas instituciones financieras pueden considerar la categoría de estado de residencia y la seguridad laboral como variables especiales.

Un modelo general multiclase utilizando programación lineal multicriterio puede ser propuesto como: *Sea* r *un conjunto de variables de comportamiento de los clientes de crédito* $a = (a_1, \cdots, a_r)$, *sea* $\mathbf{A}_i = (\mathbf{A}_{i1}, \cdots, \mathbf{A}_{ir})$ *una muestra de variables desarrolladas para la información crediticia, donde* $i = 1, \cdots, n$ *es el tamaño de la muestra. Se quiere determinar los coeficientes de las variables determinadas por* $\mathbf{B} = (x_1, \cdots, x_r)$. *Si un problema puede ser predefinido como* s *diferentes clases, entonces el límite de separación entre el* $j-$*ésimo y* $j+1-$*ésimo grupo puede ser* b_j, $j = 1, \cdots, s-1$. La separación de estas clases es:

$$\mathbf{A}_i\mathbf{X} \leq b_1, \quad \mathbf{A}_i \in \mathbf{G}_1,$$

$$b_{k-1} \leq \mathbf{A}_i\mathbf{X} \leq b_k, \quad \mathbf{A}_i \in \mathbf{G}_1 \qquad k = 2, \cdots, s-1$$

$$\mathbf{A}_i\mathbf{X} \geq b_{s-1}, \quad \mathbf{A}_i \in \mathbf{G}_s.$$

Sea α_i^j el grado de solapamiento con respecto a $\mathbf{A}_i$ en $\mathbf{G}_j$ y $\mathbf{G}_{j+1}$ y sea β_i^j la distancia desde $\mathbf{A}_i$ en $\mathbf{G}_j$ y $\mathbf{G}_{j+1}$ a sus límites ajustados. La separación de estas clases se modifica como:

$$\mathbf{A}_i\mathbf{X} = b_1 - \alpha_i^1 + \beta_i^1, \quad \mathbf{A}_i \in \mathbf{G}_i,$$

$$b_{k-1} - \alpha_i^{k-1} + \beta_i^{k-1} = \mathbf{A}_i\mathbf{X} = b_k - \alpha_i^k + \beta_i^k, \quad \mathbf{A}_i \in \mathbf{G}_k,$$

$$\mathbf{A}_i\mathbf{X} = b_{s-1} + \alpha_i^{s-1} - \beta_i^{s-1}, \quad \mathbf{A}_i \in \mathbf{G}_s,$$

El objetivo de este problema es maximizar $\sum_j \beta_i^j$ y minimizar $\sum_j \alpha_i^j$ simultáneamente:

$$\text{Minimizar} \sum_i \sum_j \alpha_i^j$$
$$\text{Maximizar} \sum_i \sum_j \beta_i^j$$

Sujeto a:

$$\begin{array}{rcccl}
\mathbf{A}_i\mathbf{X} & = & b_1 - \alpha_i^1 + \beta_i^1, & \mathbf{A}_i \in \mathbf{G}_1 & \\
\mathbf{A}_i\mathbf{X} & = & b_{k-1} + \alpha_i^{k-1} - \beta_i^{k-1}, & \mathbf{A}_i \in \mathbf{G}_k & k = 2, \cdots, s-1; \\
\mathbf{A}_i\mathbf{X} & = & b_k - \alpha_i^k + \beta_i^k, & \mathbf{A}_i \in \mathbf{G}_k & k = 2, \cdots, s-1; \\
\mathbf{A}_i\mathbf{X} & = & b_{s-1} + \alpha_i^{s-1} - \beta_i^{s-1}, & \mathbf{A}_i \in \mathbf{G}_s; & \\
 & & b_{k-1} + \alpha_i^{k-1} \leq b_k - \alpha_i^k, & & k = 2, \cdots, s-1;
\end{array}$$

donde $\mathbf{A}_i$ son dados; $\mathbf{X}$ y b_j son no restringidas; y $\alpha_i^j \geq 0$, y $\beta_i^j \geq 0$, $j = 1, \cdots, s-1$.

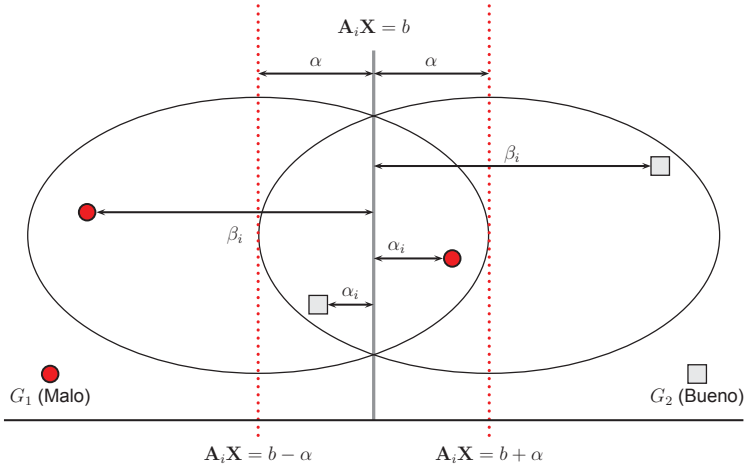
Los antecedentes del modelo anterior se basa tanto en el análisis discriminante lineal (ver figura 1.1) y varios criterios de programación lineal. En el análisis discriminante lineal, la clasificación errónea de separación de datos puede ser descrita por dos objetivos opuestos en un sistema lineal. El primero de ellos es el de maximizar las distancias mínimas (MMD) de observaciones con respecto al valor crítico.

El segundo objetivo es separar las observaciones, al minimizar la suma de las desviaciones (MSD) entre las observaciones [Freed and Glover, 1981], [Freed and Glover, 1986], [Koehler and Erenguc, 1990].

Comparando con las herramientas matemáticas tradicionales de clasificación, como el árbol de decisión, las estadísticas y las redes neuronales, éste enfoque es simple y directo, libre de los supuestos estadísticos, y flexible, al permitir que los investigadores desempeñen un papel activo en el análisis [Joachimsthaler and Stam, 1990].

Sin embargo, como el análisis discriminante lineal utiliza sólo MMD, MSD, o una combinación dada de MMD y MSD para medir los errores de clasificación, no podía encontrar la mejor solución de compromiso entre dos mediciones. Esta deficiencia se ha hecho frente con las técnicas de programación lineal multicriterio (MCLP) [Shi et al., 2002]. Usando MCLP, se puede optimizar MMD y MSD simultáneamente para identificar la mejor solución de compromiso entre MMD y MSD. La clasificación resultante produce una mejor separación de los datos en relación a los resultados del análisis discriminante lineal [Wang and Shi, 2013], [Wang and Shi, 2012], [Freed and Glover, 1981].

Figura 1.1: Análisis Discriminante Lineal: Solapamiento entre dos clases



Fuente: B. Wang, Y. Shi, *Error Correction Method in Classification by Using Multiple-Criteria and Multiple-Constraint Levels Linear Programming*. Elaborado por: Autor.

En la presente investigación se aborda este problema desde el punto de vista de las máquinas de soporte vectorial, y la programación lineal multicriterio, y se presenta una alternativa para la solución del problema anterior con la aplicación de la programación lineal multicriterio y de restricción múltiple.

1.3 Objetivos de la Investigación

1.3.1 Objetivo General

Estudiar la calidad de predicción de los diferentes modelos matemáticos desarrollados para resolver problemas de clasificación, implementando un modelo basado en programación lineal multicriterio y de restricción múltiple para la calificación crediticia.

1.3.2 Objetivos específicos

1. Estudiar herramientas (metodologías) de Técnicas de Minería de Datos para la toma de decisiones en organizaciones como el sector financiero; determinando las ventajas y desventajas de su utilización de acuerdo a los actuales avances tecnológicos.

2. Definir las características que deben incorporarse al modelo de Clasificación basado en programación lineal multicriterio y de restricción múltiple para tratar el error que se presenta en la estrategia de Clasificación para proporcionar una herramienta útil para las decisiones estratégicas de las organizaciones.

3. Definir las etapas, fases, pasos y actividades que se deben seguir para realizar un modelo de clasificación.

4. Implementar en un solver de programación lineal los diferentes modelos propuestos.

5. Comparar los diferentes modelos matemáticos que se han desarrollado como herramientas de aporte a la toma de decisiones en las organizaciones.

6. Justificar que la aplicación de la Programación Lineal que se propone puede emplearse ventajosamente para resolver problemas de naturaleza compleja, con incertidumbre y múltiples objetivos.

7. Demostrar que los modelos matemáticos que se han desarrollado para ayudar a la toma de decisiones, deben producir resultados equivalentes cuando tratan un mismo problema y los parámetros son los mismos para ambos problemas.

1.4 Justificación

En el escenario actual, las empresas participan en un mercado muy competitivo, donde los clientes se encuentran adecuadamente informados al momento de elegir entre distintas empresas. En mercados donde esto ocurre, la empresa que posea una mayor cantidad de información relevante podrá ejecutar estrategias comerciales efectivas,

sobresaliendo del resto de las empresas. Adicionalmente, la información disponible permite tomar diversas decisiones estratégicas, tales como: definir políticas de asignación de créditos en base al comportamiento histórico de clientes, diseño de nuevos productos a partir de preferencias declaradas, definir campañas que eviten que los clientes se prescindan de los servicios, diagnóstico temprano de tumores cancerígenos.

Si bien obtener información potencialmente útil es cada vez más simple, gracias al importante aumento de la capacidad de almacenaje y la disponibilidad de mejores herramientas para el manejo de datos, el proceso de extracción de información relevante a partir de los datos disponibles sigue siendo complejo y costoso. Actualmente existen técnicas que permiten analizar patrones de conducta, nichos de mercado, y muchos otros tipos de información no trivial mediante la utilización de sofisticados modelos que combinan métodos estadísticos, aprendizaje de máquinas y optimización, técnicas que se engloban bajo el concepto de minería de datos (data mining).

La diversidad y cantidad de problemas que requieren un análisis discriminante se incrementa día a día, y un número significante de ellos, son problemas de clasificación con categorías que no pueden ser definidas precisamente y tienen algún grado de solapamiento (no son excluyentes). Para este tipo de problemas, las técnicas de clasificación basadas en la lógica de Boole (verdadero, falso) presentan serias limitaciones dado que han sido diseñadas para que un objeto sólo pueda pertenecer a una y solo una clase, dando lugar a la asignación de un objeto a una clase errónea, lo que lleva a abordar el problema de minimizar datos mal clasificados, lo que implica una minimización de costos por clasificación.

En una entidad financiera si un cliente es identificado incorrectamente como malo, las consecuencias pueden ser un poco de malestar del cliente además de cumplir más requisitos para optar por un crédito, mientras que por el contrario si un cliente malo se califica como bueno, el resultado puede ser, la pérdida del monto destinado al crédito; así predecir un 0.01 % de clientes de dudoso recaudo puede salvar millones a la institución financiera, mientras que la pérdida de potenciales clientes buenos no influye mucho.

Esta investigación está enfocada al análisis de los datos a través de la programación lineal como una herramienta útil en la toma de decisiones en problemas complejos. Se entienden como tales aquellos que poseen una gran cantidad de variables que se traducen en restricciones de tipo cuantitativo y cualitativo, con numerosas interrelaciones, que corresponden a las alternativas del estudio y a los criterios de decisión. Además, son problemas que se caracterizan por tener que cumplir con varios objetivos simultá-

neamente, muchas veces opuestos o contradictorios.

La toma de decisiones comienza por intentar modelar una situación real, que además debe tener en consideración la experiencia y preferencias del centro decisor (actividad subjetiva) por lo que, considerando la cantidad de variables a manejar, los criterios para evaluar las diversas opciones o alternativas, y la complejidad del problema, se hace pertinente el empleo de modelos matemáticos que organicen la información, ejecuten determinados algoritmos de resolución y que arrojen un resultado, que puede ser aceptado, rechazado o modificado por el centro decisor.

Estas técnicas matemáticas son una ayuda para la toma de decisiones, que apoyan al centro decisor en la gestión de la información, y por otro lado facilitan su labor al permitirle analizar distintas soluciones cuando se cambian determinados parámetros, procedimiento que se conoce como análisis de sensibilidad, y que puede arrojar mucha luz sobre un caso determinado.

Por lo anterior, y dado el continuo crecimiento del mercado de crédito al consumo, la eficiente toma de decisiones es cada vez más importante, tanto en aspectos sociales (eficiencia) como privados (rentabilidad). Frente a ello, existe un creciente interés en modelar de manera más oportuna el riesgo y se han usado modelos estadísticos y máquinas de soporte vectorial como los de credit scoring o scorecard, cuyo objetivo es el de identificar la probabilidad de impago de un grupo de clientes con características similares, el objetivo que perseguimos con este proyecto es realizar está misma tarea pero de forma heurística mediante programación lineal multicriterio y de restricción múltiple (MC2LP).

1.5 Hipótesis

- Hipótesis de Equivalencia: Todos los métodos existentes deben arrojar resultados equivalentes cuando se aplican a la resolución de un mismo problema con los mismos parámetros.

  Sin embargo, puede haber métodos mejores que otros, dado que su error es menor. La hipótesis de equivalencia se considera importante porque permite comprobar, usando diferentes modelos para un mismo caso, si se alcanza una consistencia de resultados, los cuales muy probablemente corresponderán entonces a la mejor solución.

- La minimización de los costos de clasificación, es el criterio técnico que permitirá la asignación óptima de objetos a una clase apropiada mediante la aplicación de

un modelo de programación lineal multicriterio multirestrictiva en el problema de clasificación.

1.6 Estructura del proyecto de titulación

Este estudio constituye un importante aporte metodológico: El problema de clasificación, es uno de los modelos más simples de inferencia inductiva, cuyos resultados pueden ser generalizados a modelos mucho más complejos usando, con pequeñas variaciones, las mismas técnicas matemáticas estándar. De esta forma, la estructura del trabajo es la siguiente:

En este primer capítulo se ha abordado, en líneas generales las distintas metodologías de clasificación actualmente utilizadas, con estos fundamentos, se plantea el objetivo principal del trabajo.

Capítulo 2: partiendo del problema general de clasificación se explica la teoría del Análisis Discriminante Lineal como una de las técnicas estadísticas tradicionales para tratar los problemas de clasificación.

Capítulo 3: Partiendo de una sección introductoria de la Teoría de Aprendizaje Estadístico, donde se tratan conceptos como el riesgo y el riesgo empírico, se estudian los problemas de clasificación donde se ven de forma natural las principales herramientas de trabajo de la metodología de las SVMs, las condiciones de Karush-Kuhn-Tucker, conjuntos separables, hiperplanos separadores, espacios característicos, función núcleo, vectores soporte. De manera general se aborda el estudio de los problemas de clasificación binarios a partir de las máquinas de vectores soporte (SVM). Se construye una máquina lineal SVM para el caso de vectores separables. También se desarrolla las SVMs no lineales introduciendo el concepto de función núcleo. Para finalizar el capítulo se generaliza la máquina de soporte vectorial para los problemas de clasificación con más de dos etiquetas (multiclase).

Capítulo 4: se expone el problema de clasificación desde el punto de vista de la Programación Lineal, desarrollando tanto los principios del análisis discriminante lineal y de la teoría de máquinas de soporte vectorial como formulaciones de problemas de programación lineal. Además el problema de programación lineal es extendido a un modelo de programación lineal multicriterio (MCLP). Se desarrolla un modelo de programación lineal multicriterio para problemas de clasificación dicotómicos.

Capítulo 5: en este capítulo, al igual que en la teoría estadística, se estudia los erro-

res Tipo I y Tipo II que se cometen al trabajar con problemas de clasificación. De esta manera, con la finalidad de minimizar los errores en la clasificación se recurre a la programación lineal multicriterio y de múltiples niveles de restricción (MC2LP) para formular modelos que minimicen el error en problemas de clasificación.

Capítulo 6: La aplicación, comenzando con una descripción detallada sobre la operación de los créditos (como funcionan) para la mayoría de las instituciones dedicadas al microcrédito, así como una exposición de los datos reales de un portafolio con el que se trabajará en términos de número de acreditados, valor monetario, antigüedad y morosidad entre otros, para establecer las condiciones que se tendrán en el otorgamiento de créditos, mismas que servirán para estimar un score de clasificación de riesgo, para ello se implementarán las técnicas expuestas en los capítulos 2, 3, 4 y 5 para después comparar sus resultados y así analizar el error de clasificación en cada uno de los modelos desarrollados, y finalmente valuar la capacidad de clasificación del mejor modelo para futuras observaciones.

Capítulo 7: Conclusiones y Recomendaciones.

Capítulo 2

ANÁLISIS DISCRIMINANTE LINEAL

El análisis discriminante se utiliza en situaciones en las que los grupos son conocidos *a priori*. El objetivo del análisis discriminante es clasificar una observación o varias observaciones, en estos grupos conocidos. Por ejemplo, en la calificación de crédito, un banco sabe por experiencia que hay buenos clientes (quienes pagan su préstamo sin ningún problema) y clientes malos (quienes mostraron dificultades en el pago de su préstamo). Cuando un nuevo cliente solicita un préstamo, el banco tiene que decidir si otorgar o no el préstamo. El historial crediticio de un banco proporcionan observaciones multivariantes x_i sobre las dos categorías de clientes (incluyendo, por ejemplo, la edad, el salario, el estado civil, el monto del préstamo, etc.). El nuevo cliente es una nueva observación x con las mismas variables. La regla de discriminación tiene que clasificar el cliente en uno de los dos grupos existentes y el análisis discriminante debe evaluar el riesgo de una posible "mala decisión". En la mayoría de las aplicaciones, los grupos corresponden a clasificaciones naturales o a los grupos conocidos de la historia (como en el ejemplo de puntuación de crédito). Estos grupos podrían haberse formado por un análisis de conglomerados realizado en datos del pasado [Hardle, 2013], [Ayala, 2008].

2.1 Introducción

Sean Ω_1, Ω_2 dos conjuntos, y $X_1, \cdots, X_2$ variables observables, donde $\mathbf{x} = (x_1, \cdots, x_p)$ representan las observaciones de las variables sobre un individuo ω. Se trata de asignar ω a uno de los dos conjuntos. Este problema aparece en muchas situaciones: decidir si se puede conceder un crédito; determinar si un tumor es benigno o maligno; identificar la especie a que pertenece una planta, etc.

Una **regla discriminante** es un criterio que permite asignar ω conocido $(x_1, \cdots, x_p)$, y que a menudo es planteado mediante una función discriminante $D(x_1, \cdots, x_p)$. Entonces la regla de clasificación es:

$$\begin{cases} \text{Si } D(x_1, \cdots, x_p) \geq 0 & \text{asignamos } \omega \in \Omega_1 \\ \text{en caso contrario} & \text{asignamos } \omega \in \Omega_2 \end{cases}$$

Esta regla divide $\mathbb{R}^p$ en dos regiones

$$\mathbf{R}_1 = \{\mathbf{x}|D(\mathbf{x}) > 0\}, \qquad \mathbf{R}_2 = \{\mathbf{x}|D(\mathbf{x}) < 0\}$$

En la decisión de identificar ω, se comete un error de clasificación si se equivoca y se asigna ω a una población a la que no pertenece. Si estamos trabajando sólo con dos grupos, en la asignación existen dos posibles errores: el que se comete al clasificarlo en el primer grupo, cuando en realidad pertenece al segundo $P(\mathbf{R}_1|\Omega_2)$, y el que se cometería al incluirlo en el segundo grupo, cuando en realidad pertenece al primero $P(\mathbf{R}_2|\Omega_1)$. El criterio matemático de clasificación se determina de tal manera que minimice la probabilidad de error, la probabilidad de clasificación errónea (pce) es:

$$pce = P(\mathbf{R}_2|\Omega_1)P(\Omega_1) + P(\mathbf{R}_1|\Omega_2)P(\Omega_2) \tag{2.1}$$

2.2 Clasificación en dos poblaciones

2.2.1 Discriminador lineal

Sean μ_1, μ_2 los vectores de medias de las variables en Ω_1, Ω_2, respectivamente, y supongamos que la matriz de covarianzas Σ es común. Las distancias de Mahalanobis de las observaciones $\mathbf{x} = (x_1, \cdots, x_p)'$ de un individuo ω a las poblaciones son

$$M^2(\mathbf{x}, \mu_i) = (\mathbf{x} - \mu_i)'\Sigma^{-1}(\mathbf{x} - \mu_i), \qquad i = 1, 2.$$

Un criterio o regla de Clasificación consisten en asignar ω a la población más próxima:

$$\begin{cases} Si\ M^2(\mathbf{x}, \mu_1) < M^2(\mathbf{x}, \mu_2) & \text{asignamos } \omega\ a\ \Omega_1, \\ Si\ M^2(\mathbf{x}, \mu_2) \leq M^2(\mathbf{x}, \mu_1) & \text{asignamos } \omega\ a\ \Omega_2 \end{cases} \tag{2.2}$$

Expresando esta regla como una función discriminante, se tiene:

$$\begin{aligned} M^2(\mathbf{x},\mu_2) - M^2(\mathbf{x},\mu_1) &= (\mathbf{x}-\mu_2)'\Sigma^{-1}(\mathbf{x}-\mu_2) \\ &\quad -(\mathbf{x}-\mu_1)'\Sigma^{-1}(\mathbf{x}-\mu_1) \\ &= \mathbf{x}'\Sigma^{-1}\mathbf{x} + \mu_2'\Sigma^{-1}\mu_2 - 2\mathbf{x}'\Sigma^{-1}\mu_2 \\ &\quad -\mathbf{x}'\Sigma^{-1}\mathbf{x} - \mu_1'\Sigma^{-1}\mu_1 + 2\mathbf{x}'\Sigma^{-1}\mu_1 \\ &= (\mu_2-\mu_1)'\Sigma^{-1}(\mu_2+\mu_1) + 2\mathbf{x}'\Sigma^{-1}(\mu_1-\mu_2). \end{aligned}$$

Se define la función lineal discriminante:

$$L(\mathbf{x}) = \left[\mathbf{x} - \frac{1}{2}(\mu_1+\mu_2)\right]' \Sigma(\mu_1-\mu_2) \tag{2.3}$$

Entonces

$$M^2(\mathbf{x},\mu_2) - M^2(\mathbf{x},\mu_1) = 2L(\mathbf{x}) - L((\mu_1+\mu_2)/2)$$

y la regla (2.2) es

$$\begin{cases} Si\ L(\mathbf{x}) > 0 & \text{asignamos } \omega\ a\ \Omega_1, \\ Si\ L(\mathbf{x}) \leq 0 & \text{asignamos } \omega\ a\ \Omega_2. \end{cases} \tag{2.4}$$

La función lineal (2.3) es el discriminador de Fisher [Ayala, 2008].

2.2.2 Regla de máxima verosimilitud

Supongamos que $f_1(\mathbf{x})$; $f_2(\mathbf{x})$ son las densidades de $\mathbf{x}$ en Ω_1, Ω_2. La regla discriminante de máxima verosimilitud consisten en asignar ω a la población Ω_i para la cual la verosimilitud de la observación es mayor:

$$\begin{cases} Si\ f_1(\mathbf{x}) > f_2(\mathbf{x}) & \text{asignamos } \omega\ a\ \Omega_1, \\ Si\ f_1(\mathbf{x}) < f_2(\mathbf{x}) & \text{asignamos } \omega\ a\ \Omega_2. \end{cases} \tag{2.5}$$

La función discriminante es

$$V(\mathbf{x}) = \log f_1(\mathbf{x}) - \log f_2(\mathbf{x})$$

2.2.3 Regla de Bayes

En ciertas situaciones, se conocen las probabilidades *a priori* de que ω pertenezca a cada una de las poblaciones:

$$q_1 = P(\Omega_1),$$

$$q_2 = P(\Omega_2),$$

$$q_1 + q_2 = 1$$

Una vez se dispone de las observaciones $\mathbf{x} = (x_1, \cdots, x_p)$, las probabilidades *a posteriori* de que ω pertenezca a las poblaciones (teorema de Bayes) son

$$P(\Omega_i|\mathbf{x}) = \frac{q_i f_i(x)}{q_1 f_1(x) + q_2 f_2(x)}, \qquad i = 1, 2.$$

La regla discriminante de Bayes consiste en asignar ω a la población Ω_i para la que $P(\omega \in \Omega_i|\mathbf{x})$ es máxima

$$\text{Si} \begin{cases} P(\Omega_1|\mathbf{x}) > P(\Omega_2|\mathbf{x}) & \text{asignamos } \omega \in \Omega_1, \\ P(\Omega_1|\mathbf{x}) < P(\Omega_2|\mathbf{x}) & \text{asignamos } \omega \in \Omega_2, \end{cases}$$

La regla de Bayes tiene asociada la siguiente función discriminante, que se conoce como **discriminador de Bayes**:

$$B(\mathbf{x}) = \log f_1(\mathbf{x}) - \log f_2(\mathbf{x}) + \log(q_1/q_2).$$

Propiedades:

1. Cuando $q_1 = q_2 = 1/2$, entonces $B(\mathbf{x}) = V(\mathbf{x})$. Este discriminador es óptimo.
2. La regla de Bayes minimiza la probabilidad de clasificación errónea.

Teorema 1. *La regla de Bayes minimiza la probabilidad de clasificación errónea.*

Demostración. Supongamos que se dispone de otra regla que clasifica a Ω_1 si $\mathbf{x} \in \mathbf{R}_1^*$, y a Ω_2 si $\mathbf{x} \in \mathbf{R}_2^*$, donde $\mathbf{R}_1^*$, $\mathbf{R}_2^*$ son regiones complementarias del espacio muestral. Indicando $d\mathbf{x} = dx_1 \cdots dx_p$, la probabilidad de clasificación errónea es

$$\begin{aligned} pce^* &= q_1 \int_{\mathbf{R}_2^*} f_1(\mathbf{x}) d(\mathbf{x}) + q_2 \int_{\mathbf{R}_1^*} f_2(\mathbf{x}) d(\mathbf{x}) \\ &= \int_{\mathbf{R}_2^*} (q_1 f_1(\mathbf{x}) - q_2 f_2(\mathbf{x}))\, d\mathbf{x} + q_2 \left(\int_{\mathbf{R}_2} f_2(\mathbf{x}) d\mathbf{x} + \int_{\mathbf{R}_1^*} f_2(\mathbf{x}) d\mathbf{x} \right) \\ &= \int_{\mathbf{R}_2^*} (q_1 f_1(\mathbf{x}) - q_2 f_2(\mathbf{x}))\, d\mathbf{x} + q_2 \end{aligned}$$

Indiquemos $z = q_1 f_1(\mathbf{x}) - q_2 f_2(\mathbf{x})$. Esta última integral es mínima si $\mathbf{R}_2^*$ incluye todas las $\mathbf{x}$ tales que $z < 0$ y excluye todas las $\mathbf{x}$ tal que $z > 0$. Por tanto pce^* es mínima si $\mathbf{R}_2^* = \mathbf{R}_2$, siendo $\mathbf{R}_2 = \{\mathbf{x}|B(\mathbf{x}) < 0\}$. □

2.3 Clasificación en poblaciones normales

Supongamos ahora que:

$$\mathbf{X} = (X_1, \cdots, X_p) \sim N_p(\mu_1, \Sigma_1) \ en\ \Omega_1$$

$$\mathbf{X} = (X_1, \cdots, X_p) \sim N_p(\mu_2, \Sigma_2) \ en\ \Omega_2$$

es decir:

$$L_1(\mathbf{x}) = \frac{|\Sigma_i|^{-1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_i)^t \Sigma_i^{-1}(\mathbf{x}-\mu_i)\right\}, \qquad i+1, 2.$$

Caso 1: Si $\mu_1 \neq \mu_2$ y $\Sigma_1 = \Sigma_2 = \Sigma$, entonces:

a) Los clasificadores de máxima verosimilitud y lineal coinciden:

$$\begin{aligned} V(x) &= \log L_1(\mathbf{x}) - \log L_2(\mathbf{x}) \\ &= \frac{1}{2}\left((x-\mu_2)^t\Sigma^{-1}(x-\mu_2) - (x-\mu_1)^t\Sigma^{-1}(x-\mu_1)\right) \\ &= L(\mathbf{x}) \end{aligned}$$

b) Si $\mathbf{x} \in \mathbb{R}^p$ proviene de alguna de las poblaciones Ω_i, para $i = 1, 2$, entonces el discriminador lineal de Fisher tiene distribución normal:

$$L(x) = \left(\mathbf{x} - \frac{1}{2}(\mu_1+\mu_2)\right)^t \Sigma^{-1}(\mu_1-\mu_2) = (\mathbf{x}-\mu)^t\mathbf{a} = \mathbf{a}^t(\mathbf{x}-\mu),$$

donde $\mathbf{a} = \Sigma^{-1}(\mu_1-\mu_2)$ y $\mu = (\mu_1+\mu_2)/2$.

Su varianza y esperanza son:

$$var(L(\mathbf{x})) = var(\mathbf{a}^t(\mathbf{x}-\mu)) = \mathbf{a}^t\Sigma\mathbf{a} = M^2$$

$$E(L(\mathbf{x})) = \mathbf{a}^t E(\mathbf{x} - \mu) = \begin{cases} \frac{1}{2}\mathbf{a}^t(\mu_1 - \mu_2) = \frac{1}{2}M^2, & si\ \mathbf{x} \in \Omega_1, \\ -\frac{1}{2}\mathbf{a}^t(\mu_1 - \mu_2) = -\frac{1}{2}M^2, & si\ \mathbf{x} \in \Omega_2, \end{cases}$$

Por tanto,

$$L(\mathbf{x}) \sim N\left(\frac{1}{2}M^2, M^2\right)\ si\ \mathbf{x} \in \Omega_1$$

$$L(\mathbf{x}) \sim N\left(-\frac{1}{2}M^2, M^2\right)\ si\ \mathbf{x} \in \Omega_2$$

Puesto que $L(\mathbf{x})$ tiene distribución de probabilidad conocida, puede calcularse la **la probabilidad de clasificación errónea**.

Se dice que el individuo $\mathbf{x}$ se clasifica erróneamente cuando se asigna a la población Ω_1 y en realidad proviene de Ω_2, o bien cuando se asigna a la población Ω_2 y en realidad proviene de Ω_1.

Luego la probabilidad de clasificación errónea es:

$$\frac{1}{2}P(L(x) > 0|\mathbf{x} \in \Omega_2) + \frac{1}{2}P(L(x) < 0|\mathbf{x} \in \Omega_1) = \phi\left(-\frac{M}{2}\right)$$

donde ϕ es la función de distribución de la ley $N(0, 1)$.

c) Si conocemos las probabilidades *a priori* $q_1 = P(\omega \in \Omega_1)$, $q_2 = P(\omega \in \Omega_2)$, con $q_1 + q_2 = 1$, entonces el discriminador de Bayes es: $B(\mathbf{x}) = L(\mathbf{x}) + \log(q_1/q_2)$.

Caso 2: Si $\mu_1 \neq \mu_2$ y $\Sigma_1 \neq \Sigma_2$, entonces:

a) La regla de máxima verosimilitud proporciona el **discriminador cuadrático**

$$\begin{aligned} V(\mathbf{x}) &= \log f_1(\mathbf{x}) - \log f_2(\mathbf{x}) \\ &= \frac{1}{2}\mathbf{x}^t\left(\Sigma_2^{-1} - \Sigma_1^{-1}\right)\mathbf{x} + \mathbf{x}^t\left(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2\right) \\ &= +\frac{1}{2}\mu_2^t\Sigma_2^{-1}\mu_2 - \frac{1}{2}\mu_1^t\Sigma_1^{-1}\mu_1 + \frac{1}{2}\log|\Sigma_2| - \frac{1}{2}\log|\Sigma_1| \\ &= Q(\mathbf{x}) \end{aligned}$$

b) Si conocemos las probabilidades *a priori* $q_1 = P(\omega \in \Omega_1)$, $q_2 = P(\omega \in \Omega_2)$, con $q_1 + q_2 = 1$, entonces el discriminador de Bayes es: $B(\mathbf{x}) = Q(\mathbf{x}) + \log(q_1/q_2)$ [Ayala, 2008].

2.3.1 Clasificación si los parámetros son estimados

En las aplicaciones prácticas, $\mu_1, \mu_2, \Sigma_1, \Sigma_2$ son desconocidos y se deberán estimar a partir de muestras de tamaños $n_1,\ n_2$ de las dos poblaciones sustituyendo μ_1, μ_2 por los vectores de medias $\bar{\mathbf{x}_1}, \bar{\mathbf{x}_2}$, y Σ_1, Σ_2 por las matrices de covarianzas $\mathbf{S}_1, \mathbf{S}_2$. Si utilizamos el estimador lineal, entonces la estimación de Σ será:

$$\mathbf{S} = \frac{(n_1\mathbf{S}_1 + n_2\mathbf{S}_2)}{(n_1 + n_2)}$$

y la versión muestral del discriminador lineal es:

$$\hat{L}(\mathbf{x}) = \left[\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x_1}} + \bar{\mathbf{x_2}})\right]^t \mathbf{S}^{-1}(\bar{\mathbf{x_1}} + \bar{\mathbf{x_2}}).$$

La distribución muestral de $\hat{L}(\mathbf{x})$ es bastante complicada, pero la distribución asintótica es normal:

$$\hat{L}(\mathbf{x}) \text{ es } N(+\frac{1}{2}\alpha, \alpha) \text{ si } \mathbf{x} \text{ proviene de } N_p(\mu_1, \Sigma),$$

$$\hat{L}(\mathbf{x}) \text{ es } N(-\frac{1}{2}\alpha, \alpha) \text{ si } \mathbf{x} \text{ proviene de } N_p(\mu_2, \Sigma),$$

donde $\alpha = (\bar{\mathbf{x_1}} + \bar{\mathbf{x_2}})^t\mathbf{S}^{-1}(\bar{\mathbf{x_1}} + \bar{\mathbf{x_2}})$.

2.4 Clasificación en el caso de K poblaciones

Supongamos ahora que el individuo ω puede provenir de k poblaciones $\Omega_1, \Omega_2, \cdots, \Omega_k$, donde $k \geq 3$. Es necesario establecer una regla que permita asignar ω a una de las k poblaciones sobre la base de las observaciones $\mathbf{x} = (x_1, x_2, \cdots, x_p)^t$ de p variables.

2.4.1 Discriminadores lineales

Supongamos que la media de las variables en Ω_1 es μ_i; y que la matriz de covarianzas Σ es común. Si consideramos las distancias de Mahalanobis de ω a las poblaciones

$$M^2(\mathbf{x}, \mu_i) = (\mathbf{x} - \mu_i)^t\Sigma^{-1}(\mathbf{x} - \mu_i), \qquad i = 1, 2.$$

un criterio de clasificación consiste en asignar ω a la población más próxima:

$$Si\ M^2(\mathbf{x}, \mu_i) = \text{mín}\left\{M^2(\mathbf{x}, \mu_1), \cdots, M^2(\mathbf{x}, \mu_k)\right\}, \text{ asignamos } \omega\ a\ \Omega_i \qquad (2.6)$$

Introduciendo las funciones discriminantes lineales

$$L_{ij}(\mathbf{x}) = (\mu_i - \mu_j)^t \Sigma^{-1} \mathbf{x} - \frac{1}{2}(\mu_i - \mu_j)^t \Sigma^{-1}(\mu_i + \mu_j)$$

es fácil probar que (2.6) equivale a

$$Si L_{ij}(\mathbf{x}) > 0 \text{ para todo} j \neq i, \text{ asignamos } \omega \; a \; \Omega_i.$$

Además las funciones $L_{ij}(\mathbf{x})$ verifican:

1. $L_{ij}(\mathbf{x}) = \frac{1}{2}\left[M^2(\mathbf{x}, \mu_j) - M^2(\mathbf{x}, \mu_i)\right]$.
2. $L_{ij}(\mathbf{x}) = -L_{ij}(\mathbf{x})$.
3. $L_{rs}(\mathbf{x}) = L_{is}(\mathbf{x}) - L_{ir}(\mathbf{x})$

Es decir, sólo necesitamos conocer $k-1$ funciones discriminantes.

2.4.2 Regla de la máxima verosimilitud

Sea $f_i(\mathbf{x})$ la función de densidad de $\mathbf{x}$ en la población Ω_i. Podemos obtener una regla de clasificación asignando ω a la población donde la verosimilitud es más grande:

$$\text{Si } f_i(\mathbf{x}) = \text{máx}\{f_1(x), \cdots, f_k(x)\}, \text{ asignamos } \omega \text{ a } \Omega_i.$$

Este criterio es más general que el geométrico y está asociado a las funciones discriminates

$$V_{ij} = \log f_i(\mathbf{x}) - \log f_j(\mathbf{x}).$$

En el caso de normalidad multivariante y matriz de covarianzas común, se verifica $V_{ij}(\mathbf{x}) = L_{ij}(\mathbf{x})$; y los discriminadores máximo verosímiles coinciden con los lineales. Pero si las matrices de covarianzas son diferentes $\Sigma_1, \cdots, \Sigma_k$, entonces este criterio dará lugar a los discriminadores cuadráticos

$$\begin{aligned} Q_{ij} \;=\; & \frac{1}{2}\mathbf{x}^t\left(\Sigma_i^{-1}\mu_1 - \Sigma_i^{-1}\right)\mathbf{x} + \mathbf{x}^t\left(\Sigma_i^{-1}\mu_1 - \Sigma_j^{-1}\mu_2\right) \\ & + \frac{1}{2}\mu_j^t\Sigma_j^{-1}\mu_j - \frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i + \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}\log|\Sigma_i|. \end{aligned}$$

2.4.3 Reglas de Bayes

Si además de las funciones de densidad $f_i(\mathbf{x})$, se conocen las probabilidades a priori

$$q_1 = P(\Omega_1), \cdots, q_k = P(\Omega_k),$$

la regla de Bayes que asigna ω a la población tal que la probabilidad a posteriori es máxima

$$\text{Si } q_i f_i(\mathbf{x}) = \max\{q_1 f_1(\mathbf{x}), \cdots, q_k f_k(\mathbf{x})\}, \text{ asignamos } \omega \; a \; \Omega_i,$$

está asociada a las funciones discriminantes

$$B_{ij} = \log f_i(\mathbf{x}) - \log f_j(\mathbf{x}) + \log(q_i/q_j).$$

Finalmente, si $P(j/i)$ es la probabilidad de asignar ω a Ω_j cuando en realidad es de Ω_i, la probabilidad de clasificación errónea es

$$pce = \sum_{i=1}^{k} q_i \left(\sum_{j \neq i}^{k} P(j/i) \right),$$

y se demuestra que la regla de Bayes minimiza esta *pce*, [Ayala, 2008],[Cuadras, 2014].

Capítulo 3

MÁQUINAS DE SOPORTE VECTORIAL

Las máquinas de soporte vectorial o máquinas de vectores de soporte (Support Vector Machines, SVMs) son un conjunto de algoritmos de aprendizaje supervisado. Las máquinas de soporte vectorial son métodos relativamente eficientes para tratar problemas de clasificación y regresión, y que se han posicionado por encima de otras técnicas, tales como las redes neuronales. La idea básica es dado un conjunto de observaciones (muestra, *training set*) etiquetadas previamente en dos grupos, como por ejemplo, +1 y -1. Se debe encontrar un hiperplano que separe los datos perfectamente en las dos clases. Por lo tanto, la clasificación consiste en determinar a qué lado del hiperplano corresponde el vector de características de una nueva observación.

En el caso más simple, caso linealmente separable (*soft margin*) existe una distancia positiva entre ambos grupos, siendo posible determinar el hiperplano que maximiza la distancia de cada grupo a éste. Por otra parte, en el caso más complejo caso no linealmente separable, los grupos se superponen espacialmente siendo imposible la separación por medio de un hiperplano, en este caso, las SVMs recurren a la utilización de núcleos (*kernel*), es decir, obtener una transformación no lineal del espacio de entrada (espacio de características) en otro espacio de mayor dimensión (o incluso infinita) en el cual, por efecto del aumento de la dimensión y la no linealidad de la transformación, los grupos así transformados pueden ser separados por un hiperplano como el primer caso. Este tipo de transformación normalmente puede traer complejidades computacionales difíciles de resolver, pero la clave en la utilización de estas metodologías radica en que el nuevo espacio de dimensión mayor no tiene que ser tratado directamente, es suficiente con conocer cómo resolver el producto escalar en el mismo. Una manera de lograr que el hiperplano separe solo por información relevante, es tolerar que algunas observaciones queden mal clasificadas. Para ello es necesaria la introducción de las variables de holgura (*slack*) que cuantifican la distancia de un punto mal clasificado al hiperplano de separación.

3.1 Introducción

La teoría de las SVMs fue desarrollada por V.Vapnik a principios de los años 80 y se centra en lo que se conoce como Teoría de Aprendizaje Estadístico[1]. V. Vapnik propone un modelo matemático para la resolución de problemas de clasificación y regresión [Vapnik, 1998]. Se crearon aplicaciones para la solución de problemas reales, destacándose como una herramienta robusta en dominios complejos, ruidosos y con escasos datos.

3.1.1 Nociones sobre la teoría del aprendizaje estadístico

A continuación se expone brevemente el planteamiento general de las SVMs para posteriormente afrontar los problemas de clasificación desde la perspectiva de un aprendizaje supervisado, es decir, el conocimiento de las salidas de un conjunto de entradas permite cuantificar (supervisar) la bondad de los resultados del modelo.

El objetivo fundamental de este tipo de estudios es aprender a partir de los datos y para ésto se busca la existencia de alguna dependencia funcional entre un conjunto de vectores de entrada (inputs)

$$\{x_i : i = 1, \cdots, n\} \subseteq \mathcal{X} \subseteq \mathbb{R}^d$$

y valores de salidas (outputs)

$$\{y_i : i = 1, \cdots, n\} \subseteq \mathcal{Y} \subseteq \mathbb{R}$$

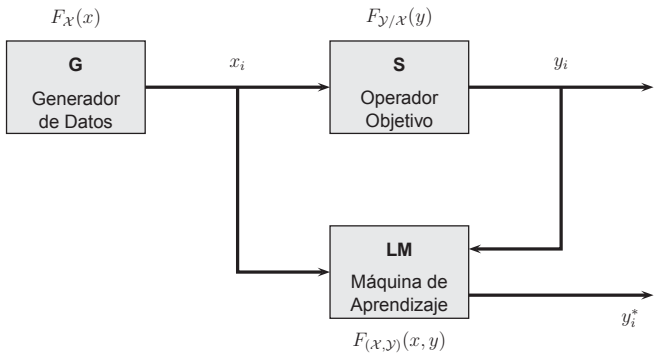
El modelo ilustrado por la figura 3.1 recoge de manera sencilla el propósito que se persigue.

En este esquema **G** representa un modelo generador de datos que proporciona los vectores $x_i \in \mathcal{X}$, independientes e idénticamente distribuidos de acuerdo con una función de distribución $F_{\mathcal{X}}(x)$ desconocida pero que se supone no cambia a lo largo del proceso de aprendizaje. Cada vector x_i es la entrada del operador objetivo **S**, el cual lo transforma en un valor y_i según una función de distribución condicional $F_{\mathcal{Y}/\mathcal{X}=x_i}(y)$. Así la máquina de aprendizaje, que denotamos **LM** (learning machine) agrupa la siguiente base de entrenamiento,

$$Z = \{(x_1, y_1), \cdots, (x_n, y_n)\} \subseteq \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$$

[1] Esta teoría busca formas de estimar dependencias funcionales a partir de una colección de datos.

Figura 3.1: Esquema de configuración de una máquina de aprendizaje.



Fuente: González Abril, Modelos de Clasificación basados en Máquinas de Vectores Soporte. Elaborado por: Autor.

el cual es obtenido independiente e idénticamente distribuido siguiendo la función de distribución conjunta:

$$F_{(\mathcal{X},\mathcal{Y})}(x,y) = F_{\mathcal{X}}(x) \cdot F_{\mathcal{Y}/\mathcal{X}=x}(y)$$

a partir del conjunto de entrenamiento Z, la máquina de aprendizaje "construye" una aproximación al operador desconocido la cual proporcione para un generador dado **G**, la mejor aproximación a las salidas proporcionadas por el supervisor. De manera formal construir un operador significa que la máquina de aprendizaje implementa una familia de funciones, de tal forma que durante el proceso de aprendizaje, elige de esta familia una función apropiada siguiendo una determinada regla de decisión.

La estimación de esta dependencia estocástica basada en un conjunto de datos trata de aproximar la función de distribución condicional $F_{\mathcal{Y}/\mathcal{X}}(y)$, esto en general conduce a un problema realmente complicado [Vapnik, 1998], [González, 2002]; asiduamente se está interesado solo en alguna de sus características. Por ejemplo se puede buscar estimar la función de esperanza matemática condicional:

$$\mathbb{E}[\mathcal{Y}/\mathcal{X} = x] =^{def} \int y dF_{\mathcal{Y}/x}(y)$$

Por lo tanto, el objetivo del problema es la construcción de una función $f(x,y)$ dentro de una determinada clase de funciones $\mathcal{F}$ elegida a priori, la cual debe cumplir un determinado criterio, formalmente el problema se puede plantear como sigue:

Dado un espacio vectorial $\mathcal{Z}$ de $\mathbf{R}^{d-1}$ donde se define una medida de probabilidad $F_{\mathcal{F}}(z)$, un conjunto $\mathcal{F} = \{f(z), z \in \mathcal{Z}\}$ de funciones reales y un funcional $\mathbf{R} : \mathcal{F} \to \mathbb{R}$. Buscar una función $f^* \in \mathcal{F}$ tal que

$$\mathbf{R}[f^*] = \min_{f \in \mathcal{F}} \mathbf{R}[f]$$

Con objeto de ser lo más general posible sería bueno elegir el funcional $\mathbf{R}$ de tal manera que se pudiese plantear con él, el mayor número de problemas posibles. Por ello se define **riesgo**, $\mathbf{R} : \mathcal{F} \to \mathbb{R}$ como sigue:

Definición 1. *Dada una clase $\mathcal{F} = \{f(z), z \in \mathcal{Z}\}$ de funciones reales y una medida de probabilidad $F_{\mathcal{Z}}(z)$ se define el **riesgo**, $\mathbf{R} : \mathcal{F} \to \mathbb{R}$, como:*

$$\mathbf{R}[f] = \int_{\mathcal{Z}} c(z, f(z)) dF_{\mathcal{Z}(z)} \tag{3.1}$$

donde c(.,.) se denomina función pérdida (o de coste) y tomará valores no negativos.

A la vista de la figura 3.1 se llega a la conclusión que los valores y_i e y_i^* no necesariamente coinciden. Cuando esto sea así, la máquina de aprendizaje habrá cometido un error que se debe cuantificar de alguna forma y este es precisamente el sentido que tiene la función de perdida.

De esta manera, en este planteamiento, dado un conjunto $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, el principal problema consiste en formular un criterio constructivo para elegir una función de $\mathcal{F}$ puesto que el funcional (3.1) por sí mismo no sirve como criterio de selección, ya que la función $F_{\mathcal{Z}}(z)$ incluida en él es desconocida. Para elaborar dicho criterio se debe tener en cuenta que el *riesgo* se define como la esperanza matemática de una variable aleatoria respecto a una medida de probabilidad, por tanto es lógico elegir como estimación, la media muestral y de aquí la siguiente definición:

Definición 2. *Dado un riesgo definido por 3.1, un conjunto de funciones $\mathcal{F}$ y una muestra $\{z_1, \cdots, z_n\}$. Al funcional $\mathbf{R}_{emp} : \mathcal{F} \to \mathbb{R}$ definido como*

$$\mathbb{R}_{emp}[f] = \frac{1}{n} \sum_{i=1}^{n} c(z_i, (f(z_i))), \qquad f \in \mathcal{F}$$

*se le denomina **riesgo empírico***

La forma clásica de abordar estos problemas es: si el valor mínimo del riesgo se alcanza con una función f_0 y el mínimo del riesgo empírico con f_n para una muestra dada de tamaño n, entonces se considera que f_n es una aproximación a f_0 en un determinado espacio métrico. El principio que resuelve este problema se denomina principio de minimización del riesgo empírico. Este es el principio utilizado en los desarrollos clásicos por

ejemplo cuando se plantea a partir de un conjunto de datos la regresión lineal mínimo cuadrática.

De lo expuesto hasta aquí, surge la pregunta: puede asegurarse que el riesgo $\mathbf{R}[f_n]$ está cerca del $\min_{f\in\mathcal{F}} \mathbf{R}[f]$? La respuesta es que en general esto no es cierto, entonces, se llega a la conclusión que la clase de funciones $\mathcal{F}$ no puede ser arbitraria, necesariamente se debe imponer algunas condiciones de regularidad a sus elementos, vía un funcional $\mathbf{Q}[f]$. Así en la elaboración del problema se debe buscar una adecuada relación entre la precisión alcanzada con un particular conjunto de entrenamiento, medido a través de $\mathbf{R}_{emp}[f]$, y la capacidad de la máquina medida por $\mathbf{Q}[f]$. Ello lleva a considera el problema de minimizar un riesgo regularizado, donde este se define para algún $\lambda > 0$ en la forma:

$$\mathbb{R}_{reg}[f] = \mathbb{R}_{emp}[f] + \lambda\mathbf{Q}[f]$$

Indicar que en las SVMs, el espacio de trabajo es

$$\mathcal{F} = \{f(x) = \omega \cdot x + b, \omega \in \mathbb{R}^d, b \in \mathbb{R}\}$$

y la restricción se impone sobre el parámetro ω, [González, 2002].

3.1.2 Máquinas de soporte vectorial para clasificación

La mejor manera de introducción a las Máquinas de Soporte Vectorial (SVMs) es considerar la tarea más simple de clasificación binaria. Muchos de los problemas del mundo real involucran predicción sobre dos clases. Así por ejemplo se puede pretender predecir si un tipo de cambio de divisas se desplazará hacia arriba (aumenta) o hacia abajo (disminuye), dependiendo de los datos económicos, o si a un cliente se le debe otorgar o no un préstamo en base a su historial crediticio. Una SVM es un una *máquina de aprendizaje* abstracta que ensayará (entrenará) de un *conjunto de datos de entrenamiento* y conseguirá *generalizar* y hacer predicciones correctas sobre futuras observaciones.

Para los datos de entrenamiento se tiene un conjunto de vectores de entrada, denotada $\mathbf{x}_i$, donde cada vector de entrada tiene una serie de *características* de los componentes. Estos vectores de entrada se emparejan con las *etiquetas* correspondientes, que se denota y_i, y existe m pares $(i = 1, \cdots, m)$. Para esta explicación, se utiliza una SVM para predecir la recaída frente a la no recaída de un cáncer en particular, basado en los datos genéticos. En este ejemplo, los casos de recaída serían etiquetados como $y_i = +1$, y por otra parte los casos de no recaída como $y_i = -1$, con la correspondiente $\mathbf{x}_i$, que son vectores de entrada que codifican los datos genéticos derivados de cada paciente

i. Por lo general, estaríamos interesados en cuantificar el desempeño de la SVM antes de cualquier uso práctico, por lo que se evaluaría una *prueba de error* basada en datos de un *conjunto de prueba*.

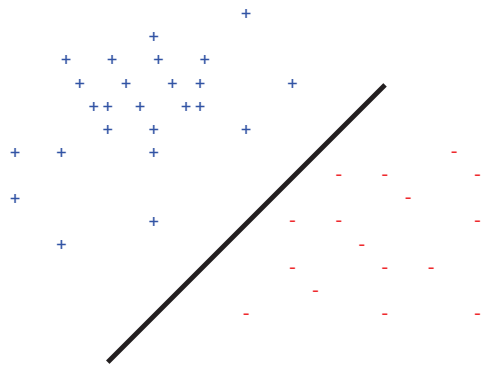


Figura 3.2: Clasificación binaria.-El argumento dentro de la función de un clasificador es $\omega \cdot \mathbf{x} + \mathbf{b}$. El hiperplano separador correspondiente a $\omega \cdot \mathbf{x} + \mathbf{b} = 0$ se muestra con una línea en este gráfico bidimensional. Este hiperplano separa las dos clases de datos, un lado etiquetado como $y_i = +1(\omega \cdot \mathbf{x} + \mathbf{b} \geq 0)$ y otro lado etiquetado como $y_i = -1(\omega \cdot \mathbf{x} + \mathbf{b} < 0)$.

Fuente: Colin C., Learning with Support Vector Machines. Elaborado por: Autor.

Los datos de entrenamiento pueden ser vistos como puntos de datos marcados en un espacio de entrada como se representa en la Figura 3.2. Para dos clases de datos bien separados, la tarea de aprendizaje consiste en encontrar un hiperplano dirigido, es decir un hiperplano orientado de tal manera que los puntos de datos en un lado serán etiquetados como $y_i = +1$ y aquellos en el otro lado como $y_i = -1$. El hiperlano dirigido encontrado por una máquina de vectores de soporte es intuitivo: es decir, que el hiperplano es la máxima distancia entre las dos clases de puntos etiquetados, ubicados a cada lado. Tales puntos más cercanos en ambos lados tienen mayor influencia sobre la posición de este hiperplano de separación y por lo tanto se denominan vectores de soporte. El hiperplano separador es dado como $\omega \cdot \mathbf{x} + b = 0$ (donde $\cdot$ representa el producto interior o escalar), b es el sesgo o desviación del hiperplano desde el origen en el espacio de entrada, $\mathbf{x}$ son puntos situados dentro del hiperplano y la normal al hiperplano, los pesos ω determinan su orientación.

Por su puesto, esta imagen es demasiado simple para muchas aplicaciones. En la figura 3.2 se muestran dos grupos etiquetados que son fácilmente separables por un hiperplano que es simplemente una línea en esta ilustración bidimensional. En realidad, los dos grupos podrían ser altamente entrelazados con la superposición (solapamiento) de puntos de datos: así el conjunto de datos no es linealmente separable. Esta situación

es una de las motivaciones para introducir el concepto de los núcleos más adelante en este capítulo. También podemos ver que los puntos de datos perdidos podrían actuar como vectores de soporte anómalos con un impacto significativo en la orientación del hiperplano: que por lo tanto necesitan tener un mecanismo de gestión de puntos de datos ruidosos y anómalos. Por otra parte, tenemos que ser capaces de manejar datos multiclase [Colin and Yiming, 2011].

3.2 Máquinas de soporte vectorial para clasificación lineal binaria

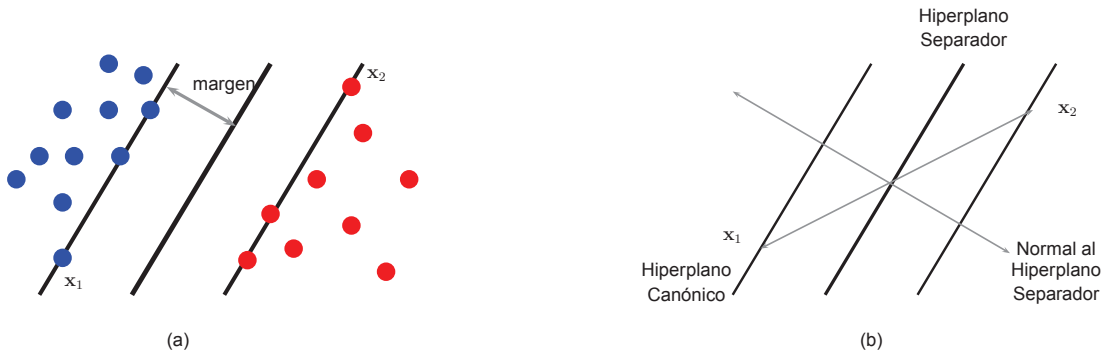


Figura 3.3: Clasificación binaria, caso linealmente separable. (a): La distancia perpendicular entre el hiperplano separador y un hiperplano a través de los puntos más cercanos (los vectores de soporte) es llamado el *margen*, γ. $\mathbf{x}_1$ y $\mathbf{x}_2$ son ejemplos de *vectores de soportes* de signos opuestos. Los hiperplanos que pasan a través de los vectores de soporte son los *hiperplanos canónicos* y la región entre los hiperplanos canónicos es la *banda de margen*. (b): la proyección del vector $(\mathbf{x}_1 - \mathbf{x}_2)$ sobre la normal a la hiperplano de separación $(\omega/||\omega||_2)$ es 2γ.

Fuente: Colin C., Learning with Support Vector Machines. Elaborado por: Autor.

La teoría del aprendizaje estadístico es una aproximación teórica a la comprensión del aprendizaje y la capacidad de las máquinas de aprendizaje para generalizar. Desde la perspectiva de la teoría del aprendizaje estadístico, la motivación para considerar SVMs como clasificador binario viene de un límite superior teórico en la *generalización de error*, es decir, la predicción teórica del error cuando se aplica el clasificador a observaciones futuras. Este límite para la generalización del error tiene dos características importantes:

- el límite se reduce al mínimo mediante la maximización del margen, γ, es decir, la distancia mínima entre el hiperplano que separa las dos clases y los puntos de datos más cercanos al hiperplano (ver figura 3.3),

- el límite no depende de la dimensionalidad del espacio.

3.2.1 Caso separable linealmente

Dado un conjunto finito de datos de entrenamiento (ensayo)

$$S = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$$

donde $x_i \in \mathbb{R}^d$. El objetivo es encontrar una función π que separe perfectamente las clases y_i para todos los puntos de entrenamiento x_i y que a la vez sea tan "plana" como sea posible. En el caso más simple, dado un conjunto de vectores, se debe encontrar una función lineal π de la forma:

$$\pi(x) = \langle \omega, x \rangle + b$$

donde $\langle ., . \rangle$ es el producto interior o escalar. En este contexto "plana" significa que se debe encontrar un ω pequeño.

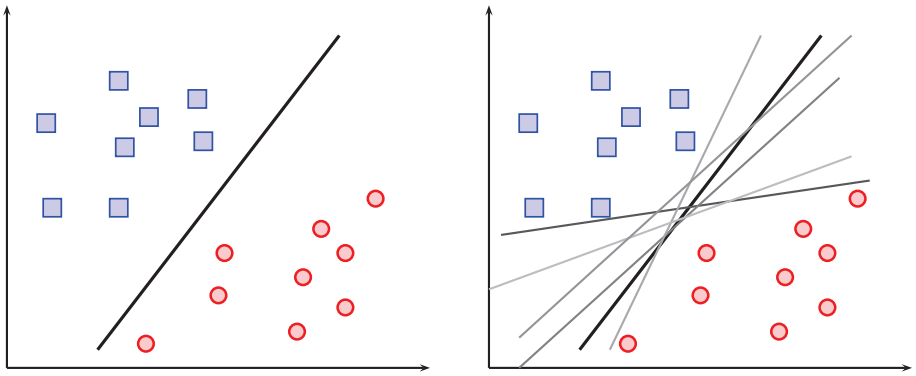


Figura 3.4: Hiperplanos de separación en un espacio bidimensional, de un conjunto de ejemplos separables en dos clases:(a) ejemplo de hiperplano de separación (b) otros ejemplos de hiperplanos de separación, de entre los infinitos posibles.

Fuente: Carmona Suárez, Tutorial sobre Máquinas de Vectores Soporte (SVM). Elaborado por: Autor.

Definición 3. *(Conjunto Separable). Sea* $S = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, $x_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$, $i = 1, \cdots, n$, S *se dice separable si existe algún hiperplano (hiperplano separable) en* $\mathbb{R}^d$ *que separe los vectores* $\mathbb{X} = \{x_1, \cdots, x_n\}$ *con valor* $y_i = +1$ *de aquellos con valor* $y_i = -1$.

Dado un conjunto separable de vectores de ensayo $S = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, donde $x_i \in \mathbb{R}^d$ e $y_i \in \{+1, -1\}$, se puede definir un hiperplano de separación (ver figura 3.4a) como una función lineal que es capaz de separar dicho conjunto sin error:

$$\pi(x) = (\omega_1 x_1 + \cdots + w_n x_n) + b = \langle \omega, x \rangle + b \tag{3.2}$$

donde ω y $\mathbf{b}$ son coeficientes reales. El hiperplano de separación cumplirá las siguientes restricciones para todo x_i del conjunto de ensayo:

$$\begin{cases} \langle \omega, x_i \rangle + b \geq 0 & \text{si } y_i = +1 \\ \langle \omega, x_i \rangle + b \leq 0 & \text{si } y_i = -1, i = 1, \cdots, n \end{cases} \tag{3.3}$$

que es equivalente a:

$$y_i \left(\langle \omega, x_i \rangle + b \right) \geq 0, \ i = 1, \cdots, n \tag{3.4}$$

de forma más compacta

$$y_i \pi(x_i) \geq 0, \ i = 1, \cdots, n \tag{3.5}$$

como se puede deducir fácilmente de la figura 3.4 (b), el hiperplano que permite separar las observaciones no es único, es decir, existen infinitos hiperplanos separables, representados por todos aquellos hiperplanos que son capaces de cumplir las restricciones impuestas por cualquiera de las expresiones equivalentes 3.3-3.5. Sin embargo, surge la pregunta sobre si es posible establecer algún criterio complementario que permita definir un hiperplano de separación óptimo. De esta manera, primero, se define el concepto de margen de un hiperplano de separación, denotado por τ, como la mínima distancia entre dicho hiperplano y el vector de ensayo más cercano de cualquiera de las dos clases (ver figura 3.5 (a)). A partir de esta definición, un hiperplano de separación se denominará óptimo si su margen es de tamaño máximo (figura 3.5 (b)).

La distancia entre un hiperplano de separación $\pi(x)$ y un vector de ensayo x' viene dada por

$$\frac{|\pi(x')|}{||\omega||} \tag{3.6}$$

siendo $|\cdot|$ el valor absoluto, $||\cdot||$ el operador norma de un vector, y ω el vector que, junto con el parámetro $\mathbf{b}$, define el hiperplano $\pi(x)$ y que, además, tiene la propiedad de ser perpendicular al hiperplano considerado. En consecuencia, de las expresiones (3.5) y (3.6), todos los vectores de entrenamiento cumplirán que:

$$\frac{y_i \pi(x_i)}{||\omega||} \geq \tau, \ i = 1, \cdots, n \tag{3.7}$$

de donde, se deduce que encontrar el hiperplano óptimo es equivalente a encontrar

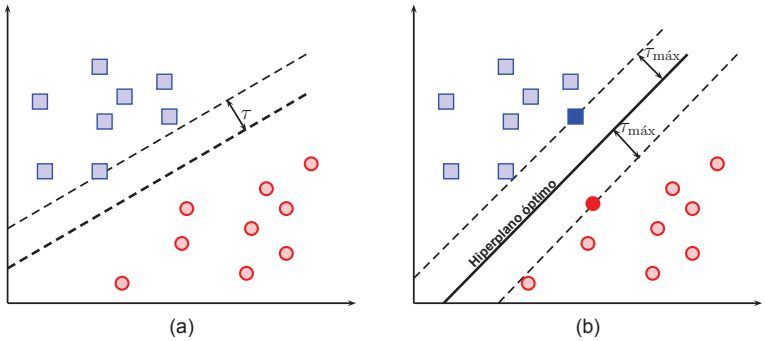


Figura 3.5: Margen de un hiperplano de separación: (a) hiperplano de separación no-óptimo y su margen asociado (no máximo), (b) hiperplano de separación óptimo y su margen asociado (máximo)

.

Fuente: Carmona Suárez, Tutorial sobre Máquinas de Vectores Soporte (SVM). Elaborado por: Autor.

el valor de ω que maximiza el margen. Sin embargo, existen infinitas soluciones que difieren solo en la escala de ω. Así, por ejemplo, todas las funciones lineales $\lambda(< \omega, x > +b)$, con $\lambda \in \mathbb{R}$, representan el mismo hiperplano. Para limitar, por tanto, el número de soluciones a una sola, y teniendo en cuenta que (3.7) se puede expresar también como:

$$y_i \pi(x_i) \geq \tau ||\omega||, \ i = 1, \cdots, n \tag{3.8}$$

la escala del producto de τ y la norma de ω se fija, de forma arbitraria, a la unidad, es decir

$$\tau ||\omega|| = 1 \tag{3.9}$$

por lo que se concluye que aumentar el margen es equivalente a disminuir la norma de ω, ya que la expresión anterior se puede expresar como

$$\tau = \frac{1}{||\omega||} \tag{3.10}$$

Por lo tanto, de acuerdo a su definición, un *hiperplano de separación óptimo* (ver figura 3.6) será aquel que posee un margen máximo y, por consiguiente, un valor mínimo de $||\omega||$ y, además, está sujeto a la restricción dada por (3.8), junto con el criterio expresado por (3.9), es decir,

$$y_i \pi(x_i) \geq 1, \ i = 1, \cdots, n \tag{3.11}$$

equivalente a:

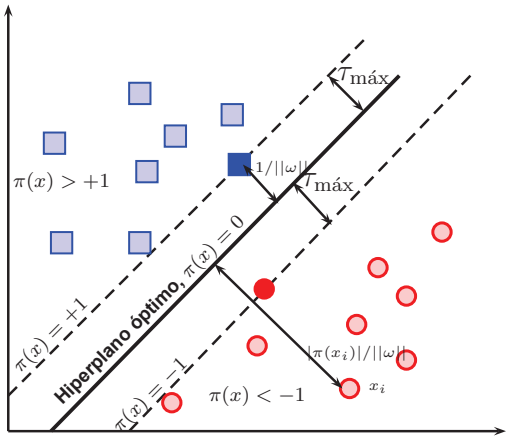


Figura 3.6: Hiperplano de separación óptimo.- La distancia de cualquier ejemplo, x_i, al hiperplano de separación óptimo viene dada por $|\pi(x_i)|/||\omega||$. En particular, si dicho ejemplo pertenece al conjunto de vectores soporte (identificados por siluetas sólidas), la distancia a dicho hiperplano será siempre $1/||\omega||$. Además, los vectores soporte aplicados a la función de decisión siempre cumplen que $|\pi(x)| = 1$.

Fuente: Carmona Suárez, Tutorial sobre Máquinas de Vectores Soporte (SVM). Elaborado por: Autor.

$$y_i(\langle \omega, x_i \rangle + b) \geq 1, \ i = 1, \cdots, n \tag{3.12}$$

La definición de *margen* máximo está directamente relacionado con la capacidad de generalización del hiperplano de separación, de manera que, a mayor margen, existirá mayor separación entre las dos clases. Los vectores de ensayo que están situados a ambos lados del hiperplano óptimo y que definen el margen o, lo que es lo mismo, aquellos para los que la restricción (3.12) es una igualdad, reciben el nombre de *vectores de soporte* (ver figura 3.6). Debido a que estos vectores son los más cercanos al hiperplano de separación, serán los más difíciles de clasificar y, por tanto, son los únicos vectores a considerar a la hora de construir dicho hiperplano.

La búsqueda del hiperplano óptimo para el caso separable puede ser formalizado como el problema de encontrar el valor de ω y $\mathbf{b}$ que minimiza el funcional $f(\omega) = ||\omega||$ sujeto a las restricciones (3.12), o de forma equivalente

$$\begin{cases} \text{mín } f(\omega) = \frac{1}{2}||\omega||^2 = \frac{1}{2}\langle \omega, \omega \rangle \\ \text{sujeto a:} \\ y_i(\langle \omega, x_i \rangle + \ b) \geq 1, \ i = 1, \cdots, n \end{cases} \tag{3.13}$$

Este problema de optimización con restricciones pertenece a un problema de programa-

ción cuadrático y es tratable mediante la teoría de la optimización [2]. De esta manera, se puede demostrar que el problema de optimización dado por (3.13) satisface el criterio de convexidad y, por tanto, tiene un dual. En estas condiciones, y aplicando los resultados descritos en el apéndice A. En el desarrollo de este proyecto se ha determinado un esquema para resolver el problema principal. Así, se deben seguir los siguientes pasos:

Primer paso, construir un problema de optimización sin restricciones utilizando la función Lagrangiana a minimizar:

$$L(\omega, b, \alpha) = \frac{1}{2} \langle \omega, \omega \rangle - \sum_{i=1}^{n} \alpha_i \left[y_i (\langle \omega, x_i \rangle + b) - 1 \right] \tag{3.14}$$

donde los $\alpha_i \geq 0$ [3] son los multiplicadores de Lagrange.

Segundo paso, aplicar las condiciones de *Karush-Kuhn-Tucker* (apéndice A.3.1), también conocidas como condiciones **KKT**:

$$\frac{\partial L(\omega, b, \alpha)}{\partial \omega} = \omega - \sum_{i=1}^{n} \alpha_i y_i x_i = 0, \ i = 1, \cdots, n \tag{3.15}$$

$$\frac{\partial L(\omega, b, \alpha)}{\partial b} = - \sum_{i=1}^{n} \alpha_i y_i = 0, \ i = 1, \cdots, n \tag{3.16}$$

$$\alpha_i [1 - y_i (\langle \omega, x_i \rangle) + b] = 0, \ i = 1, \cdots, n \tag{3.17}$$

Las restricciones representadas por (3.15-3.16) corresponden al resultado de aplicar la primera condición **KKT**, y las expresadas en (3.17), al resultado de aplicar la denominada condición complementaria (segunda condición **KKT**). Las primeras permiten expresar los parámetros de ω y **b** en términos de α_i:

[2] La teoría de optimización establece que un problema de optimización, denominado primal, tiene una forma dual si la función a optimizar y las restricciones son funciones estrictamente convexas. En estas circunstancias, resolver el problema dual permite obtener la solución del problema primal.

[3] El problema que se debe resolver es minimizar la función $L(\omega, b, \alpha_i)$ respecto a ω y b, y simultáneamente requerir que las derivadas parciales de L con respecto a los multiplicadores de Lagrange α_i sean todas nulas, todo ello sujeto al conjunto de restricciones $C_1 = \{\alpha_i \geq 0, i = 1, \cdots, n\}$; a este problema se le denomina problema primal. Así, el problema inicial queda como un problema de programación cuadrática donde la función objetivo es convexa, y los puntos que satisfacen las restricciones forman un conjunto convexo. Esto significa que se puede resolver el siguiente problema dual asociado al problema primal: maximizar la función $L(\omega, b, \alpha_i)$ respecto a las variables duales α_i sujeta a las restricciones impuestas para que los gradientes de L con respecto a ω y b sean nulos, y sujeta también al conjunto de restricciones $C_1 = \{\alpha_i \geq 0, i = 1, \cdots, n\}$. Esta particular formulación del problema dual se denomina problema dual de Wolfe, y verifica que el máximo de $L(\omega, b, \alpha_i)$ respecto de la variables duales, sujeto a las restricciones C_2, coincide con los mismos valores para ω, b y α_i, que el mínimo de $L(\omega, b, \alpha_i)$ respecto a ω y b sujeto a las restricciones de C_1, es decir, la solución al problema planteado es un punto silla de la función $L(\omega, b, \alpha_i)$.

$$\omega = \sum_{i=1}^{n} \alpha_i y_i x_i,\ i = 1, \cdots, n \tag{3.18}$$

y, además, establecen restricciones adicionales para los coeficientes α_i:

$$\sum_{i=1}^{n} \alpha_i y_i = 0,\ i = 1, \cdots, n \tag{3.19}$$

a partir de las nuevas relaciones obtenidas, se construirá el problema dual. De esta forma, bastará usar (3.18) para expresar la función Lagrangiana únicamente en función de α_i. Para lo cual, se puede reescribir (3.14) como

$$L(\omega, b, \alpha) = \frac{1}{2} \langle \omega, \omega \rangle - \sum_{i=1}^{n} \alpha_i y_i \langle \omega, x_i \rangle - b \sum_{i=1}^{n} \alpha_i y_i + \sum_{i=1}^{n} \alpha_i$$

Considerando que, según la condición (3.19), el tercer término de la expresión anterior es nulo, la substitución de (3.18) en dicha expresión resulta ser

$$L(\alpha) = \frac{1}{2} \left(\sum_{i=1}^{n} \alpha_i y_i x_i \right) \left(\sum_{j=1}^{n} \alpha_j y_j x_j \right) - \left(\sum_{i=1}^{n} \alpha_i y_i x_i \right) \left(\sum_{j=1}^{n} \alpha_j y_j x_j \right) + \sum_{i=1}^{n} \alpha_i$$

$$L(\alpha) = -\frac{1}{2} \left(\sum_{i=1}^{n} \alpha_i y_i x_i \right) \left(\sum_{j=1}^{n} \alpha_j y_j x_j \right) + \sum_{i=1}^{n} \alpha_i$$

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \tag{3.20}$$

En consecuencia, se ha transformado el problema de minimización (primal 3.13), en el problema dual, consistente en maximizar (3.20) sujeto a las restricciones (3.19), junto a las asociadas originalmente a los multiplicadores de Lagrange:

$$\begin{cases} \max L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\ \text{sujeto a:} \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \alpha_i \geq 0,\ i = 1, \cdots, n \end{cases} \tag{3.21}$$

Del mismo modo que el problema primal, este problema es abordable mediante técnicas estándar de programación cuadrática[4].

[4]Sin embargo, como se puede comprobar, el tamaño del problema de optimización dual escala con el número de muestras, n, mientras que el problema primal lo hace con la dimensionalidad, d. Por tanto, aquí radica la ventaja del problema dual, es decir, el costo computacional asociado a su resolución es

La solución del problema dual, α^*, permitirá obtener la solución del problema primal. Por lo tanto, bastará sustituir dicha solución en la expresión (3.18) y, finalmente, sustituir el resultado así obtenido en (3.2), es decir:

$$\pi(x) = \sum_{i=1}^{n} \alpha_i^* y_i \langle x, x_i \rangle + b^* \tag{3.22}$$

considerando las restricciones (3.17), resultantes de aplicar la segunda condición KKT, se puede afirmar que si $\alpha_i > 0$ entonces el segundo factor de la parte izquierda de dicha expresión deberá ser cero y, por tanto

$$y_i(\langle w^*, x_i \rangle + b^*) = 1 \tag{3.23}$$

es decir, el correspondiente vector de ensayo, $(x_i; y_i)$, satisface la correspondiente restricción del problema primal (3.13), pero considerando el caso "igual que". Por definición, los vectores de ensayo que satisfacen las restricciones expresadas en (3.13), considerando el caso "igual que", son los vectores soporte y, por tanto, se puede afirmar que sólo los vectores de ensayo que tengan asociado un $\alpha_i > 0$ serán vectores soporte. De este resultado, también puede afirmarse que el hiperplano de separación (3.22) se construirá como una combinación lineal de los vectores soporte del conjunto de ensayo, ya que el resto de vectores de ensayo tendrán asociado un $\alpha_j = 0$.

Para que la definición del hiperplano (3.22) sea completa, es preciso determinar el valor del parámetro b^*. Su valor se calcula despejando b^* de (3.23):

$$b^* = y_{vs} - \langle w^*, x_{vs} \rangle \tag{3.24}$$

donde (x_{vs}, y_{vs}) representa la tupla correspondiente al vector de soporte junto con su respectivo valor (etiqueta) de la clase a la cual corresponde. Finalmente, haciendo uso de (3.18) en (3.23), o en (3.24), permitirá también calcular el valor de b^* en función de la solución del problema dual [Carmona, 2014], [Vapnik, 1998].

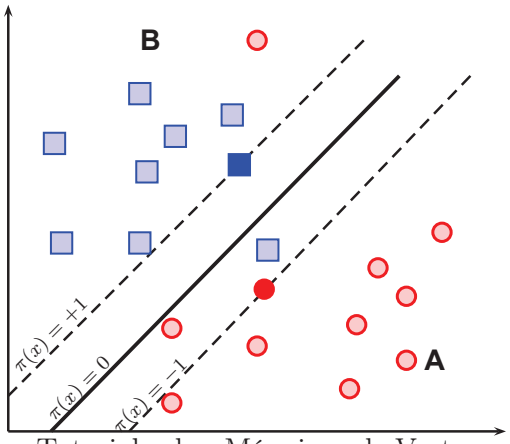
3.2.2 Caso linealmente no separable

El problema desarrollado en la sección anterior tiene escaso interés práctico porque los problemas reales se caracterizan normalmente por poseer observaciones ruidosas, es decir, no ser perfecta y linealmente separables.

Por ejemplo en la figura 3.7 se observa que existe un punto A dentro de la región correspondiente a los puntos B que nunca podrá ser separado de ellos por medio de

factible incluso para problemas con un número muy alto de dimensiones.

Figura 3.7: Caso no linealmente separable



Fuente: Carmona Suárez, Tutorial sobre Máquinas de Vectores Soporte (SVM). Elaborado por: Autor.

hiperplanos. En este caso se dice que es un conjunto no separable. Surge entonces la inquietud de si es posible extender la idea desarrollada en la sección anterior para encarar conjuntos no separables. La estrategia para tratar este tipo de problemas es relajar el grado de separabilidad del conjunto de ensayo, permitiendo que haya errores de clasificación en algunos de los vectores del conjunto de entrenamiento. Sin embargo, sigue siendo el objetivo, encontrar un hiperplano óptimo para el resto de casos que sí son separables.

Desde el punto de vista de la formulación desarrollada en la sección anterior, un vector de ensayo es no separable si no cumple la condición (3.12). Luego se pueden observar dos casos. En el primero caso, el vector de ensayo cae dentro del margen asociado a la clase correcta, de acuerdo a la frontera de decisión que define el hiperplano de separación. En el otro caso, el vector de ensayo cae al otro lado de dicho hiperplano. En ambos casos se dice que el vector de características es no separable, sin embargo en el primer caso es clasificado de forma correcta y, en el segundo, no lo es (ver figura 3.8).

La idea para afrontar este nuevo problema es introducir, en la condición (3.12) un costo adicional (es decir, un aumento en la función objetivo primal), un conjunto de variables reales positivas, denominadas variables de holgura, $\xi_i,\ i, \cdots, n$; esto permitirá cuantificar el número de vectores de ensayo no-separables que se está dispuesto a admitir, es decir:

$$y_i(\langle \omega, x_i \rangle) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \cdots, n \tag{3.25}$$

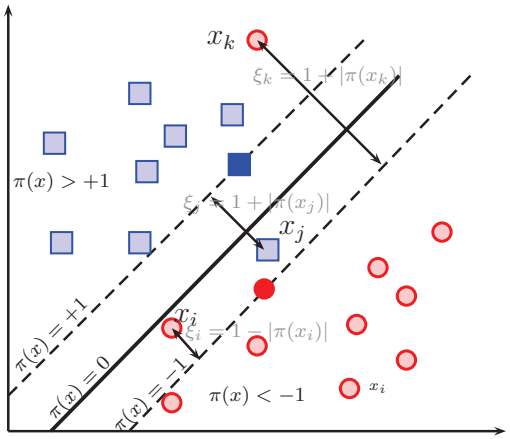


Figura 3.8: Casos no linealmente separable.- en el caso de ejemplos no-separables, las variables de holgura miden la desviación desde el borde del margen de la clase respectiva. Así, los puntos de entrada x_i, x_j y x_k son, cada uno de ellos, no-separables $(\xi_i, \xi_j, \xi_k) > 0$. Sin embargo, x_i está correctamente clasificado, mientras que x_j y x_k están en el lado incorrecto de la frontera de decisión y, por tanto, mal clasificados.

Fuente: Carmona Suárez, Tutorial sobre Máquinas de Vectores Soporte (SVM). Elaborado por: Autor.

En consecuencia, para un vector de ensayo (x_i, y_i), su variable de holgura asociada, ξ_i, representa la desviación del caso separable, medida desde el borde del margen que corresponde a la clase y_i (ver figura 3.8). De acuerdo a esta definición, las variables de holgura de valor cero corresponden a vectores de ensayo separables, mayores que cero corresponden a vectores de ensayo no separables y mayores que uno corresponden a puntos no separables y, además, mal clasificados. Además, la suma de todas las variables de holgura, $\sum_{i=1}^{n} \xi_i$, permite cuantificar el costo asociado al número de observaciones no separables . De esta forma, en una primera aproximación, cuanto mayor sea el valor de esta suma, mayor será el número de observaciones no separables.

Una vez relajadas las restricciones, según (3.25), entonces no basta con plantear como único objetivo maximizar el margen, ya que podríamos lograrlo a costa de clasificar erróneamente muchas observaciones. Por lo tanto, la función a optimizar debe incluir, de alguna forma, los errores de clasificación que está cometiendo el hiperplano de separación, es decir:

$$f(\omega, \xi) = \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{n} \xi_i \tag{3.26}$$

donde **C** es un parámetro (constante), suficientemente grande, elegida por el investigador, que permite controlar en qué grado influye el término del costo de observaciones

no separables en la minimización de la norma, es decir, permitirá regular el compromiso entre el grado de sobreajuste del clasificador final y la proporción del número de observaciones no separables. De manera que, si se toma un valor de **C** grande, permitiría valores de ξ_i muy pequeños. Por el contrario, un valor de **C** muy pequeño permitiría valores de ξ_i muy grandes, es decir, estaríamos admitiendo un número muy elevado de puntos mal clasificados.

Entonces, el nuevo problema de optimización consistirá en encontrar el hiperplano, definido por ω y **b**, que minimiza el funcional (3.26) y sujeto a las restricciones dadas por (3.25):

$$\begin{cases} \text{mín}\, \dfrac{1}{2}\langle \omega, \omega \rangle + C\sum_{i=1}^{n} \xi_i \\ \text{sujeto a:} \\ y_i(\langle \omega, x_i \rangle + b) + \xi_i - 1 \geq 0 \\ \xi_i \geq 0,\ i = 1, \cdots, n \end{cases} \tag{3.27}$$

El hiperplano así definido recibe el nombre de *hiperplano de separación de margen blando* (*soft margin*), en oposición al obtenido en el caso perfectamente separable, también conocido como *hiperplano de separación de margen duro* (*hard margin*). De manera similar al caso de la sección anterior, el problema de optimización a ser resuelto puede ser transformado a su forma dual. El procedimiento para obtener el hiperplano de separación es similar al caso perfectamente separable. Por tanto, en este caso, sólo se reproducirán de forma esquemática y secuencial los pasos necesarios para realizar dicha transformación.

Primer paso, obtención de la función Lagrangiana a minimizar:

$$L(\omega, b, \xi, \alpha, \beta) = \frac{1}{2}\langle \omega, \omega \rangle + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i [y_i(\langle \omega, x_i \rangle - b) + \xi_i - 1] - \sum_{i=1}^{n} \beta_i \xi_i \tag{3.28}$$

el siguiente paso es aplicar las condiciones KKT:

$$\frac{\partial L(\omega, b, \xi, \alpha, \beta)}{\partial \omega} = \omega - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \tag{3.29}$$

$$\frac{\partial L(\omega, b, \xi, \alpha, \beta)}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3.30}$$

$$\frac{\partial L(\omega, b, \xi, \alpha, \beta)}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \tag{3.31}$$

$$\alpha_i[1 - y_i(\langle \omega, x_i \rangle + b) - \xi_i] = 0,\ i = 1, \cdots, n \tag{3.32}$$

$$\beta_i \xi_i = 0,\ i = 1, \cdots, n \tag{3.33}$$

como paso final se deben, establecer las relaciones entre las variables del problema primal (ω, b, ξ) con las del problema dual (α, β). Para ello, hacemos uso de la relación (3.29):

$$\omega = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{3.34}$$

Establecer restricciones adicionales de las variables duales. Así se hace uso de las relaciones (3.30-3.31):

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3.35}$$

$$C = \alpha_i + \beta_i \tag{3.36}$$

Del resultado obtenido en (3.34), eliminar las variables primales de la función Lagrangiana para obtener el problema dual que queremos maximizar:

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \tag{3.37}$$

Finalmente, se obtiene la formalización buscada del problema dual:

$$\begin{cases} \text{máx} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\ \text{sujeto a:} \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C,\ i = 1, \cdots, n \end{cases} \tag{3.38}$$

De esta manera, la solución del problema dual nos permitirá expresar el *hiperplano de separación óptima* en términos de α^*. Por tanto, bastará tener en cuenta dicha solución y sustituir la expresión (3.34) en (3.2), es decir:

$$\pi(x) = \sum_{i=1}^{n} \alpha_i^* y_i \langle x, x_i \rangle + b^* \tag{3.39}$$

Para obtener el hiperplano final es necesario determinar el valor del parámetro b^* (sesgo), pero antes de obtener una expresión para el cálculo de b^*, se considera algunos resultados interesantes.

Considerando la restricción 3.36 se tiene que:

- Si $\alpha_i = 0$, entonces $C = \beta_i$, que conjuntamente con la restricción (3.33) se deduce que $\xi_i = 0$. Entonces, se puede afirmar que todos los puntos x_i cuyo α_i asociado sea igual cero corresponden a puntos separables ($\xi_i = 0$).

- Por otra parte, todo punto no separable, x_i, se caracteriza por tener asociado un $\xi_i > 0$ (ver figura 3.8). En este caso, y de la restricción (3.33), se deduce que $\beta_i = 0$. A su vez, de este último resultado y la restricción (3.36), se deduce que $\alpha_i = C$. Entonces, se puede afirmar que todos los puntos x_i cuyo $\alpha_i = C$ corresponderán a ejemplos no separables ($\xi_i > 0$).Además, dado que, en este caso, $\alpha_i \neq 0$, de la restricción (3.32) se deduce que

  $$1 - y_i(\langle \omega^*, x_i \rangle + b^*) - \xi_i = 0$$

  es decir

  $$1 - y_i \pi(x_i) = \xi_i$$

  Donde se pueden considerar dos casos (ver figura 3.8).

  - En el primer caso, el punto, x_i, aunque siendo no separable, está bien clasificado, es decir, $y_i \pi(x_i) \geq 0$, entonces $\xi_i = 1 - |\pi(x_i)|$
  - En el segundo caso, el punto, x_i, siendo no separable, está mal clasificado, es decir, $y_i \pi(x_i) < 0$, entonces $\xi_i = 1 + |\pi(x_i)|$

- Finalmente, se considera el caso: $0 < \alpha_i < C$, en esta situación, la restricción (3.36) permite afirmar que $\beta_i \neq 0$ y, de este resultado y la restricción (3.33), se deduce que $\xi_i = 0$. Igualmente, si $0 < \alpha_i < C$, de la restricción (3.32) y del resultado obtenido anteriormente ($\xi_i = 0$), se deduce que

  $$1 - y_i(\langle \omega^*, x_i \rangle + b^*) = 0$$

  Por tanto, se puede afirmar que un punto, x_i, es *vector soporte* si y solo si $0 < \alpha_i < C$.

De la expresión anterior, se está en disposición de calcular el valor b^*, es decir

$$b^* = y_i - \langle \omega^*, x_i \rangle \quad \forall_i \, t.q. \, 0 < \alpha_i < C \tag{3.40}$$

Obsérvese que, a diferencia del caso perfectamente separable, ahora, para el cálculo de b^*, no es suficiente con elegir cualquier vector de ensayo (observación) x_i que tenga asociado un $\alpha_i > 0$. Ahora, se habrá de elegir cualquier vector ensayo x_i que tenga asociado un α_i que cumpla la restricción $0 < \alpha_i < C$. Finalmente, haciendo uso de (3.29), es posible expresar b^* en términos de las variables duales:

$$b^* = y_i - \sum_{j=1}^{n} \alpha^* y_i \langle x_j, x_i \rangle \quad \forall_i \, t.q. \, 0 < \alpha_i < C \tag{3.41}$$

donde los coeficientes α_i^*, $i = 1, \cdots, n$, corresponden a la solución del problema dual.

En resumen, en el caso de vectores de ensayo (observaciones) no separables, hay dos tipos de vectores de ensayo para los que los $\alpha_i^* \neq 0$. Aquellos, para los que $0 < \alpha_i^* < C$, que corresponderían a vectores soporte "normales" y, aquellos para los que $\alpha_i^* = C$, asociados a casos no separables. Estos últimos reciben el nombre de *vectores soporte "acotados" (bounded support vectors)*. Ambos tipos de vectores (observaciones) intervienen en la construcción del hiperplano de separación. El problema dual del caso no separable (3.38) y el correspondiente al caso perfectamente separable, (3.21), son prácticamente iguales. La única diferencia radica en la inclusión de la constante C (función de pérdida) en las restricciones del primero. [Vapnik, 1998], [Carmona, 2014], [González, 2002], [Jiménez y Rengifo, 2010].

3.2.3 Máquinas no lineales de vectores soporte

En las primeras secciones de este capítulo se ha trabajado, tanto en el caso separable como el caso no separable, considerando conjuntos de funciones lineales en los parámetros[5], las mismas salvo el caso de las funciones constantes, constituyen el conjunto de funciones más simples posible. En esta sección se estudia el problema de generalizar los desarrollos anteriores para el caso de conjuntos de funciones no necesariamente lineales en los parámetros. La idea es utilizar eficientemente conjuntos de funciones base, no lineales, para construir espacios transformados de alta dimensionalidad y buscar hiperplanos de separación óptimos en los espacios así transformados; cada uno de estos espacios es llamado *espacio de características*, para diferenciar del espacio de entrada (inputs, espacio-x).

[5]La relación que liga los parámetros del modelo $\omega \in \mathbb{R}^d$ y $b \in \mathbb{R}$ se expresa en términos de sumas y diferencias.

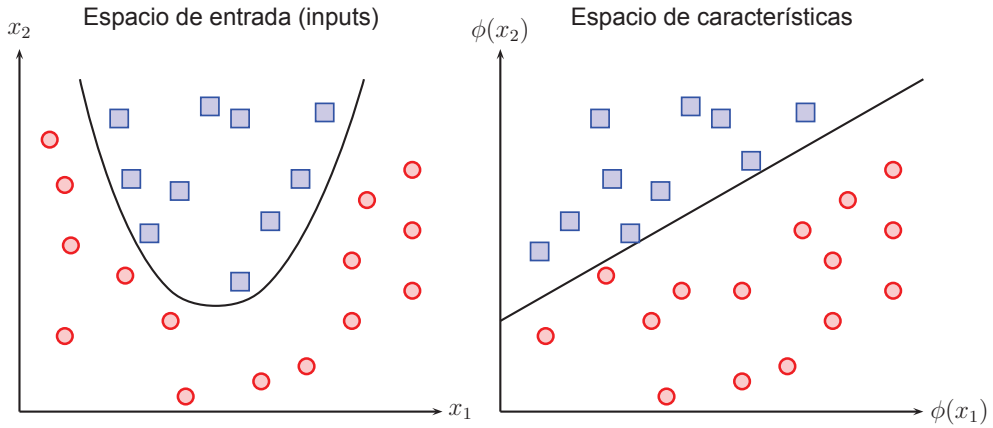


Figura 3.9: Máquinas no lineales de vectores soporte.- El problema de la búsqueda de una función de decisión no lineal en el espacio del conjunto de ejemplos original (espacio de entradas), se puede transformar en un nuevo problema consistente en la búsqueda de una función de decisión lineal (hiperplano) en un nuevo espacio transformado (espacio de características).

Fuente: Carmona Suárez, Tutorial sobre Máquinas de Vectores Soporte (SVM). Elaborado por: Autor.

Para conseguir tal fin, primeramente se observa que los vectores del conjunto de ensayo (inputs o entradas) forman parte de la solución (3.39) del problema de clasificación, a través de los productos escalares, $\langle x_i, x_j \rangle\ ,\ i, j = 1, \cdots, n$

Las ideas dadas por varios autores, permiten el siguiente desarrollo: Sea una aplicación (transformación), a la que se denota por Φ, del conjunto de entrada $\mathbb{X} \in \mathbb{R}^d$ en un espacio vectorial $\mathcal{F}$ (en la práctica se utilizan espacios de dimensión mucho mayor que d) dotado de un producto escalar:

$$\Phi : \mathbb{X} \in \mathbb{R}^d \rightarrow \mathcal{F} \tag{3.42}$$

que hace corresponder a cada vector del conjunto de ensayo x_i con un punto $\Phi(x_i)$ en el espacio $\mathcal{F}$. Entonces, en lugar de considerar el conjunto de vectores $\{x_1, \cdots, x_n\}$ se consideran los vectores transformados $\{\Phi(x_1), \cdots, \Phi(x_n)\}$ y así, si se plantea el problema de optimización original a estos vectores, es decir, se cambia de vectores de ensayo y de espacio de entrada, entonces se tiene que los nuevos vectores entran a formar parte de la solución del problema, solo a través del producto escalar definido en $\mathcal{F}$ como funciones de la forma $\langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{F}}$. Por tanto si se considera una función

$$k : \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{R}$$

a la que se denominará **función núcleo o kernel**, tal que

$$k(x_i, x_j) = \Phi(x_i) \bullet \Phi(x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle \,,$$

para resolver el algoritmo no se necesita conocer la forma explícita la aplicación Φ, solo es necesario conocer la función núcleo.

La idea entonces es construir un hiperplano de separación lineal en este nuevo espacio. La frontera de decisión lineal obtenida en el *espacio de características* se transformará en una frontera de decisión no lineal en el espacio original de entradas (ver figura 3.9). En este sentido, la función de decisión (3.2) en el espacio de características vendrá dada por [6]

$$\pi(x) = (w_1 \Phi(x_1) + \cdots + w_m \Phi(x_m)) = \langle \omega; \Phi(x) \rangle \tag{3.43}$$

y, en su forma dual, la función de decisión se obtiene transformando convenientemente la expresión de la frontera de decisión (3.39) en:

$$\pi(x) = \sum_{i=1}^{n} \alpha_i^* y_i k(x, x_i) \tag{3.44}$$

donde $K(x; x_i)$ es la *función kernel*. Por tanto, una función kernel puede sustituir convenientemente el producto escalar en (3.39). Así, dado el conjunto de funciones base, $\Phi(x) = \Phi(x_1), \cdots, \Phi(x_m)$, el problema a resolver en (3.44) sigue siendo encontrar el valor de los parámetros $\alpha_i^*,\ i = 1, \cdots, n$ que optimiza el problema dual (3.38), pero expresado ahora como:

$$\begin{cases} \text{máx} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \\ \text{sujeto a:} \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C,\ i = 1, \cdots, n \end{cases} \tag{3.45}$$

Así, resolver el problema, de la máquina de soporte vectorial, en un espacio de dimensión superior al del espacio de entrada, donde se trabaja con un conjunto de funciones lineales, la solución resultante es lineal en este espacio, pero no necesariamente es lineal en el espacio de entrada $\mathbb{X}$, de esta manera se está generalizando el problema básico a conjuntos de funciones no lineales. En la sección anterior se trabajó con funciones lineales de la forma:

[6] Obsérvese que se ha prescindido del termino **b** puesto que éste puede ser representado incluyendo en la base de funciones de transformación la función constante $\Phi(x) = 1$

$$\pi(x) = x \cdot \omega = \langle \omega, x \rangle + b$$

donde el vector solución ω estaba dado por $\omega = \sum_{i=1}^{n} \alpha_i y_i x_i$, en término de los vectores soporte $\omega = \sum_{i=1}^{N_{vs}} \alpha_i y_i s_i$. Si se lleva a cabo la transformación de los datos, el vector solución ω queda

$$\omega = \sum_{i=1}^{n} \alpha_i y_i \Phi(x_i)$$

y en términos de los vectores soporte

$$\omega = \sum_{i=1}^{N_{vs}} \alpha_i y_i \Phi(s_i) \tag{3.46}$$

donde por $\Phi(s_i)$ denotamos los vectores soporte del conjunto $\{\Phi(x_1), \cdots, \Phi(x_n)\}$, [Jiménez y Rengifo, 2010],[Carmona, 2014].

Claramente, se observa que la aplicación Φ aparece explícitamente en la solución del vector $\omega \in \mathcal{F}$ dada en (3.46), pero cuando sobre un nuevo vector de entrada x se realiza la fase de prueba, no es necesario tener identificada la transformación Φ ya que la solución viene dada por:

$$\pi(x) = \sum_{i=1}^{N_{vs}} \alpha_i y_i \langle \Phi(s_i), \Phi(x) \rangle + b \tag{3.47}$$

y escrita en términos de la función núcleo:

$$\pi(x) = \sum_{i=1}^{N_{vs}} \alpha_i y_i k(s_i, x) + b \tag{3.48}$$

3.3 Máquinas de Soporte Vectorial para la multiclasificación

En la sección preliminar, se ha considerado problemas de clasificación con solo dos clases (binaria). Sin embargo, comúnmente en los problemas reales es necesario discriminar entre más de dos clases (l>2), así es necesario hablar de problemas de clasificación multiclase. El paso a un problema de multiclasificación no es tan evidente como, a primera vista, puede parecer y se ha de elaborar una metodología precisa que nos permita resolver este problema de la forma más adecuada posible. Por ello, en esta sección se propone una máquina de vectores soporte para la multiclasificación que basada en la

máquina *l*-SVCR que proporciona una salida probabilística que resuelve el problema de asignación de etiqueta en caso de empate y nos proporciona un grado de confianza de la fiabilidad que un investigador deposita en el modelo.

3.3.1 Introducción

Sea Z un conjunto de vectores de entradas (inputs):

$$Z = \{x_1, \cdots, x_n\} \subset \mathcal{X} \subset \mathbb{R}^n$$

e $\mathcal{Y} = \{\theta_1, \theta_2, \cdots, \theta_l\}$ el conjunto de todas las posibles etiquetas, donde *l*>2 (para *l*=2 se tiene las SVMs expuestas en las secciones anteriores). Dentro del conjunto de entrenamiento $Z = \{(x_i, y_i)\}_{i=1}^n$ es útil realizar una partición a partir de los conjuntos

$$Z_k = \{(x_i, y_i), \text{ tales que } y_i = \theta_k\},$$

por lo que se tiene:

$$\bigcup_{k=1}^{l} Z_k = Z$$

y para cualquier $k \neq h$ se sigue $Z_k \cap Z_h = \emptyset$, como es fácil verificar. Sea n_k el número de vectores de entrenamiento del conjunto Z_k con lo que se tiene: $n = n_1 + n_2 + \cdots + n_l$; y por I_k el conjunto de índices i tales que $(x_i, y_i) \in Z_k$ de donde se tiene que $\bigcup_{i \in I_k} \{(x_i, y_i)\} = Z_k$.

La forma más habitual de utilización de las máquinas de vectores soporte para resolver problemas de multiclasificación, admite dos tipos de estructura:

- Máquinas biclasificadoras SV generalizadas: Construyen una función clasificadora global a partir de un conjunto de funciones clasificadoras dicotómicas (biclasificadoras).
- Máquinas multiclasificadoras SV: Construyen una función clasificadora global directamente considerando todas las clases a la vez.

3.3.2 Máquinas biclasificadoras generalizadas

Este tipo de máquinas da solución al problema de la multiclasificación transformando las *l* particiones del conjunto de entrenamiento en un conjunto de *L* biparticiones, en las cuales construye la correspondiente función discriminadora (es lo que se denomina **esquema de descomposición**) obteniendo $f_1, \cdots, f_L$ clasificadores dicotómicos o biclasificadores. A continuación, mediante un **esquema de reconstrucción**, realiza la

fusión de los biclasificadores f_i, $i = 1, \cdots, L$ con objeto de proporcionar como salida final, una de las *l* clases posibles, $\{\theta_1, \cdots, \theta_l\}$.

Dentro del esquema de descomposición, las máquinas más utilizadas son:

- Máquinas de vectores soporte: 1-v-r (*one-versus-rest*). Máquinas de vectores soporte, donde cada función clasificadora parcial f_i, enfrenta la clase θ_i contra el resto de las clases.
- Máquinas de vectores soporte: 1-v-1 (*one- versus- one*). Máquinas de vectores soporte, donde cada función clasificadora parcial f_{ij}, enfrenta la clase θ_i contra la clase θ_j, sin considerar las clases restantes.

Una vez construidas las *L* máquinas biclasificadoras, según estos métodos, se ha de determinar la respuesta global de la máquina frente a un nuevo input x. Para ello, se tiene en cuenta principalmente las *L* etiquetas proporcionadas (se incluye una nueva etiqueta $\theta_0 = \emptyset$, para aquellos casos en los que la máquina no selecciona una etiqueta concreta, contabilizando de esta forma todos los “No"votos en esta etiqueta artificial) por las funciones discriminadoras f_k, así como sus correspondientes salidas numéricas $f_k(x),\ k = 1, 2, \cdots, L$.

El método de reconstrucción más habitual, es el **esquema de votación**, donde se tiene en cuenta exclusivamente las etiquetas proporcionadas por las *L* máquinas biclasificadoras, $\{f_1, \cdots, f_L\}$. De esta forma, el esquema de reconstrucción parte de un conjunto formado por las etiquetas $\{\theta_k^*\}_{k=1}^L$ donde $\theta_k^* = \theta_i$ para algún $i = 0, 1, \cdots, l$. A partir de este conjunto realiza un recuento de todas las etiquetas (sin tener en cuenta, cuando aparezcan, la etiqueta $\theta_0 = \emptyset$):

Etiquetas	Votos
θ_1	m_1
$\vdots$	$\vdots$
θ_k	m_k
$\vdots$	$\vdots$
θ_l	m_l
	$L - m_0$

donde m_i, con $i = 1, \cdots, l$ es el número de veces que las máquinas f_k, $k = 1, \cdots, L$ asignan sus votos a la etiqueta θ_i; y m_0 es el número de veces que las máquinas f_k, $k = 1, \cdots, L$ no asignan ninguna etiqueta concreta.

Un esquema de reconstrucción posible, es la **votación por unanimidad**. En este esquema se toma como salida global de la máquina aquella etiqueta que haya obtenido todos

los votos posibles. A veces, es más adecuado considerar un esquema de **votación por mayoría absoluta** donde se toma como salida global de la máquina aquella etiqueta que haya obtenido más de la mitad de los votos posibles. Otra posibilidad, es considerar un esquema de **votación por mayoría simple** donde se toma como salida global de la máquina aquella etiqueta que haya obtenido más votos (la moda de la distribución de las etiquetas $\{\theta_k^*\}_{k=1}^L$).

En este último esquema de votación se puede presentar empates entre etiquetas. Este problema permite diferentes soluciones, según el tipo de arquitectura de la máquina, pero actualmente no existe una solución aceptada como válida por todos los investigadores [González, 2002].

Máquinas de vectores soporte 1-v-r

Esta aproximación del problema fue dada por Vapnik en [Vapnik, 1998]. Este tipo de máquina multiclasificadora construye *l* bi-clasificadores donde la función discriminante f_k, $k = 1, 2, \cdots, l$ discrimina los vectores de entrenamiento de la clase k, Z_k, del resto de vectores de las otras clases, $Z \backslash Z_k$, esto es, si el biclasificador f_k lleva a cabo la discriminación de las clases sin error, entonces $sign(f_k(x_i)) = 1$, si el vector $x_i \in Z_k$ y $sign(f_k(x_i)) = -1$, si el vector $x_i \in Z \backslash Z_k$.

De esta forma, dado una nueva entrada x, la salida numérica de la máquina $f_k(x)$ se interpreta de la siguiente forma:

$$\Theta(f_k(x)) = \begin{cases} \theta_k \; si \; sign(f_k(x)) = 1 \\ \theta_0 \; si \; sign(f_k(x)) = -1 \end{cases}$$

es decir, la función $\Theta(\cdot)$ etiqueta cada vector input x, en función del valor $f_k(x)$ dado por la máquina.

En este esquema, por construcción, ninguna etiqueta puede tener dos votos puesto que aparece explícitamente en una sola máquina. Si existe un único voto entonces la máquina global asignara aquella clase que haya obtenido este voto, puesto que en estas circunstancias todos los métodos de votación coinciden. El problema se planteará cuando haya más de una etiqueta con votos.

Las características de este multiclasificador son las siguientes:

- Se necesitan estimar *l* funciones biclasificadoras, es decir, se han de resolver *l* problemas SVM estándar.
- En la construcción de los biclasificadores intervienen todos los elementos del conjunto de ensayo *Z*, es decir, los biclasificadores disponen de toda la información

proporcionada por los datos.

- Es conocido, y está contrastado, que este procedimiento normalmente proporciona buenos resultados, [González, 2002].

Los dos principales inconvenientes que presentan este tipo de máquinas multiclasificadoras son:

- En caso de tener dos o más etiquetas empatadas en número de votos, no se encuentra dentro de la construcción un indicador que nos permita discriminar entre ellas. Podría pensarse en la utilización de las salidas numéricas $f_k(x)$, pero éstas no son adecuadas, por la propia naturaleza de las SVMs, puesto que la solución SVM se construye obligando a que la separación entre las dos clases sea la unidad.
- No es posible asignar un nivel de confianza a la salida global, a partir de los *l* biclasificadores, [González, 2002].

Máquinas de Vectores Soporte 1-v-1

En esta aproximación del problema de multiclasificación se construyen $L = \frac{l(l-1)}{2}$ biclasificadores donde la función discriminante f_{kh}, $1 \leq k < h \leq l$ discrimina los vectores de entrenamiento de la clase k, Z_k, de los vectores de entrenamiento de la clase h, Z_h, esto es, si el biclasificador f_{kh} lleva a cabo la discriminación de las clases sin error, entonces $sign(f_{kh}(x_i)) = 1$, si el vector $x_i \in Z_k$ y $sign(f_{kh}(x_i)) = -1$, si el vector $x_i \in Z_h$. Los restantes vectores de entrenamiento $Z \backslash \{Z_k \bigcup Z_h\}$ no se consideran en la construcción del problema de optimización.

De esta forma, dado un nuevo input x, la salida numérica de la máquina $f_{kh}(x)$ se interpreta de la siguiente forma:

$$\Theta(f_{kh}(x)) = \begin{cases} \theta_k \; si \; sign(f_{kh}(x)) = 1 \\ \theta_h \; si \; sign(f_{kh}(x)) = -1. \end{cases}$$

En esta construcción, el número máximo de votos que puede tener una determinada clase θ_k es el $l - 1$, ya que es el número de veces que aparece en la construcción de las funciones f_{kh}. De esta forma, si se utiliza el esquema de votación por unanimidad, se considera como salida global de la multiclasificación aquella etiqueta θ_k que haya obtenido $l - 1$ votos, que por construcción, si existe, debe ser única. Si no es posible aplicar este esquema, porque no existe tal etiqueta, se puede aplicar el esquema de votación por mayoría absoluta, en este caso se toma como salida global, la etiqueta θ_k tal que $m_k > m_h$ donde $h \neq k$ para todo h y $m_k > \frac{l-1}{2}$. Si se adopta el esquema de

votación por mayoría simple entonces se toma como salida global, la etiqueta θ_k tal que $m_k > m_h$ donde $h \neq k$ para todo h. Si finalmente, se presenta una situación con dos o más etiquetas empatadas en número de votos, habrá necesariamente que acudir a alguna otra característica procedente de la máquina 1-v-1, que nos permita elegir entre ellas.

Las características, más significativas, de este multiclasificador son las siguientes:

- Se necesita estimar $\frac{l(l-1)}{2}$ funciones biclasificadoras, es decir, es necesario entrenar $\frac{l(l-1)}{2}$ máquinas SV estándar, aunque con un conjunto de entrenamiento más reducido.
- Es posible asignar un nivel de confianza a la salida global, utilizando la interpretación probabilística de las SVMs.
- Es conocido que este procedimiento es, generalmente, preferido al esquema 1-v-r como así lo demuestran diferentes estudios empíricos.

Los dos principales inconvenientes que presentan este multiclasificador son:

- Cada uno de los biclasificadores es entrenado con datos extraídos de solo dos clases del conjunto de entrenamiento por lo que la varianza es mayor y no proporciona información sobre el resto de clases. Además, cada máquina f_{kh} entrenada, no utiliza la información disponible en los datos que quedan fuera de las etiquetas θ_k y θ_h, lo que supone una preocupante pérdida de información.
- El número de clasificadores, en comparación con las máquinas 1-v-r es alto, si el número de etiquetas *l* es grande. Por ejemplo, si $l = 6$ se tendría que entrenar 15 máquinas, [González, 2002].

3.3.3 Máquinas multiclasificadoras

Dentro de las máquinas de vectores soporte, también, es posible obtener de forma directa un multiclasificador, incorporando todas las etiquetas directamente en la configuración de un único problema de optimización.

Una primera aproximación se da en [Vapnik, 1998]. Siguiendo la notación dada en la anterior referencia, se denota los vectores de entrenamiento por

$$x_1^1, \cdots, x_{n_1}^1, \cdots, x_1^l, \cdots, x_{n_l}^l$$

donde el superíndice k en x_i^k denota que el vector pertenece a la clase k. Se considera el conjunto de funciones lineales

$$f_k(x) = \omega_k \cdot x + b_k, \quad k = 1, \cdots, l$$

El objetivo es construir l funciones (obtener l pares (ω_k, b_k) de parámetros) tales que para cada vector de entrada x, el clasificador

$$m = arg \max_{k=1,\cdots,l} \{\omega^k \cdot x + b_k\}$$

discrimine adecuadamente todos los vectores de entrenamiento sin error. Esto es, que las desigualdades

$$\omega^k \cdot x_i + b_k - \omega^m \cdot x_i - b_m \geq 1$$

sean ciertas para todo $k = 1, \cdots, l,\ m \neq k\ e\ i = 1, \cdots, n$. Si existen soluciones a este problema, se elige entre ellas, el par de parámetros (ω^k, b_k), $k = 1, \cdots, l$ para el cual, el funcional

$$\sum_{k=1}^{l} ||\omega^k||^2 = \sum_{k=1}^{l} \omega^k \cdot \omega^k$$

sea mínimo. Si, por el contrario, el conjunto de entrenamiento no puede ser discriminado sin provocar error en la clasificación, entonces el objetivo es minimizar el funcional

$$\sum_{k=1}^{l} ||\omega^k||^2 + C \sum_{k=1}^{l} \sum_{i=1}^{n} \xi_i^k$$

sujeto a las restricciones

$$\omega^k \cdot x_i + b_k - \omega^m \cdot x_i - b_m \geq 1 - \xi_i^k,$$

para $i = 1, \cdots, n$ y $1 \leq k, m \leq l$. Para resolver este problema de optimización se usa las mismas técnicas de optimización que en el caso de las SVMs con dos clases y se obtiene que:

1. La función $f_k(x)$ presenta el siguiente desarrollo en términos de los vectores soporte:

$$f_k(x) = \sum_{m \neq k} \sum^{n_k} i = 1 \alpha_i(k, m) x \cdot x_i^k - \sum_{m \neq k} \sum^{n_m} j = 1 \alpha_i(m, k) x \cdot x_i^m + b_k$$

2. Los coeficientes $\alpha_i(k, m)$, $k = 1, \cdots, m$, $m \neq k$, $i = 1, \cdots, n_k$, $j = 1, \cdots, n_m$ de este desarrollo tienen que maximizar la forma cuadrática:

$$
\begin{aligned}
W(\alpha) \;=\; & \sum_{k=1}^{l}\sum_{m\neq k}\Bigg[\sum_{i=1}^{n_k}\alpha_i(k,m) \\
& -\frac{1}{2}\sum_{m^*\neq k}\Bigg(\sum_{i,j=1}^{n_k}\alpha_i(k,m^*)\alpha_j(k,m)(x_i^k\cdot x_j^k) \\
& +\sum_{i=1}^{n_m}\sum_{i=1}^{n_m^*}\alpha_i(m,k)\alpha_j(m^*,k)(x_i^m,x^{m_j^*}) \\
& -2\sum_{i=1}^{n_k}\sum_{i=1}^{n_m}\alpha_i(k,m^*)\alpha_j(m,k)(x_i^k\cdot x_j^k)\Bigg)\Bigg]
\end{aligned}
$$

sujetos a las restricciones

$$0 \leq \sum_{m\neq k}\alpha_i(k,m) \leq C,$$

$$\sum_{m\neq k}\sum_{i=1}^{n_k}\alpha_i(k,m) = \sum_{m\neq k}\sum_{j=1}^{n_m}\alpha_j(m,k),$$

$$k = 1,\cdots,l$$

Así, se tiene que para $l > 2$ se han de estimar simultáneamente $n(l-1)$ parámetros $\alpha_i(k,m)$, con $i = 1,\cdots,n_k$, $m \neq k$, $k = 1,\cdots,l$, donde $n = \sum_{k=1}^{l} n_k$.

Como en el caso del biclasificador SVM, para construir la máquina de vectores soporte no lineal basta con sustituir el producto escalar $x_i^r \cdot x_j^s$ por una función núcleo $k(x_i^r, x_j^s)$ en las ecuaciones correspondientes. Otra aproximación a estos problemas de multiclasificación, aparece en [Weston and Watkins, 1998], donde según los autores, es una aproximación más natural que las dadas por los esquemas de descomposición y reconstrucción, seguidos con los biclasificadores, ya que se considera todas las clases a la vez. Esta aproximación sigue un camino muy similar al dado en [Vapnik, 1998] donde el problema de optimización consiste en minimizar la función

$$\frac{1}{2}\sum_{k=1}^{l}||\omega^k||^2 + C\cdot\sum_{i=1}^{n}\sum_{m\neq k}\xi_i^m$$

sujeto a las siguientes restricciones:

$$\omega^k\cdot x_i + b_k \geq \omega^m\cdot x_i + b_m + 2 - \xi_i^m,$$

$$\xi_i^m \geq 0, \quad i = 1, \cdots, n, \quad m \in \{1, 2, \cdots, l\} \backslash k.$$

La función de decisión viene dada por:

$$f(x) = arg\, \max_k (\omega^k \cdot x_i + b_k), \qquad k = 1, \cdots, l.$$

La resolución del problema de optimización proporciona la siguiente solución:

$$f(x) = arg\, \max_k \left(\sum_{i:y_i=k} A_i (x_i \cdot x) - \sum_{i:y_i \neq k} \alpha_i^n (x_i \cdot x) + b_k \right),$$

donde

$$A_i = \sum_{k=1}^{l} \alpha_i^k, \quad c_i^k = \begin{cases} 1 \; si \; y_i = k \\ 0 \; si \; y_i \neq k \end{cases}, \quad \sum_{i=1}^{n} \alpha_i^k = \sum_{i=1}^{n} c_i^k A_i, \quad k = 1, \cdots, l$$

y

$$0 \leq \alpha_i^k \leq C, \quad \alpha_i^k = 0, \quad i = 1, \cdots, n, \quad m \in \{1, 2, \cdots, l\} \backslash k.$$

Claramente, como en todas las máquinas de vectores soporte, si se reemplaza el producto escalar $(x_i \cdot x_j)$ por una función núcleo $k(x_i, x_j)$, se tiene una máquina no lineal. Como se apunta y comprueba empíricamente en [Weston and Watkins, 1998], esta máquina de multiclasificación, tiene resultados similares, en términos de porcentaje de errores, a las máquinas (1-v-1) y (1-v-r).
Las características de estos dos multiclasificadores, que incorporan todas las clases a la vez, son las siguientes:

- Se necesita estimar una única función multiclasificadora, pero sobre un problema de optimización que resulta mucho más complejo que en los problemas de biclasificación.
- La salida que resulta de la máquina no necesita ser interpretada en términos de un esquema de votación, y por tanto, el investigador no tiene que establecer un método de votación a priori, ya que va recogido dentro de la configuración de la máquina.

Los dos principales inconvenientes que presentan estos multiclasificadores son:

- Pensamos que el mayor inconveniente que presentan estas configuraciones de “todas las clases a la vez”, es él de ser una caja negra, en el sentido de no poder, evaluar a través de una salida intermedia, cómo se ha llegado a la salida última y medir la bondad de la misma.

- No es posible asignar un nivel de confianza a la salida global proporcionada por la máquina.

Veamos más detenidamente el inconveniente dado anteriormente en primer lugar. Si se nos presenta una nueva entrada x cuyo etiquetado, en caso de realizarse mal, pueda traer desagradables consecuencias, lo correcto será estudiar en profundidad cómo el multiclasificador ha llegado al etiquetado final, con objeto de corregir o tener en cuenta donde se ha producido el error. Esta posibilidad de trabajo es posible realizarla con las máquinas SV de multiclasificación siguiendo un esquema 1-v-1, o un esquema 1-v-r, sin más que exigir a la máquina que nos presente los resultados intermedios.

Por todo ello, consideramos más adecuado trabajar con máquinas multiclasificadoras que siguen un esquema de descomposición y reconstrucción, puesto que con éstas, podemos obtener como salidas los resultados de todas y cada una de las máquinas implementadas, y de esta forma disponer de un conjunto de resultados que nos permita tener una mayor capacidad de evaluación de la funcionalidad global.

Por otro lado, como empíricamente se ha demostrado que las máquinas 1-v-1 proporcionan mejores resultados que las máquinas 1-v-r, nosotros optamos por aquellas. Sin embargo, hemos apuntado anteriormente una serie de inconvenientes que sería adecuado paliar en lo posible, modificando la configuración inicial de estas máquinas, antes de su utilización.[González, 2002].

3.4 Funciones Núcleos

El objetivo de esta sección es generalizar los problemas de optimización de las máquinas de soporte vectorial sobre clases de funciones no necesariamente lineales. Como ya se había comentado anteriormente, para realizar esta generalización al caso no lineal, se ha de definir una aplicación real de dos variables con ciertas características determinadas, a la que se denominará *función núcleo*. De entre todas, se destaca que la principal característica de este tipo de función es que debe ser expresada a través de un producto escalar de una transformación de los vectores de entrada $\mathcal{X}$ en un espacio característico de dimensión superior, $\mathcal{H}$.
En las diferentes máquinas de soporte vectorial, que se ha estudiado, se resolvía un problema de optimización cuadrática sujeto a un conjunto de restricciones lineales. Así, en el problema de máquina biclasificadora estándar:

$$\min_{\omega\in\mathbb{R}^d} \frac{1}{2}||\omega||^2 + C\cdot\sum_{i=1}^{n}\xi_i \text{sujeto a: } \begin{cases} y_i\cdot(x_i\cdot\omega+b)-1+\xi_i\geq 0, \quad \forall_i \\ \xi_i\geq 0, \quad \forall_i \end{cases} \tag{3.49}$$

el objetivo es encontrar un hiperplano discriminador (la solución) dentro del siguiente conjunto de funciones lineales (se busca un hiperplano separador en $\mathbb{R}^d$):

$$\mathcal{F} = \{f : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R} / f(x) = \langle \omega, x \rangle + b, \text{ donde} \omega \in \mathbb{R}^d, b \in \mathbb{R}\} \tag{3.50}$$

donde cada una de las funciones queda determinada a partir de dos parámetros $\omega \in \mathbb{R}^d$ y $b \in \mathbb{R}$. El parámetro ω se expresa en la forma (ver ecuación (3.25)):

$$\omega = \sum_{i=1}^{N_{SV}} \alpha_i y_i s_i = \sum_{i=1}^{n} \beta_i x_i$$

donde, como se indicó anteriormente por N_{SV} se denota el número de vectores de soporte y por s_i los vectores de soporte del conjunto de ensayo $\{x_1, \cdots, x_n\}$. Además, se recuerda que los β_i son nulos salvo en los vectores de soporte y se verifica que

$$\sum_{i=1}^{n} \beta_i = 0$$

Por otro lado, se tiene que el parámetro b se determina a partir de las condiciones de Karush-Kuhn-Tucker, resolviendo para cada vector de soporte una ecuación lineal y tomando como valor b un valor promedio de todas esas soluciones. Por tanto, la solución lineal al problema de optimización se expresa a través de una función, que puede escribirse como sigue por la linealidad del producto escalar:

$$f(x) = \left\langle \sum_{i=1}^{n} \beta_i x_i, x \right\rangle + b = \sum_{i=1}^{n} \beta_i \langle x_i, x \rangle + b.$$

Así, en la representación de la función de decisión (función discriminadora), solo se necesita el producto escalar de los vectores de entrenamiento $\{x_1, x_2, \cdots, x_n\}$ con el nuevo vector de entrada $x \in \mathcal{X}$.

Como ya se indicó, lo más importante de esta expresión de la solución es que en ella intervienen los vectores de entrada exclusivamente a través del producto escalar, lo cual permite utilizar una técnica introducida por primera vez en [ABR64] que consigue generalizar el problema (3.49) a conjuntos de funciones no lineales.

Sea un conjunto de vectores de entrenamiento:

$$Z = \{(x_1, y_1), \cdots, (x_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y}$$

siguiendo las ideas expuestas en las secciones anteriores, lo práctico sería encontrar un hiperplano separador óptimo en un determinado espacio tal que en mencionado espacio el conjunto de entrenamiento fuese separable, con lo que no se tendría ninguna pérdida en la función objetivo.

Por ello, desde el punto de vista del problema de optimización, sería muy conveniente dado un conjunto de vectores de entrenamiento Z, no necesariamente separable, realizar una transformación ϕ de los vectores inputs $\mathcal{X} = \{x_1, \cdots, x_n\}$ tal que convierta el conjunto transformado,

$$Z_\phi = \{(\phi(x_1), y_1), \cdots, (\phi(x_n), y_n), \} \tag{3.51}$$

en un conjunto separable, en un espacio adecuado. Aun cuando este problema no siempre tendrá solución, si permitirá ampliar el campo de trabajo al poder considerar otras clases de funciones clasificadoras distintas de las lineales. Sin embargo, la imposibilidad de poder obtener siempre conjuntos separables se ha de tener en cuenta en el problema de la generalización. El hiperplano solución no puede ser cualesquiera, hay que buscar una relación entre suavidad y ajuste que necesariamente lleva, en los problemas prácticos, a tener que plantear el problema con alguna función de pérdida.

Formalmente, el problema se plantea como sigue: dado el espacio de los vectores de entrada $\mathcal{X}$ se considera una transformación Φ de este espacio en un espacio vectorial, que se denotará por $\mathcal{H}$ y llamará espacio característico, en la forma:

$$\Phi : \mathcal{X} \subset \mathbb{R}^d \to \mathcal{H} \subset \mathbb{R}^{d'} \tag{3.52}$$

donde normalmente la dimensión de $\mathcal{H}(d')$ es muy superior a la dimensión del espacio $\mathcal{X}$ $(d << d')$. A partir de esta transformación Φ, en lugar de la clase de funciones dada en (3.50), se considera la clase

$$\mathcal{F}_\phi = \{f : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R} / f(x) = \langle \omega, \phi(x) \rangle_{\mathcal{H}} + b, \text{ donde } \omega \in \mathbb{R}^{d'}, b \in \mathbb{R}\} \tag{3.53}$$

donde $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denota un producto escalar definido en $\mathcal{H}$.

Planteando el problema de optimización de la SVM a los vectores transformados (3.51), se obtiene un problema de optimización que proporciona, según la clase de funciones dada en (3.53), una solución lineal en el espacio característico, pero no necesariamente lineal en el espacio de las entradas.

Para una mayor simplicidad en los desarrollos y con intención de hacerlos más operativos, se utiliza la siguiente notación para el producto escalar en $\mathcal{H}$:

$$\langle \phi(x), \phi(x') \rangle = k(x, x')$$

y, si se reemplaza en todos los desarrollos del problema (3.49), $\langle x, x' \rangle$ por $k(x, x')$, se obtiene el siguiente problema de optimización convexa no lineal:

$$\min_{\omega\in\mathbb{R}^{d'}} \frac{1}{2}||\omega||^2 + C\cdot\sum_{i=1}^{n}\xi_i \text{sujeto a: } \begin{cases} y_i\cdot(\langle\omega\cdot\phi(x_i)\rangle - b) - 1 + \xi_i \geq 0, \quad \forall_i \\ \xi_i \geq 0, \quad \forall_i \end{cases} \tag{3.54}$$

cuya solución viene dada por:

$$\omega = \sum_{i=1}^{n}\beta_i\phi(x_i)$$

y proporciona la siguiente función en $\mathcal{F}_\phi$:

$$f(x) = \sum_{i=1}^{n}\beta_i k(x, x_i) + b. \tag{3.55}$$

sin más que cambiar el producto escalar $\langle x_i, x\rangle$ por $k(x_i, x)$ para $i = 1, \cdots, n$.

De esta forma se tiene una función $k(x_i, x)$ que juega un papel muy importante no solo en las máquinas de soporte vectorial, sino en toda la teoría del aprendizaje estadístico, y que es objeto de muchas investigaciones en la actualidad. Así, se da la siguiente definición:

Definición 4. *(**de núcleo de Mercer**) Una función núcleo es una función real de dos variables, denotada por k, que verifica:*

$$k : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$$

$$(x, x') \mapsto k(x, x') = \langle\phi(x), \phi(x')\rangle_{\mathcal{H}}$$

donde ϕ es una aplicación definida como en (3.52).

Por todo lo anteriormente expuesto, se sigue que la solución al problema de optimización (3.54) se expresa en términos de la función núcleo $k(\cdot,\cdot)$, sin tener en ningún momento que conocer la transformación ϕ, ya que ésta aparece implícitamente en (3.55) a través del producto escalar definido en $\mathcal{H}$. Por tanto, si la función núcleo $k(x, x')$ es fácil de evaluar, parece razonable utilizar k sin tener que utilizar para nada la aplicación ϕ, ya que entre otras razones:

- suele ser difícil de tratar con ella,
- no siempre es posible determinarla de forma explícita.

Un resumen de lo visto hasta ahora se puede observar en la figura 6.1. En esta figura, se observa que la idea inicial, en las máquinas de vectores soporte, es transformar los vectores de entrada $\mathcal{X} = \{x_1, \cdots, x_n\}$en unos nuevos vectores de entrada $\phi(\mathcal{X}) =$

$\{\phi(x_1), \cdots, \phi(x_n)\}$ dentro de un espacio característico $\mathcal{H}$ de dimensión muy superior, siguiendo una transformación no lineal elegida a priori. De esta forma se consigue tener un mayor grado de libertad para poder actuar sobre los datos. A continuación y dentro de este nuevo espacio, se busca el hiperplano separador óptimo (función discriminadora, función de clasificación, función de decisión) como un desarrollo lineal de funciones núcleos donde una de las dos componentes es un vector de entrada del conjunto de entrenamiento. Claramente, por definirse a partir de un producto escalar, una función núcleo debe verificar necesariamente que:

1. (Simétrica) $k(x, z) = k(z, x)$, para todo $x, z \in \mathcal{X}$.
2. (Desigualdad de Cauchy-Schwarz) $|k(x, z)|^2 \leq k(x, x)^2 \cdot k(z, z)^2$, para todo $x, z \in \mathcal{X}$.

Como ya se ha indicado, este nuevo enfoque está dirigido principalmente a problemas donde la dimensión del espacio de entrada es grande, es por ello por lo que se debe destacar que, a pesar de ser el espacio característico de una dimensión muy alta, la ejecución de las SVMs no depende de esta dimensión. Por todo ello la situación planteada, en el ejemplo anterior, no es desesperante ya que una de las características más interesante de utilizar las funciones núcleos está en que se puede tratar implícitamente con espacios $\mathcal{H}$ de dimensión arbitrariamente grande sin tener que calcular la transformación ϕ explícitamente, y por tanto no tener que manejar vectores de dimensión tan alta. [Aizerman et al., 1964][González, 2000][Scholkopf et al., 1997].

Capítulo 4

PROGRAMACIÓN LINEAL MULTICRITERO

La programación lineal (LP) es uno de los enfoques más clásicos y ampliamente utilizados en la solución de los problemas de optimización y, específicamente, en el problema de clasificación de dos grupos. Freed y Glover, [Freed and Glover, 1981], introdujeron una serie de enfoques de clasificación basados en la programación lineal para reducir la clasificación errónea en la separación de datos a través de dos objetivos en un sistema lineal. El objetivo más específico es el de maximizar las distancias mínimas (MMD) entre las observaciones y el valor crítico (discriminador). Por otra parte, el segundo objetivo es el de minimizar la suma de las desviaciones (MSD) entre las observaciones y el valor crítico. Dentro de la teoría de la programación lineal, los objetivos expuestos anteriormente no se pueden alcanzarse simultáneamente. Entonces, Shi propone un modelo de Programación Lineal de Múltiples Criterios (MCLP) [Shi et al., 2001], [Shi, 2001] que mejora los resultados minimizando la suma de las desviaciones externas y maximizando la suma de las desviaciones internas, simultáneamente. Estos modelos de clasificación basados en la programación lineal son muy poderosos cuando los datos son linealmente separables. Sin embargo, en el mundo real es más probable que los datos se presenten de manera linealmente no separables. Frente a esta situación, los modelos de clasificación basados en programación lineal (LP) no son soluciones de gran alcance o no aplicables. En consecuencia, algunos modelos entre ellos modelos no lineales [Kou et al., 2004], son propuestos para mejorar el poder de clasificación cuando se manipulan conjuntos de datos que no son linealmente separables. En esencia, ni los modelos lineales ni los modelos no lineales consideran la interacción entre dos o más atributos arbitrarios en un conjunto de datos para la clasificación, porque se supone que las contribuciones de todos los atributos hacia la clasificación son la suma de las contribuciones de cada atributo individual.

4.1 Programación Lineal

El propósito de esta sección es mostrar cómo se utiliza la programación lineal (PL) para resolver problemas de clasificación. Específicamente, se mostrará cómo formular el problema de encontrar un discriminador lineal como un problema de programación lineal.

El problema inicial que se quiere abordar es el de discriminar o clasificar entre los elementos de dos conjuntos con sus respectivas características, a manera de un ejemplo sencillo, se considera la tarea de asignar un solicitante de crédito a una clasificación de riesgo. Un cliente puede ser clasificado dentro de las categorías de riesgo como: *cliente bueno, cliente regular o cliente malo* de acuerdo a los datos requeridos en una aplicación estándar de crédito.

El problema puede ser descrito formalmente (matemáticamente) como: *Dados los puntos* $\mathbf{x}_i$ *y los conjuntos* $\mathbf{G}_j$*, la tarea consiste en encontrar una transformación lineal* ω*, y los límites apropiados (subdivisiones del intervalo)* b_j^L *y* b_j^U*, para caracterizar "correctamente" cada* $\mathbf{x}_i$ *(Los límites* b_j^L *y* b_j^U *representan respectivamente los límites inferior y superior para los puntos asignados al grupo* j*). Por tanto la tarea es determinar un predictor lineal o sistema de ponderación* ω *y puntos de corte* b_j^L *y* b_j^U*, tal que*

$$b_j^L \leq \mathbf{x}_k\omega \leq b_j^U \Leftrightarrow \mathbf{x}_k \in \mathbf{G}_j$$

y

$$b_1^L < b_1^U < b_2^L < b_2^U < \cdots < b_g^U.$$

los puntos $\mathbf{x}_i$ pueden, por supuesto, estar distribuidos de tal manera que la diferenciación completa de grupos sea imposible (por ejemplo, cuando los atributos de algunos solicitantes de crédito desafían la clasificación por categoría de riesgo).

Por lo tanto, es importante dotar al sistema de ponderación con el poder de establecer la diferenciación de los grupos anteriores, con una excepción mínima. Dos formulaciones de Programación Lineal útiles y directas para lograr tal objetivo se sugieren a continuación:

Alternativa 1. Determinar un predictor ω tal que:

$$\mathbf{x}_i\omega \geq b_j^L, \qquad \mathbf{x}_i\omega \leq b_j^U$$

para todo $\mathbf{x}_i \in \mathbf{G}_j$ y, además se impone como restricciones metas:

$$b_j^U < b_{j+1}^L + \xi_j \qquad para\ j = 1, \cdots, g-1,$$

donde g es el número de grupos asignados, el objetivo es

$$\text{mín} \sum c_j \xi_j$$

Alternativa 2. Esta formulación debe imponer los límites de separación como una restricción común, definiendo como una meta la inclusión de observaciones con límites apropiados. De esta manera;

$$b_j^U \leq b_{j+1}^L, \quad para\ j = 1, \cdots, g-1,$$

donde g es el número de grupos designados.

$$\mathbf{x}_i\omega \geq b_j^L - \xi_j, \qquad \mathbf{x}_i\omega \geq b_j^U + \xi_j$$

para todo $\mathbf{x}_i \in \mathbf{G}_j$, con el objetivo

$$\text{mín} \sum \xi_j \alpha_j.$$

En adelante se generaliza la idea anterior para la discriminación entre dos grupos, como se detalla a continuación:

Dados dos grupos, $\mathbf{G}_1$ y $\mathbf{G}_2$, determinar un vector apropiado ω y un valor límite b tal que, en la medida que sea posible,

$$\mathbf{x}_i\omega \leq b, \quad \mathbf{x}_i \in \mathbf{G}_1, \qquad \mathbf{x}_i\omega \geq b, \quad \mathbf{x}_i \in \mathbf{G}_2, \tag{4.1}$$

Se introduce ξ_i para medir el grado, para el cual los miembros del grupo $\mathbf{x}_i$ incumplen el límite crítico (la frontera entre los dos grupos), además se debe garantizar una solución en la cual:

$$\mathbf{x}_i\omega \leq b + \xi_i, \quad \mathbf{x}_i \in \mathbf{G}_1, \qquad \mathbf{x}_i\omega \geq b - \xi_i, \quad \mathbf{x}_i \in \mathbf{G}_2, \tag{4.2}$$

y la suma de los límites incumplidos ξ_i (o alguna suma ponderada de los límites incumplidos $h_i\xi_i$) es minimizada.

Además, el hiperplano de separación, $\mathbf{x}\omega = b$, será seleccionado de manera que los puntos que se encuentran dentro de la frontera, en la medida que sea posible, estén dentro de los límites, de este modo agudizar la diferenciación entre los grupos. Si bien por lo general no será posible anticipar que puntos se encuentran dentro de la frontera "verdadera" (es decir, aquellos puntos que satisfacen $\mathbf{x}_i\omega \leq b$ para $\mathbf{x}_i \in \mathbf{G}_1$ o $\mathbf{x}_i\omega \geq b$

para $\mathbf{x}_i \in \mathbf{G}_2$), es evidente que todos los puntos se encuentran dentro de los límites "ajustados". Es decir todos los puntos satisfarán $\mathbf{x}_i\omega \leq b + \xi_i$ para $\mathbf{x}_i \in \mathbf{G}_1$ o $\mathbf{x}_i\omega \geq -b + \xi_i$ para $\mathbf{x}_i \in \mathbf{G}_2$. Sea d_i la distancia de un punto $\mathbf{x}_i$ a su límite ajustado, se puede eficazmente combinar el objetivo de minimizar las desviaciones de límite con el objetivo de maximizar la suma (ponderada) de estas distancias $(\sum k_i d_i)$. Adaptar el problema en un contexto minimización, donde maximizar $\sum k_i d_i$ corresponde a minimizar $-\sum k_i d_i$, el objetivo combinado tendrá la forma

$$\begin{cases} \text{mín} \sum h_i \xi_i - \sum k_i d_i, \\ \text{sujeto a:} \\ \mathbf{x}_i\omega + d_i = b + \xi_i, \qquad \mathbf{x}_i \in \mathbf{G}_1, \\ -\mathbf{x}_i\omega + d_i = -b + \xi_i, \qquad \mathbf{x}_i \in \mathbf{G}_2, \end{cases} \tag{4.3}$$

Observación 1. *Se debe tener en cuenta que las distancias d_i son precisamente las variables de holgura que cambian las desigualdades de (4.2) en igualdades.*

Es importante destacar que, el procedimiento dará lugar a una solución en la que $d_i = 0$ siempre que el peso para reducir al mínimo el incumplimiento de los límites sobrepasa el peso para maximizar la distancia de $\mathbf{x}_i$ desde el límite ajustado y $\mathbf{x}_i$ incumple el límite verdadero; es decir, si $\xi_i > 0$, entonces $d_i = 0$ para todo $h_i > k_i$. Además, ξ_i nunca se verá obligada a tomar un valor más grande de lo necesario con el fin de permitir que la variable de holgura d_i sea positiva.(El recíproco también es cierto, para $h_i < k_i$, ξ_i será empujado a su límite con el fin de aumentar el valor de la variable de holgura d_i.) Cuando $\xi_i = 0$, es decir, cuando $\mathbf{x}_i$ está dentro del límite "verdadero", d_i se enfrentará a su valor más grande para cualquier k_i positivo.

Una reducción significativa en el número de variables requeridas en (4.3) se puede lograr mediante la agregación de términos. Por ejemplo, mediante la sustitución del conjunto de las variables ξ_i en un solo un término ξ, el modelo se convierte en:

$$\begin{cases} \text{mín } H\xi - \sum k_i d_i, \\ \text{sujeto a:} \\ \mathbf{x}_i\omega + d_i = b + \xi, \qquad \mathbf{x}_i \in \mathbf{G}_1, \\ -\mathbf{x}_i\omega + d_i = -b + \xi, \qquad \mathbf{x}_i \in \mathbf{G}_2, \end{cases} \tag{4.4}$$

Aquí, ξ mide el máximo límite incumplido asociado con un candidato a solución discriminante (es decir, $\xi_i < \xi$ para todo i). La relación entre ξ y d_i es similar a la que existe entre ξ_i y d_i descrita anteriormente en (4.3). Si $\xi > 0$, $H > \sum K_i$ asegura que el más pequeño d_i siempre será 0.

La incorporación de $d_i's$ en lugar de $\xi_i's$ produce el problema de programación lineal:

$$\begin{cases} \text{mín} \sum h_i\alpha_i - Kd, \\ \text{sujeto a:} \\ \mathbf{x}_i\omega + d = b + \xi_i, \qquad \mathbf{x}_i \in \mathbf{G}_1, \\ -\mathbf{x}_i\omega + d = -b + \xi_i, \qquad \mathbf{x}_i \in \mathbf{G}_2, \end{cases} \tag{4.5}$$

en el cual d mide la distancia mínima de cualquier miembro del grupo a la frontera (hiperplano) $\mathbf{x}\omega = b$, cuando todo $\xi_i = 0$. Si $\xi_i > 0$ para algún i y $\sum h_i > K$, entonces $d = 0$. $\sum h_i < K$ asegura que los $\xi_i's$ tienden a su valor máximo.

La formulación completa, una variante extrema de la formulación anterior, en la que se sustituye el conjunto de variables ξ_i por ξ y el conjunto de variable d_i con d, proporcionan un problema de programación lineal en el que la tarea es:

$$\begin{cases} \text{mín } H\xi - Kd \\ \text{sujeto a:} \\ \mathbf{x}_i\omega + d = b + \xi, \qquad \mathbf{x}_i \in \mathbf{G}_1, \\ -\mathbf{x}_i\omega + d = -b + \xi, \qquad \mathbf{x}_i \in \mathbf{G}_2. \end{cases} \tag{4.6}$$

$H > K$ asegura que el objetivo de minimizar el máximo solapamiento (ξ) domina el objetivo. Para $\xi > 0$, d=0. Para $\xi = 0$, d es la medida de la distancia mínima a cualquier punto del hiperplano de separación $\mathbf{x}\omega = b$. $H < K$ asegura que ξ tiende a su valor máximo.

Conforme a lo expuesto anteriormente, el uso de programas lineales para determinar los coeficientes de las funciones discriminantes lineales ha sido ampliamente estudiado [Freed and Glover, 1981], [Gochet et al., 1997], [Joachimsthaler and Stam, 1990], [Mangasarian, 1965]. Uno de los primeros modelos de programación lineal para afrontar el problema de clasificación fue propuesto por Mangasarian [Mangasarian, 1965], el mismo que se fundamenta en la construcción de un hiperplano de separación de dos grupos de datos. Dos conjuntos de puntos pueden ser no separables por un hiperplano o superficie a través de un enfoque de programación lineal de un simple paso, pero se pueden separar estrictamente por más planos o superficies a través de un enfoque de programación lineal multi-paso ([Mangasarian, 1968]).

Diversos estudios de algoritmos de programación lineal para afrontar el problema discriminante, en la década de 1980, fueron desarrollados por Hand [Hand, 1981], Freed y Glover [Freed and Glover, 1981], y Bajgier y Hill [Bajgier and Hill, 1982]. Los mismos que propusieron los modelos de programación lineal para el problema de clasificación

de dos grupos, que incluyen minimizar la suma de las desviaciones (MSD), minimizar la desviación máxima (MMD), y minimizar la suma de las distancias interiores (MSID).

Para las formulaciones de algunos de estos modelos se utilizará la siguiente notación. Los subíndices i, j y k se utilizan para notar las observaciones, atributos y grupos, respectivamente, entonces x_{ij} es el valor del atributo j de la observación i. Sea p el número de atributos, k es el número de grupos, G_k representan el conjunto de datos del grupo k,

Los modelos de programación lineal para la resolución de problemas de clasificación son usados para determinar los coeficientes ω_j $j = 1, 2, \cdots, p$, para el atributo j y el término constante b. Estos diferentes modelos son capaces de determinar la función discriminante, la cual debe ser usada para clasificar nuevas observaciones en uno de los dos grupos. Por ejemplo, para una nueva observación i, x_{ij}, $j = 1, 2, \cdots, p$ es el valor del atributo j, y además:

(i) $\sum_{j=i}^{p} \omega_j x_{ij} > b$: la observación es clasificada en el grupo1.

(ii) $\sum_{j=i}^{p} \omega_j x_{ij} < b$: la observación es clasificada en el grupo2.

(iii) $\sum_{j=i}^{p} \omega_j x_{ij} = b$: la observación no puede ser clasificada.

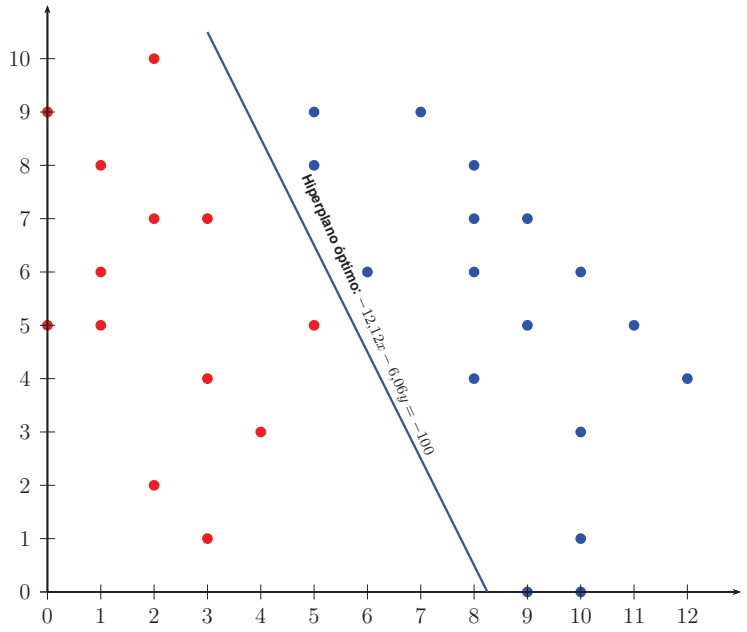
4.1.1 Modelo MMD

Freed y Glover en 1981 y 1986 sugirieron el modelo MMD (Maximize the Minimun of Deviations) para resolver el problema de clasificación. Su formulación es

$$(MMD) \begin{cases} \max d \\ sujeto\ a: \\ \sum_{j=1}^{p} \omega_j x_{ij} - d \geq b, \quad \forall i \in G_1 \\ \sum_{j=1}^{p} \omega_j x_{ij} + d \leq b, \quad \forall i \in G_2 \\ \sum_{j=1}^{p} \omega_j x_{ij} = 1 \end{cases} \tag{4.7}$$

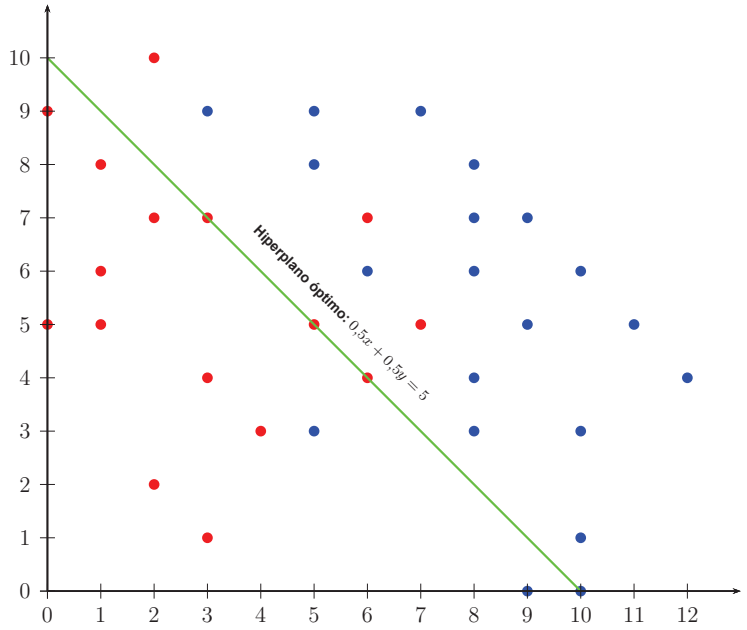
donde, ω_j y b son libres.

Figura 4.1: Modelo MMD para un conjunto de datos separables.



Fuente: O. L. Mangasarian, W. N. Street, and W. W. Wolberg, "Breast Cancer Diagnosis and Prognosis Via Linear Programming" Operations Research. 43 (1995), 570-577. Elaborado por: Autor.

Figura 4.2: Modelo MMD para un conjunto de datos no separables.



Fuente: O. L. Mangasarian, W. N. Street, and W. W. Wolberg, "Breast Cancer Diagnosis and Prognosis Via Linear Programming" Operations Research. 43 (1995), 570-577. Elaborado por: Autor.

4.1.2 Modelo MSD

El modelo MSD (Minimize the Sums of Deviations) fue propuesto por Bajgier and Hill [Bajgier and Hill, 1982] y Freed y Glover [Freed and Glover, 1986]. El objetivo de esta formulación es minimizar la suma total de solapamientos. Este modelo puede ser formulado como sigue:

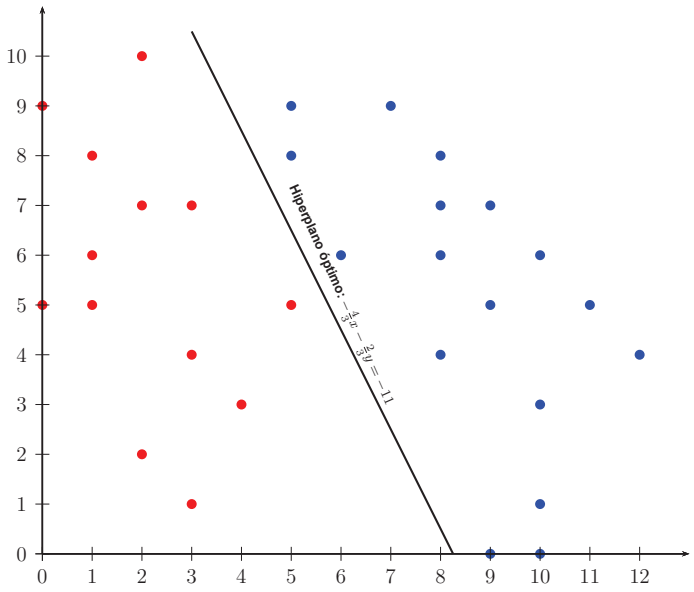
$$(MSD)\begin{cases} \text{mín} \sum_{i=1}^{n} d_i \\ sujeto\ a: \\ \sum_{j=1}^{p} \omega_j x_{ij} + d_i \geq b, \quad \forall i \in G_1 \\ \sum_{j=1}^{p} \omega_j x_{ij} - d_i \leq b, \quad \forall i \in G_2 \\ \sum_{j=1}^{p} \omega_j x_{ij} = 1 \end{cases} \tag{4.8}$$

donde, ω_j y b son libres y $d_i \geq 0$.
La regla de clasificación del modelo MSD es:

$d_i = 0$ si la observación i es correctamente clasificada

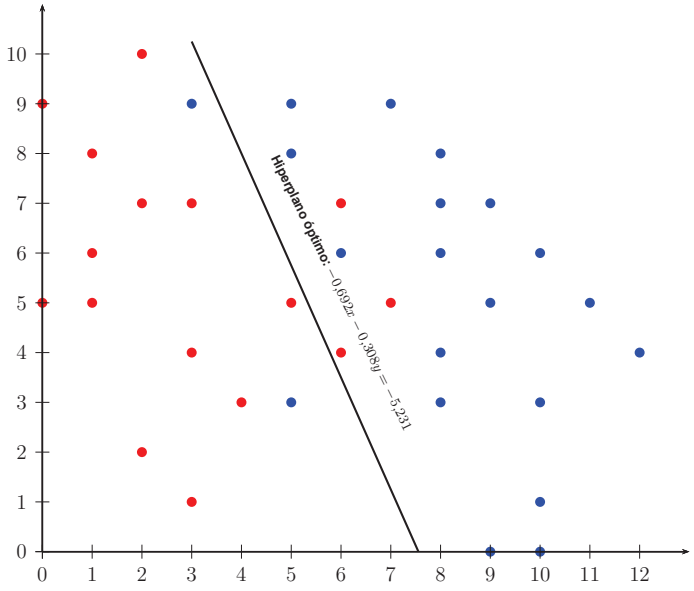
$d_i = \sum_{j=1}^{p} \omega_j x_{ij}$ si la observación i es clasificada erróneamente

Figura 4.3: Modelo MSD para un conjunto de datos separables.



Fuente: O. L. Mangasarian, W. N. Street, and W. W. Wolberg, "Breast Cancer Diagnosis and Prognosis Via Linear Programming" Operations Research. 43 (1995), 570-577. Elaborado por: Autor.

Figura 4.4: Modelo MSD para un conjunto de datos no separables.



Fuente: O. L. Mangasarian, W. N. Street, and W. W. Wolberg, "Breast Cancer Diagnosis and Prognosis Via Linear Programming" Operations Research. 43 (1995), 570-577. Elaborado por: Autor.

4.2 Comparación de las máquinas de soporte vectorial y la programación multicriterio

Dado un conjunto de n variables a cerca de los registros $\mathbb{X}^T = (x_1, \cdots, x_l)$, sea $x_i = (x_{i1}, \cdots, x_{in})^T$ la muestra tratada de la data para las variables, donde $i = 1, \cdots, l$ y l es el tamaño de la muestra. Se supone que se dispone de un caso linealmente separable y se quiere determinar los coeficientes o pesos para un subconjunto apropiado de variables (o atributos), denotado por $\omega = (\omega_1, \cdots, \omega_n)^T$, y una frontera (límite) b para separar en dos clases: por ejemplo $\mathbf{B}$ (Bueno) y $\mathbf{M}$ (malo); es decir

$$(\omega \cdot x_i) \leq b, \quad \text{para } x_i \in \mathbf{B} \tag{4.9}$$

y

$$(\omega \cdot x_i) \geq b, \quad \text{para } x_i \in \mathbf{M} \tag{4.10}$$

donde $(\omega \cdot x_i)$ es el producto interior.

En la formulación de Máquinas de Soporte de Vectores (SVM), la pertenencia de cada x_i en la clase $+1$ (Malo) o -1 (Bueno) especificada por una matriz diagonal $l \times l$, $Y = \{y_{ii}\}$ con entradas $+1$ y -1. Dados dos planos acotados $(\omega \cdot x_i) = \pm 1$, el problema descrito anteriormente sería:

$$(\omega \cdot x_i) \leq b - 1, \quad y_{ii} = -1, \tag{4.11}$$

$$(\omega \cdot x_i) \geq b + 1, \quad y_{ii} = +1. \tag{4.12}$$

esto se muestra como

$$\mathbb{Y}(\mathbb{X}\omega - eb) \geq e, \tag{4.13}$$

donde e es un vector de unos.
Una formulación de Máquinas de Soporte de Vectores estándar, que puede ser abordado mediante programación cuadrática, es el siguiente:
Si se define ξ_i como una variable de holgura, entonces $\xi = (\xi_1, \cdots, \xi_t)^T$ es un vector de holgura. Una SVM se establece como

$$\text{mín } \frac{1}{2}||\omega||_2^2 + C||\xi||_2^2 \tag{4.14}$$

$$\text{sujeto a:} \mathbb{Y}(\mathbb{X}\omega - eb) \geq e - \xi \tag{4.15}$$

$$\xi \geq 0 \tag{4.16}$$

donde e es un vector de unos y $C > 0$.
En la formulación de un problema de programación multicriterio, la variable ξ_i es vista como el solapamiento (superposición) con respecto a la muestra de entrenamiento x_i. Sea β_i la distancia de la muestra de entrenamiento x_i al discriminador $(\omega \cdot x) = b$ (hiperplano separador). Las restricciones para los dos grupos se pueden escribir como:

$$(\omega \cdot x_i) \leq b + \xi_i - \beta_i, \quad y_{ii} = -1 \text{ (Bueno)}, \tag{4.17}$$

y

$$(\omega \cdot x_i) \geq b - \xi_i + \beta_i, \quad y_{ii} = +1 \text{ (Malo)}. \tag{4.18}$$

Esto se puede escribir como $\mathbb{Y}(\mathbb{X}\omega - eb) \geq \beta - \xi$, donde e es un vector de unos. Además, un problema de programación cuadrática multicriterio puede ser formulada como

$$\text{mín}\, ||\xi||_2^2 - ||\beta||_2^2, \tag{4.19}$$

$$\text{sujeto a:} \mathbb{Y}(\mathbb{X}\omega - eb) \geq \beta - \xi, \tag{4.20}$$

$$\xi, \beta \leq 0. \tag{4.21}$$

Comparando las dos formulaciones anteriores, se puede ver que el modelo de programación de múltiples criterios es similar al modelo de máquina de vectores de soporte en cuanto a la formulación, considerando la reducción al mínimo de la superposición de los datos. Sin embargo, los antiguos intentos de medir todas las distancias posibles β de la muestra de entrenamiento xi al hiperplano de separación, mientras que las segundas correcciones de la distancia como 1 (a través de los planos de delimitación $(\omega \cdot x) = b \pm 1)$ a partir de los vectores de soporte. Aunque la interpretación puede variar, el modelo de programación de criterios múltiples aborda más parámetros de control que la máquina de vectores de soporte, que puede proporcionar una mayor flexibilidad para una mejor separación de los datos en el marco de la programación matemática.

4.3 Programación Lineal Multicriterio MCLP

En esta sección se presta atención a la discusión sobre la formulación de la programación de criterios múltiples. En el análisis discriminante lineal, la separación de los datos se puede lograr mediante dos objetivos opuestos. El primer objetivo es *separar las ob-*

servaciones al minimizar la suma de las desviaciones (MSD) entre las observaciones. El segundo objetivo es *maximizar las distancias mínimas (MMD) de las observaciones desde el valor crítico* [Freed and Glover, 1986]. Como se comentó en la sección anterior, la superposición de los datos ξ debe ser reducido al mínimo, mientras que la distancia β tiene que ser maximizada. Sin embargo, es difícil para la programación lineal tradicional optimizar *MMD* y *MSD* simultáneamente. De acuerdo con el concepto de óptimo de Pareto, podemos buscar el mejor compromiso de las dos medidas [Shi et al., 2002] [Shi et al., 2001] . El primer modelo de programación lineal multicriterio (MCLP) se puede describir como sigue:

$$\text{mín} \sum_{i=1}^{l} \xi_i, \tag{4.22}$$

$$\text{máx} \sum_{i=1}^{l} \beta_i, \tag{4.23}$$

$$\text{sujeto a: } (\omega \cdot x_i) = b + y_i(\xi_i - \beta_i), \quad i = 1, \cdots, l \tag{4.24}$$

$$\xi, \beta \geq 0 \tag{4.25}$$

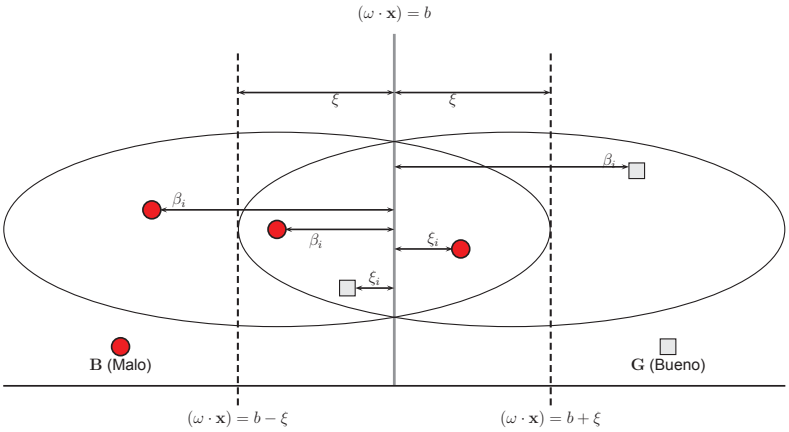
donde, ξ_i es la superposición y β_i es la distancia desde la muestra de entrenamiento x_i al discriminador $(\omega \cdot x_i) = b$ (hiperplano de separación de clasificación). Si $y_i \in \{+1, -1\}$ denota la etiqueta de x_i y l es el tamaño de la muestra, un conjunto de entrenamiento puede ser interpretado como un par $\{x_i, y_i\}$, donde los x_i son los valores del vector de variables y $y_i \in \{+1, -1\}$. La interpretación geométrica del modelo se muestra en la Figura 4.5.

Los objetivos del problema determinado por las ecuaciones (4.22)-(4.25), la forma original del MCPL, son difíciles de optimizar. Para facilitar los cálculos, la solución de compromiso [184] es empleada para reformar el modelo anterior, de modo que, se puede identificar sistemáticamente el mejor compromiso entre $-\sum \xi_i$ y $\sum \beta_i$ para una solución óptima. Tener en cuenta que ya que se debe considerar el espacio objetivo de este problema de dos criterios, en lugar de minimizar $\sum \xi_i$, se debe maximizar $-\sum \xi_i$. Se supone constante el "valor ideal" de $-\sum \xi_i$ y $\sum \beta_i$, $\xi^* > 0$ y $\beta^* > 0$, respectivamente. Entonces, si $-\sum \xi > \xi^*$, se define la medida de penalización como $-d_\xi^+ = \sum \xi_i + \xi^*$; para otros casos esto es 0. Si $-\sum \xi < \xi^*$, la medida de penalización se define como $d_\xi^- = \xi^* + \sum \xi_i$; para otros casos esto es 0. Por lo tanto se tiene:

(i) $\xi^* + \sum \xi_i = d_\xi^- - d_{\xi_i}^+$,

(ii) $|\xi^* + \sum \xi_i| = d_\xi^- + d_\xi^+$, y

Figura 4.5: Interpretación Geométrica de MCLP



Fuente:Shi Y., Tian Y., Kou G., Peng Y., Li J., Optimization Based Data Mining: Theory and Applications. pag:121. Elaborado por: Autor.

(iii) $d_\xi^-, d_\xi^+ \geq 0.$

Análogamente, se tiene

(i) $\beta^* - \sum \beta_i = d_\beta^- - d_{\beta_i}^+,$

(ii) $|\beta^* - \sum \beta_i| = d_\beta^- + d_\beta^+,$ y

(iii) $d_\beta^-, d_\beta^+ \geq 0.$

De esta manera el modelo MCLP de dos clases ha evolucionado hasta el siguiente modelo:

$$\text{mín}\, d_\xi^+ + d_\xi^- + d_\beta^+ + d_\beta^-, \tag{4.26}$$

sujeto a:

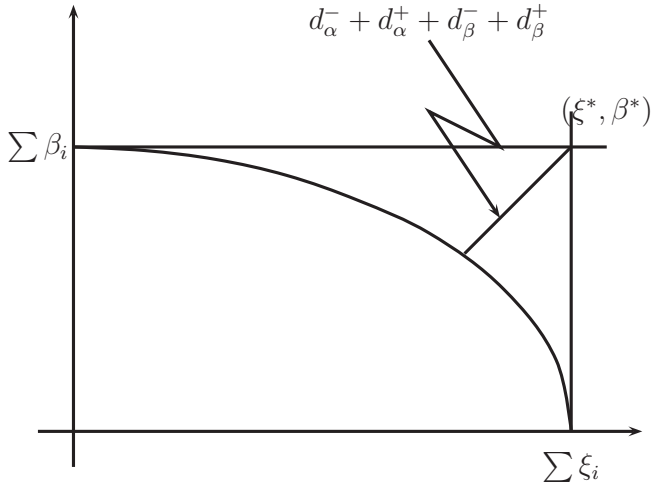
$$\xi^* + \sum_{i=1}^{l} \xi_i = d_\xi^- + d_\xi^+, \tag{4.27}$$

$$\beta^* - \sum_{i=1}^{l} \beta_i = d_\beta^- - d_\beta^+, \tag{4.28}$$

$$(\omega \cdot x_i) = b + y_i(\xi - \beta_i), \quad i = 1, \cdots, l \tag{4.29}$$

$$\xi, \beta \geq 0, \qquad d_\xi^+, d_\xi^-, d_\beta^+, d_\beta^- \geq 0. \tag{4.30}$$

Figura 4.6: Interpretación geométrica del compromiso de solución de MCLP

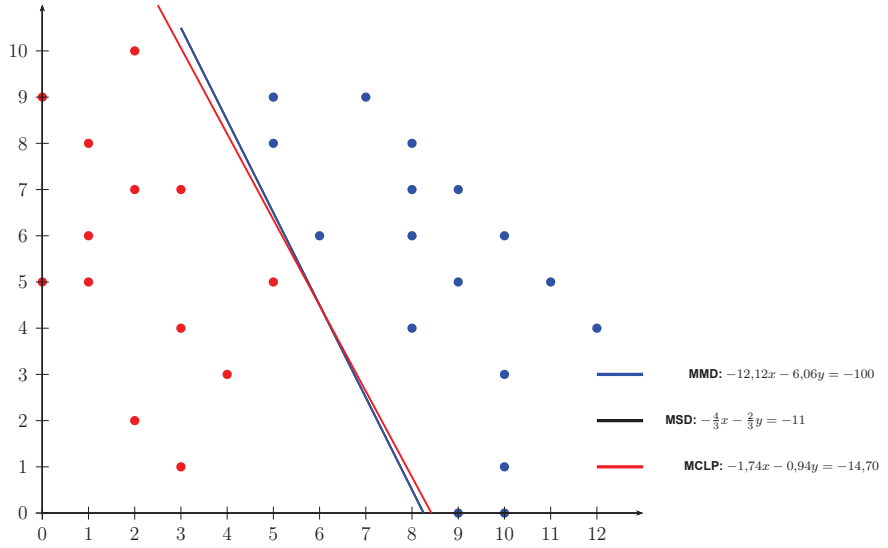


Fuente:Shi Y., Tian Y., Kou G., Peng Y., Li J., Optimization Based Data Mining: Theory and Applications. pag:122. Elaborado por: Autor.

Aquí ξ^* y β^* son dados, ω y b no son restringidos. El significado geométrico del modelo (4.26)-(4.30) se muestra en la figura 4.6.

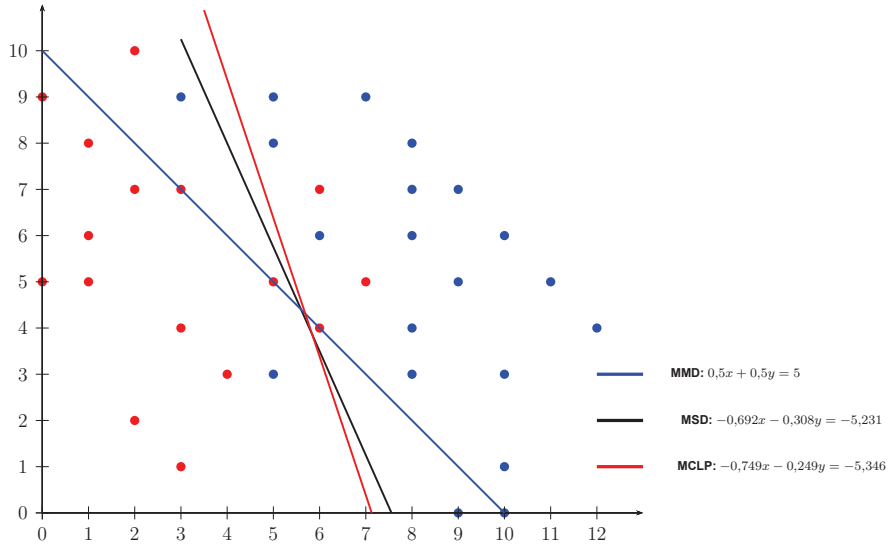
En el problema (4.26)-(4.30) se puede utilizar $\omega > 1$ para evitar una solución trivial. El valor de b también afectará a la solución. Sin embargo b es una variable en el problema (4.26)-(4.30), por tanto para algunas aplicaciones, el usuario puede elegir un valor fijo de b para obtener una solución como el clasificador.

Figura 4.7: Modelo MCLP para un conjunto de datos separables.



Fuente: O. L. Mangasarian, W. N. Street, and W. W. Wolberg, "Breast Cancer Diagnosis and Prognosis Via Linear Programming" Operations Research. 43 (1995), 570-577. Elaborado por: Autor.

Figura 4.8: Modelo MCLP para un conjunto de datos no separables.



Fuente: O. L. Mangasarian, W. N. Street, and W. W. Wolberg, "Breast Cancer Diagnosis and Prognosis Via Linear Programming" Operations Research. 43 (1995), 570-577. Elaborado por: Autor.

4.4 Programación Lineal Multicriterio para Múltiples Clases

Un problema multi-clase de clasificación de datos usando programación lineal multicriterio puede ser descrito como:

Dado un conjunto de n variables o atributos, sea $x_i = (x_{i1}, \cdots, x_{in})^T \in \mathbb{R}^n$ la muestra de observaciones de los datos para las variables, donde $i = 1, \cdots, l$ y l es el tamaño de la muestra. Si un problema dado puede ser predefinido como s clases diferentes, $\mathbf{G}_1, \cdots, \mathbf{G}_s$, entonces el límite entre la j-ésima clase y la (j+1)-ésima clase puede ser b_j, $j = 1, \cdots, s-1$. Se quiere determinar los coeficientes para un apropiado subconjunto de variables, denotados por $\omega = (\omega_1, \cdots, \omega_n)^T$ y escalares b_j tal que la separación de estas clases puede ser descrita como sigue:

$$(\omega \cdot x_i) \leq b_1, \qquad \forall x_i \in \mathbf{G}_1, \tag{4.31}$$

$$b_{k-1} \leq (\omega \cdot x_i) \leq b_k, \qquad \forall x_i \in \mathbf{G}_k, \tag{4.32}$$

$$(\omega \cdot x_i) \geq b_{s-1}, \qquad \forall x_i \in \mathbf{G}_s, \tag{4.33}$$

donde $\forall x_j \in \mathbf{G}_k$, $k = 1, \cdots, s$, significa que el caso de datos x_j pertenece a la clase $\mathbf{G}_k$.

En la separación de los datos, $(\omega \cdot x_i)$ es llamado *el score (puntuación)* del dato del caso i, que es una combinación lineal de los valores ponderados de las variables de atributo ω. Por ejemplo, en el caso del análisis de un portafolio de tarjetas de crédito, $(\omega \cdot x_i)$ puede representar el valor agregado de la puntuación del titular de la i-ésima tarjeta de crédito, en consideración de sus atributos edad, salario, educación, y residencia.

A pesar de que el límite b_j se define como un *escalar* en la separación de los datos anteriores, generalmente, b_j puede ser tratada como "variable" en la formulación. Sin embargo, si no hay ninguna solución factible sobre b_j "variable" en el análisis de datos reales, debe ser predeterminada como un parámetro de control de acuerdo a la experiencia del analista.

La calidad de la clasificación se mide por minimizar la superposición total de datos y maximizar las distancias de todos los datos a su límite de clases simultáneamente. Sea ξ_i^j el grado de superposición (solapamiento) con respecto a los datos del caso x_i dentro $\mathbf{G}_j$ y $\mathbf{G}_{j+1}$ a su límite ajustado.

Incorporando tanto ξ_i^j y β_i^j en las desigualdades de separación, un modelo de programación lineal multicriterio (MCLP) se puede definir como:

$$\text{mín} \sum_i \sum_j \xi_i^j \quad y \quad \text{máx} \sum_i \sum_j \beta_i^j, \tag{4.34}$$

sujeto a:

$$(\omega \cdot x_i) = b_1 + \xi_i^1 - \beta_i^1, \quad \forall x_i \in \mathbf{G}_1, \tag{4.35}$$

$$b_{k-1} - \xi_i^{k-1} + \beta_i^{k-1} = (\omega \cdot x_i) = b_k + \xi_i^k - \beta_i^k, \quad \forall x_i \in \mathbf{G}_k,\ k = 2, \cdots, s-1, \tag{4.36}$$

$$(\omega \cdot x_i) = b_{s-1} - \xi_i^{s-1} + \beta_i^{s-1}, \quad \forall x_i \in \mathbf{G}_s, \tag{4.37}$$

$$b_{k-1} + \xi_i^{k-1} \leq b_k - \xi_i^k, \quad k = 2, \cdots, s-1,\ i = 1, \cdots, l, \tag{4.38}$$

$$\xi_i^j, \beta_i^j \geq 0, \quad j = 1, \cdots, s-1,\ i = 1, \cdots, l, \tag{4.39}$$

donde los x_i son dados, ω y b_j sin restricciones. Se debe tener en cuenta que las restricciones (4.38) aseguran la existencia de los límites.

Si minimizar la superposición total de datos, maximizar las distancias de cada uno de los datos a su límite de la clase, o una combinación dada de ambos criterios se considera por separado, modelo (4.34) - (4.39) se reduce a la programación lineal (LP) de clasificación (conocido como análisis discriminante lineal), que fue iniciado por Freed y Glover. Sin embargo, el único criterio LP no puede determinar la “mejor solución” de las dos mediciones de clasificación errónea. Por lo tanto, el modelo anterior es potencialmente mejor que el modelo de clasificación basado en programación lineal (LP) en la identificación de la mejor compensación de los errores de clasificación para la separación de datos.

Para facilitar el cálculo sobre datos de la vida real, una aproximación para una solución de compromiso es emplear una reformulación del modelo (4.34) - (4.39) para el “mejor compromiso” entre $\sum_i \sum_j \xi_i^j$ y $\sum_i \sum_j \beta_i^j$. Se supone que el “valor ideal” para los $s-1$ solapamientos de clases $(-\sum_i \xi_i^1, \cdots, -\sum_i \xi_i^{s-1})$ es $(\xi_*^1, \cdots, \xi_*^{s-1}) > 0$, y el “valor ideal” de $(\sum_i \beta_i^1, \cdots, -\sum_i \beta_i^{s-1})$ es $(\beta_*^1, \cdots, \beta_*^{s-1}) > 0$. La selección de los valores ideales depende de la naturaleza y formato de los datos del problema. Cuando $-\sum_i \xi_i^j > \xi_*^j$, se define la medida de penalización como $-d_{\xi j}^+ = \xi_*^j + \sum_i \xi_i^j$; para otros casos, esto es 0, donde $j = 1, \cdots, s-1$. Cuando $-\sum_i \xi_i^j < \xi_*^j$, se define la medida de la penalización con $d_{\xi j}^- = \xi_*^j + \sum_i \xi_i^j$; para otros casos, esto es 0, donde $j = 1, \cdots, s-1$. Por lo tanto,

se tiene:

Teorema 2.

$$(i) \qquad \xi_*^j + \sum_i \xi_i^j = d_{\xi_j}^- - d_{\xi_j}^+ \tag{4.40}$$

$$(ii) \qquad \left| \xi_*^j + \sum_i \xi_i^j \right| = d_{\xi_j}^- + d_{\xi_j}^+ \tag{4.41}$$

$$(iii) \qquad d_{\xi_j}^-, d_{\xi_j}^+ \geq 0, \quad j = 1, \cdots, s-1 \tag{4.42}$$

Similarmente, se puede derivar:

Corolario 1.

$$(i) \qquad \beta_*^j + \sum_i \beta_i^j = d_{\beta_j}^- - d_{\beta_j}^+ \tag{4.43}$$

$$(ii) \qquad \left| \beta_*^j + \sum_i \beta_i^j \right| = d_{\beta_j}^- + d_{\beta_j}^+ \tag{4.44}$$

$$(iii) \qquad d_{\beta_j}^-, d_{\beta_j}^+ \geq 0, \quad j = 1, \cdots, s-1 \tag{4.45}$$

La aplicación de los resultados anteriores en el modelo (4.34)-(4.39), se reformuló como:

$$\text{mín} \sum_{j=1}^{s-1} (d_{\xi_j}^- + d_{\xi_j}^+ + d_{\beta_j}^- + d_{\beta_j}^+) \tag{4.46}$$

sujeto a:

$$\xi_*^j + \sum_i \xi_i^j = d_{\xi_j}^- - d_{\xi_j}^+, \quad j = 1, \cdots, s-1 \tag{4.47}$$

$$\beta_*^j + \sum_i \beta_i^j = d_{\beta_j}^- - d_{\beta_j}^+, \quad j = 1, \cdots, s-1 \tag{4.48}$$

$$(\omega \cdot x_i) = b_1 + \xi_i^1 - \beta_i^1, \quad \forall x_i \in \mathbf{G}_1, \tag{4.49}$$

$$b_{k-1} + \xi_i^{k-1} - \beta_i^{k-1} = (\omega \cdot x_i) = b_k + \xi_i^k - \beta_i^k, \quad \forall x_i \in \mathbf{G}_k,\ k = 2, \cdots, s-1 \tag{4.50}$$

$$(\omega \cdot x_i) = b_{s-1} - \xi_i^{s-1} - \beta_i^{s-1}, \quad \forall x_i \in \mathbf{G}_s, \tag{4.51}$$

$$b_{k-1} + \xi_i^{k-1} \geq b_k - \xi_i^k, \quad k = 2, \cdots, s-1,\ i = 1, \cdots, l, \tag{4.52}$$

$$\xi_i^j, \beta_i^j \geq 0, \quad j = 1, \cdots, s-1, \ i = 1, \cdots, l, \tag{4.53}$$

$$d_{\xi_j}^-, d_{\xi_j}^+, d_{\beta_j}^-, d_{\xi_j}^+ \geq 0, \quad j = 1, \cdots, s-1 \tag{4.54}$$

donde x_i, ξ_*^j y β_*^j son dados, ω y b_j son libres.

Una vez que los límites ajustados $b_{k-1} + \xi_i^{k-1} \leq b_k - \xi_i^k$, $k = 2, \cdots, s-1$, $i = 1, \cdots, l$ son debidamente elegidos, el modelo (4.46)-(4.54) relaja las condiciones de separación de datos de modo que pueda considerar el mayor número de datos superpuestos como sea posible en el proceso de clasificación. Se puede decir que el modelo (4.46)-(4.54) es una "fórmula de separación débil", Con esta motivación, podemos construir una "fórmula de separación media" en los límites de clase absolutos en el siguiente modelo (4.55) -(4.63) y una "fórmula de separación fuerte", que contiene el menor número de datos superpuestos como sea posible en el siguiente modelo (4.64)-(4.72).

$$\text{mín} \sum_{j=1}^{s-1} (d_{\xi_j}^- + d_{\xi_j}^+ + d_{\beta_j}^- + d_{\beta_j}^+) \tag{4.55}$$

sujeto a:

$$\xi_*^j + \sum_i \xi_i^j = d_{\xi_j}^- - d_{\xi_j}^+, \quad j = 1, \cdots, s-1 \tag{4.56}$$

$$\beta_*^j + \sum_i \beta_i^j = d_{\beta_j}^- - d_{\beta_j}^+, \quad j = 1, \cdots, s-1 \tag{4.57}$$

$$(\omega \cdot x_i) = b_1 - \beta_i^1, \quad \forall x_i \in \mathbf{G}_1, \tag{4.58}$$

$$b_{k-1} + \beta_i^{k-1} = (\omega \cdot x_i) = b_k - \beta_k^1, \quad \forall x_i \in \mathbf{G}_k, \ k = 2, \cdots, s-1, \tag{4.59}$$

$$(\omega \cdot x_i) = b_{s-1} - \beta_i^{s-1}, \quad \forall x_i \in \mathbf{G}_s, \tag{4.60}$$

$$b_{k-1} + \epsilon \leq b_k - \xi_i^k, \quad k = 2, \cdots, s-1, \ i = 1, \cdots, l, \tag{4.61}$$

$$\xi_i^j, \beta_i^j \geq 0, \quad j = 1, \cdots, s-1, \quad i = 1, \cdot, l \tag{4.62}$$

$$d_{\xi_j}^-, d_{\xi_j}^+, d_{\beta_j}^-, d_{\beta_j}^+ \geq 0, \quad j = 1, \cdots, s-1, \tag{4.63}$$

donde x_i, ϵ, ξ_*^j y β_*^j son dados, ω y b_j son libres.

$$\text{mín} \sum_{j=1}^{s-1} (d_{\xi_j}^- + d_{\xi_j}^+ + d_{\beta_j}^- + d_{\beta_j}^-) \tag{4.64}$$

sujeto a

$$\xi_*^j + \sum_i \xi_i^j = d_{\xi_j}^- - d_{\xi_j}^+, \quad j = 1, \cdots, s-1 \tag{4.65}$$

$$\beta_*^j + \sum_i \beta_i^j = d_{\beta_j}^- - d_{\beta_j}^+, \quad j = 1, \cdots, s-1 \tag{4.66}$$

$$(\omega \cdot x_i) = b_1 - \xi_i^1 - \beta_i^1, \forall x_i \in \mathbf{G}_1, \tag{4.67}$$

$$b_{k-1} + \xi_i^{k-1} + \beta_i^{k-1} = (\omega \cdot x_i) = b_k - \xi_i^k - \beta_i^k, \forall x_i \in \mathbf{G}_k, \ k = 2, \cdots, s-1 \tag{4.68}$$

$$(\omega \cdot x_i) = b_{s-1} + \xi_i^{s-1} + \beta_i^{s-1}, \quad \forall x_i \in \mathbf{G}_s, \tag{4.69}$$

$$b_{k-1} + \xi_i^{k-1} \geq b_k - \xi_i^k, \ k = 2, \cdot, s-1, \ i = 1, \cdots, l, \tag{4.70}$$

$$\xi_i^j, \beta_i^j \geq 0, \quad j = 1, \cdots, s-1, \ i = 1, \cdots, l, \tag{4.71}$$

$$d_{\xi_j}^-, d_{\xi_j}^+, d_{\beta_j}^-, d_{\beta_j}^+ \geq 0, \quad j = 1, \cdots, s-1, \tag{4.72}$$

donde x_i, ξ_*^j y β_*^j son dadas, ω y b_j son libres.

Una relación de pérdida de los tres modelos anteriores está dada por:

Teorema 3.

(i) Si un dato del caso x_i es clasificado en un grupo dado $\mathbf{G}_j$ por el modelo (4.64)-(4.72), entonces este puede estar en $\mathbf{G}_j$ usando los modelos (4.55)-(4.63) y (4.46)-(4.54)

(ii) Si un dato del caso x_i es clasificado en un grupo dado $\mathbf{G}_j$ por el modelo (4.55)-(4.63), entonces este puede estar en $\mathbf{G}_j$ usando el modelo (4.46)-(4.54)

Demostración. Esto se desprende el hecho de que para un cierto valor de $\epsilon > 0$, las soluciones factibles del modelo (4.64)-(4.72) es las soluciones factibles de modelos (4.55) - (4.63) y (4.46) -(4.54), y las soluciones factibles de modelo (4.55) - (4.63) es éstos de modelo (4.46) -(4.54). □

4.5 Programación Lineal Milticriterio Penalizada

En muchas de las aplicaciones de la minería de datos en problemas de la vida real, los tamaños de las muestras sobre las diferentes clases varían (son distintas); es decir, el conjunto de entrenamiento es desequilibrado. Normalmente, dado un conjunto de datos para el caso binario (Bueno vs. Malo), hay mucho más registros de los buenos que los registros de los malos. En el proceso de formación, el mejor clasificador sería difícil de encontrar si se usa el modelo (4.26)-(4.30). Para superar la dificultad del enfoque MCLP, se propone el siguiente método MCLP penalizado (4.73)-(4.77) en el tratamiento del problema de la puntuación de crédito de la vida real.

$$\min d_{\xi}^{-} + d_{\xi}^{+} + d_{\beta}^{-} + d_{\beta}^{+}, \tag{4.73}$$

sujeto a:

$$\xi^{*} + p\frac{n_2}{n_1}\sum_{i\in\mathbf{B}}\xi_i + \sum_{i\in\mathbf{G}}\xi_i = d_{\xi}^{-} - d_{\xi}^{+}, \tag{4.74}$$

$$\beta^{*} + p\frac{n_2}{n_1}\sum_{i\in\mathbf{B}}\beta_i + \sum_{i\in\mathbf{G}}\beta_i = d_{\beta}^{-} - d_{\beta}^{+}, \tag{4.75}$$

$$(\omega \cdot x_i) = b + y_i(\xi_i - \beta_i), \quad i = 1, \cdot, l, \tag{4.76}$$

$$\xi^{*}, \beta^{*} \geq 0, \quad d_{\xi}^{-}, d_{\xi}^{+}, d_{\beta}^{-}, d_{\beta}^{+} \geq 0, \tag{4.77}$$

donde n_1 y n_2 son los números de muestras correspondientes a las dos clases, y $p \geq 1$ es el parámetro penalizado.

La distancia está en equilibrio sobre los dos lados de b con los parámetros n_1/n_2, aun cuando existen menos registros de la clase “Malo” (+1) a la derecha del hiperplano separador (score de crédito) b. El valor de p aumenta el efecto de la distancia “Malo” y penaliza mucho más si se quiere más “Malos” a la derecha del hiperplano de separación. Si $n_1 = n_2$, $p = 1$, el modelo anterior degenera al modelo MCLP original (4.22)-(4.25). Si $n_1 < n_2$, entonces $p \geq 1$ es utilizado para hacer que la tasa de captura “Malos” de PMCLP sea mayor que la de MCLP con el mismo n_1, n_2.

4.6 Programación Lineal Milticriterio Regularizada para la Clasificación

Dados: una matriz $\mathbb{A} \in \mathbb{R}^{m\times n}$ y los vectores $c, d \in \mathbb{R}^m_+$, la programación lineal multicriterio tiene la siguiente versión:

$$\begin{cases} \text{mín}_{u,v}\, d^T u - c^T v \\ \text{sujeto a:} \\ a_i x - u_i + v_i = b, \quad i = 1, 2, \cdots, l \\ a_i x + u_i - v_i = b, \quad i = l+1, l+2, \cdots, m \\ u, v \geq 0. \end{cases} \tag{4.78}$$

donde a_i es la i-ésima fila de $\mathbb{A}$ que contiene todos los datos dados.
El modelo MCLP es un programa especial lineal, y se ha utilizado con éxito en la minería de datos para un número de aplicaciones con grandes conjuntos de datos [Shi et al., 2005],[Shi et al., 2001],[Shi et al., 2008],[Zhang et al., 2004].
Sin embargo, no se puede garantizar que este modelo siempre tenga una solución. Obviamente, el conjunto factible de MCLP es no vacío, como el vector cero es un punto factible. Para $c \geq 0$, la función objetivo puede no tener un límite inferior en el conjunto factible. En este trabajo, para garantizar la existencia de una solución, se añade términos de regularización en la función objetivo, y se considera el siguiente MCLP regularizada

$$\begin{cases} \text{mín}_z\, \frac{1}{2}\omega^T \mathbb{H}\omega + \frac{1}{2}u^T \mathbb{Q}u + d^T u - c^T v, \\ \text{sujeto a:} \\ (\omega \cdot x_i) + u_i - v_i = b, \quad i = 1, 2, \cdots, l \\ (\omega \cdot x_i) - u_i + v_i = b, \quad i = l+1, l+2, \cdots, m \\ u, v \geq 0, \end{cases} \tag{4.79}$$

donde $z = (\omega, u, v, b) \in \mathbb{R}^{n+m+m+1}$,$\mathbb{H} \in \mathbb{R}^{n\times n}$ y $\mathbb{Q} \in \mathbb{R}^{m\times m}$ son matrices simétricas definidas positivas. El MCLP regularizado es un programa cuadrático convexo. Aunque la función objetivo

$$f(z) := \frac{1}{2}x^T \mathbb{H}x + \frac{1}{2}u^T \mathbb{Q}u + d^T u - c^T v$$

no es una función estrictamente convexa, podemos demostrar que (4.79) siempre tiene una solución. Además, el conjunto solución de (4.79) es acotado si $\mathbb{H}, \mathbb{Q}, d, c$ se eligen adecuadamente.
Sean $\mathbb{I}_1 \in \mathbb{R}^{l\times l}$, $\mathbb{I}_2 \in \mathbb{R}^{(m-l)\times(m-l)}$ matrices identidad,

$$\mathbb{A}_1 = \begin{pmatrix} a_1 \\ \vdots \\ a_l \end{pmatrix}, \qquad \mathbb{A}_2 = \begin{pmatrix} a_{l+1} \\ \vdots \\ a_m \end{pmatrix}, \tag{4.80}$$

$$\mathbb{A} = \begin{pmatrix} \mathbb{A}_1 \\ \mathbb{A}_2 \end{pmatrix}, \qquad \mathbb{E} = \begin{pmatrix} \mathbb{I}_1 & 0 \\ 0 & -\mathbb{I}_2 \end{pmatrix}, \tag{4.81}$$

y $e \in \mathbb{R}^m$ es el vector cuyos elementos son todos 1. Sea

$$\mathbb{B} = \begin{pmatrix} \mathbb{X} & \mathbb{E} & -\mathbb{E} & -e \end{pmatrix} \tag{4.82}$$

El conjunto factible de (4.79) esta dado por

$$\mathcal{F} = \{z | \mathbb{B}z = 0,\ u \geq 0,\ v \geq 0\}. \tag{4.83}$$

Puesto que (4.79) es un programa convexo con restricciones lineales, las conocidas condiciones de KKT es una condición necesaria y suficiente para la optimalidad. Para demostrar que $f(z)$ está acotada en $\mathcal{F}$, se considerará el sistema KKT de (4.79).

4.6.1 Conjunto solución de RMCLP

Sin pérdida de generalidad, se supone que $l > 0$ y $m - l > 0$.

Teorema 4. *El conjunto solución de RMCLP(4.79) es no vacío.*

Demostración. Se muestra que bajo la suposición que $l > 0$, $m - l > 0$, la función objetivo tiene un límite inferior. Tener en cuenta que los primeros términos en la función objetivo son no negativos. Si existe una sucesión z^k en $\mathcal{F}$ tal que $f(z^k) \longrightarrow -\infty$, entonces existe i tal que $v_i^k \to \infty$, que, conjuntamente con las restricciones de (4.79), implica que debe ser j tal que $|x_j^k| \to \infty$ o $|u_j^k| \to \infty$. Sin embargo, la función objetivo tiene términos cuadráticos en x y u que son más grandes que los términos lineales cuando $k \to \infty$. Esto contradice $f(z^k) \to -\infty$. Por lo tanto, por el teorema Frank-Wolfe, el modelo MCLP regularizado (4.79) siempre tiene una solución. Lo que completa la demostración. □

Ahora se demostrará que el conjunto solución del problema (4.79) está acotada si los parámetros $\mathbb{H}, \mathbb{Q}, d, c$ se eligen adecuadamente.

Teorema 5. *Se supone que* $\mathbb{A}\mathbb{H}^{-1}\mathbb{A}^T$ *es no singular. Sea* $\mathbb{G} = (\mathbb{A}\mathbb{H}^{-1}\mathbb{A}^T)^{-1}$, $\mu = 1/e^T$ *y*

$$\mathbb{M} = \begin{pmatrix} \mathbb{Q} + \mathbb{E}\mathbb{G}\mathbb{E} - \mu\mathbb{E}\mathbb{G}ee^T\mathbb{G}\mathbb{E} & -\mathbb{E}\mathbb{G}\mathbb{E} + \mu\mathbb{E}\mathbb{G}ee^T\mathbb{G}\mathbb{E} \\ -\mathbb{E}\mathbb{G}\mathbb{E} + \mu\mathbb{E}\mathbb{G}ee^T\mathbb{G}\mathbb{E} & \mathbb{E}\mathbb{G}\mathbb{E} - \mu\mathbb{E}\mathbb{G}ee^T\mathbb{G}\mathbb{E} \end{pmatrix}, \tag{4.84}$$

$$q = \begin{pmatrix} d \\ -c \end{pmatrix}, \quad q = \begin{pmatrix} u \\ v \end{pmatrix}. \tag{4.85}$$

Entonces el problema (4.79) es equivalente al problema de complementariedad lineal

$$\mathbb{M}y + q \geq 0, \qquad y \geq 0, \qquad y^T(\mathbb{M}y + q) = 0. \tag{4.86}$$

Si se elige $\mathbb{Q}$ y $\mathbb{H}$ tal que $\mathbb{M}$ sea una matriz semidefinida positiva y c,d satisfacen

$$d + 2\mathbb{Q}e > (\mu\mathbb{E}\mathbb{G}ee^T\mathbb{G}\mathbb{E} - \mathbb{E}\mathbb{G}\mathbb{E})e > c. \tag{4.87}$$

entonces el problema (4.79) tiene un conjunto solución no vacío y acotado.

Demostración. Se considera la condición de KKT asociada al problema (4.79)

$$\mathbb{H}\omega + \mathbb{A}^T\lambda = 0, \tag{4.88}$$

$$-c - \mathbb{E}\lambda - \beta = 0, \tag{4.89}$$

$$\mathbb{Q}u + \mathbb{E}\lambda + d - \alpha = 0, \tag{4.90}$$

$$\mathbb{B}z = 0, \tag{4.91}$$

$$e^T\lambda = 0, \tag{4.92}$$

$$u \geq 0, \qquad \alpha \geq 0, \qquad \alpha^T u = 0, \tag{4.93}$$

$$v \geq 0, \qquad \beta \geq 0, \qquad \beta^T v = 0. \tag{4.94}$$

de las primeras tres igualdades en la condición KKT, se tiene

$$\omega = -\mathbb{H}^{-1}\mathbb{A}^T\lambda, \tag{4.95}$$

$$\beta = -c - \mathbb{E}\lambda, \tag{4.96}$$

$$\alpha = \mathbb{Q}u + \mathbb{E}\lambda + d. \tag{4.97}$$

Sustituyendo ω en la cuarta igualdad en la condición KKT da

$$\lambda = \mathbb{G}(\mathbb{E}u - \mathbb{E}v - eb). \tag{4.98}$$

Además, de la quinta igualdad en la condición KKT, se obtiene
$b = \mu e^T \mathbb{G}\mathbb{E}(u - v)$. Por lo tanto, βy α pueden ser definidos por u, v como

$$\beta = -c - \mathbb{E}\mathbb{G}(\mathbb{E}u - \mathbb{E}v - b) = -c - \mathbb{E}\mathbb{G}(\mathbb{E}u - \mathbb{E}v - \mu e e^T \mathbb{G}\mathbb{E}(u - v)) \tag{4.99}$$

y

$$\alpha = \mathbb{Q}u + \mathbb{E}\mathbb{G}(\mathbb{E}u - \mathbb{E}v - eb) = \mathbb{Q}u + \mathbb{E}\mathbb{G}(\mathbb{E}u - \mathbb{E}\mathbb{E} - \mu e e^T \mathbb{G}\mathbb{E}(u - v)). \tag{4.100}$$

Esto implica que la condición KKT se puede escribir como el problema de complementariedad lineal (4.86). A partir del problema (4.79) es un problema convexo, es equivalente al problema de complementariedad lineal (4.86).
Sea $u = 2e$, $v = e$ y $y_0 = (2e, e)$. Entonces de (4.87) se tiene

$$\mathbb{M}y_0 + q = \begin{pmatrix} 2\mathbb{Q}e + \mathbb{E}\mathbb{G}\mathbb{E}e - \mu\mathbb{E}\mathbb{G}ee^T\mathbb{G}\mathbb{H}e + d \\ \mu\mathbb{E}\mathbb{G}ee^T\mathbb{G}\mathbb{E}e - \mathbb{E}\mathbb{G}\mathbb{E}e - c \end{pmatrix} > 0, \tag{4.101}$$

lo que implica que y_0 es un punto de (4.86) estrictamente factible. Por lo tanto, cuando $\mathbb{M}$ es una matriz semidefinida positiva, el conjunto solución de (4.86) es no vacío y acotado. Sea $y^* = (u^*, v^*)$ una solución de (4.86), entonces $z^* = (x^*, u^*, v^*, b^*)$ con

$$b^* = \mu e^T \mathbb{G}\mathbb{E}(u^* - v^*)\ y$$

$$\omega^* = -\mathbb{H}\mathbb{A}^T\mathbb{G}(\mathbb{E}u^* - \mathbb{E}v^* - \mu e e^T \mathbb{G}\mathbb{E}(u^* - v^*))$$

es una solución de (4.79). Por otra parte, de la condición KKT, es fácil verificar que la acotación del conjunto solución de (4.86) implica la acotación del conjunto solución de (4.79).

□

Capítulo 5

PROGRAMACIÓN LINEAL MULTICRITERIO Y DE RESTRICCIÓN MÚLTIPLE

Como se ha expuesto en el capítulo anterior, en los problemas de clasificación basados en programación lineal multicriterio (MCLP), se debe encontrar la solución óptima del problema MCLP como un clasificador (discriminador). En este capítulo, a partir de la teoría de la dualidad, los criterios múltiples pueden cambiar a múltiples niveles de restricción y viceversa, se explicará como un problema de programación lineal multicriterio (MCLP) se puede ampliar de forma lógica a un problema de programación lineal multicriterio y de múltiples niveles de restricción (MC2LP). En muchas aplicaciones de la vida real, como por ejemplo en el ámbito financiero, en la clasificación de cuentas de tarjetas de crédito, como enfrentar los tipos de errores es una cuestión clave. Los errores pueden ser causados por un límite fijo entre dos grupos, un grupo “Bueno” y un grupo “Malo”. Los dos tipos de errores pueden ser corregidos sistemáticamente mediante el uso de la estructura de MC2LP, que permite encontrar dos puntos de corte alterables. Para ello, se impone una penalización (o costo) para encontrar la solución potencial de entre todas las posibles soluciones del problema MC2LP.

5.1 Introducción

Tomando ventaja de varios algoritmos de clasificación establecidos en materia de optimización, los datos de las transacciones recogidas por una institución financiera pueden ser analizados de muchas maneras. Como resultado, el beneficio de nuevos usuarios se puede predecir, es decir, cuáles usuarios se inclinan a la quiebra (pérdida) y cuáles tienen un buen comportamiento de crédito. Sin embargo, para muchos de los problemas de la vida real, los sujetos o individuos no pueden ser separados por el hiperplano (discriminador) a pesar de la utilización de algunas técnicas avanzadas que involucran la utilización de núcleos (kernel) [Zhang et al., 2010]. Por lo tanto, cómo disminuir el número de sujetos mal clasificados se convierte en un gran problema al momento de tomar decisiones.

Como ya se estudió en secciones anteriores, la programación lineal (LP) es una herramienta útil para el análisis discriminante de un problema dados los grupos apropiados (por ejemplo, “bueno” y “malo”) [Freed and Glover, 1981]. La programación lineal multicriterio (MCLP) ha mejorado el resultado minimizando la suma de las desviaciones externos y la maximizando de la suma de las desviaciones internas simultáneamente, pero las soluciones obtenidas mediante la programación lineal no son invariantes bajo transformaciones lineales de los datos[Thomas, Edelman & Crook, 2002],[He et al., 2010]. Al darse cuenta de estos problemas, algunos investigadores hicieron muchos esfuerzos en este tema. En consecuencia, en este capítulo se estudia un nuevo modelo basado en programación lineal multicriterio y de múltiples niveles de restricción (MC2LP) [Shi, 2010],[Shi et al., 2011].En particular, resuelve el problema de clasificación dos veces. La desviación externa máxima se encuentra en una primera fase, mientras que MC2LP se explota para buscar el hiperplano óptimo basado en la minimización de los dos tipos de error en una segunda fase.

5.2 Programación Lineal Multicriterio y de Múltiples niveles de restricción para la Clasificación

5.2.1 Programación Lineal Multicriterio

La solución de compromiso [He et al., 2004] en la programación lineal multicriterio localiza las mejores soluciones de compromiso entre MMD y MSD para todas las opciones posibles.

$$\begin{cases} \text{mín} \sum_i \xi_i \\ \text{máx} \sum_i \beta_i \\ \text{sujeto a} \\ (x_i \cdot \omega) = b + \xi_i - \beta_i, \quad x_i \in M, \\ (x_i \cdot \omega) = b - \xi_i + \beta_i, \quad x_i \in N, \end{cases} \tag{5.1}$$

donde los x_i son dados, ω es libre, $\xi_i \geq 0$ y $\beta_i \geq 0$, $i = 1, \cdots, n$.

Un valor de frontera b (punto de corte) se utiliza a menudo para separar dos grupos, donde b es libre (sin restricciones). Uno de los problemas causados al tratar a b como una variable es incurrir en muchos casos sin solución. Para algunas aplicaciones, el investigador puede elegir un valor fijo de b $(b = 1)$ para obtener una solución como clasificador. Los esfuerzos para mejorar la tasa de precisión de clasificación se han limitado en gran medida a las características de libre elección de b (es decir, conocido x y dado un determinado valor para b, se resuelve el sistema para encontrar los coeficien-

tes ω) en base a la experiencia del usuario frente al conjunto de datos en tiempo real [He et al., 2010]. En tal procedimiento, el objetivo de encontrar la solución óptima para la pregunta de clasificación se sustituye por la tarea de probar el límite b. Es decir, si b es dado, se puede encontrar un clasificador utilizando una solución óptima al resolver el modelo anterior. Sin embargo, a continuación se señalan los inconvenientes de manejar un modelo de esta manera. En principio, fijar $b = 1$ permitirá soluciones diferentes bajo la suma de vectores (un tipo de transformación lineal). Además la solución cambiará si se cambian las clases de “Bueno” y “Malo” lo que parece ser ilógico. Formalmente, esto significa que las soluciones obtenidas mediante la programación lineal no son invariantes bajo transformaciones lineales de los datos. Esto significa que cuando los datos cambian, no se puede simplemente mantener b como un número fijo, por ejemplo 1, y luego resolver el problema MCLP para diferentes conjuntos de datos. Un enfoque alternativo para resolver este problema es añadir una constante como ζ a todos los valores, pero esto afectará los pesos del resultado y el rendimiento de su clasificación. La adicción de una brecha (espacio) entre las dos regiones puede superar el problema anterior. Sin embargo, si un sujeto cae en esta brecha (espacio), se debe determinar a qué clase debe pertenecer [Thomas, Edelman & Crook, 2002].

Para simplificar el problema, se utiliza una combinación lineal de b^{λ} para reemplazar b. Entonces se puede obtener el mejor clasificador como $\omega^{*}(\lambda)$. En acuerdo a la discusión anterior, un b no fijo es muy importante para afrontar el problema. Al mismo tiempo, por simplicidad y existencia de la solución, b debe fijarse en algún intervalo. Por tanto, ahora se puede suponer que se tiene un límite superior b_u y un límite inferior b_l. En lugar de encontrar la mejor frontera (límite) b aleatoriamente, se puede encontrar la mejor combinación lineal para el mejor clasificador. Es decir, además de considerar el espacio de criterios que contiene el mejor punto de corte de múltiples criterios en (MSD), la estructura de la programación lineal MC2 tiene un espacio de restricción por niveles que muestra todas las posibles ventajas y desventajas de los niveles disponibles de recursos (es decir, la compensación entre el límite superior b_u y el límite inferior b_l). Se puede probar el valor del intervalo tanto para b_u y b_l usando los métodos de interpolación clásica como Lagrange, Newton, Hermite y Golden. No es necesario establecer valores negativos y positivos para b_l y b_u por separado, pero es mejor establecer el valor inicial de b_l como el valor mínimo y el valor inicial de b_u como el valor máximo. Y luego reducir al intervalo $[b_l, b_u]$.

Con el límite ajustado, MSD y MMD se pueden cambiar desde la programación lineal estándar a la programación lineal con múltiples restricciones.

$$\begin{cases} \text{mín} \sum_i \xi_i \\ \text{sujeto a:} \\ (x_i \cdot \omega) \leq \lambda_1 \cdot b_l + \lambda_2 \cdot b_u + \xi_i, \quad x_i \in \mathbf{B}, \\ (x_i \cdot \omega) > \lambda_1 \cdot b_l + \lambda_2 \cdot b_u - \xi_i, \quad x_i \in \mathbf{G}, \\ \lambda_1 + \lambda_2 = 1, \\ 0 \leq \lambda_1, \lambda_2 \leq 1, \end{cases} \tag{5.2}$$

donde x_i, b_u, b_l son dados, ω es libre, $\xi_i \geq 0$.

$$\begin{cases} \text{máx} \sum_i \beta_i \\ \text{sujeto a:} \\ (x_i \cdot \omega) \geq \lambda_1 \cdot b_l + \lambda_2 \cdot b_u - \beta_i, \quad x_i \in \mathbf{B}, \\ (x_i \cdot \omega) < \lambda_1 \cdot b_l + \lambda_2 \cdot b_u + \beta_i, \quad x_i \in \mathbf{G}, \\ \lambda_1 + \lambda_2 = 1, \\ 0 \leq \lambda_1, \lambda_2 \leq 1, \end{cases} \tag{5.3}$$

donde x_i, b_u, b_l son dados, ω es libre, $\beta_i \geq 0$.

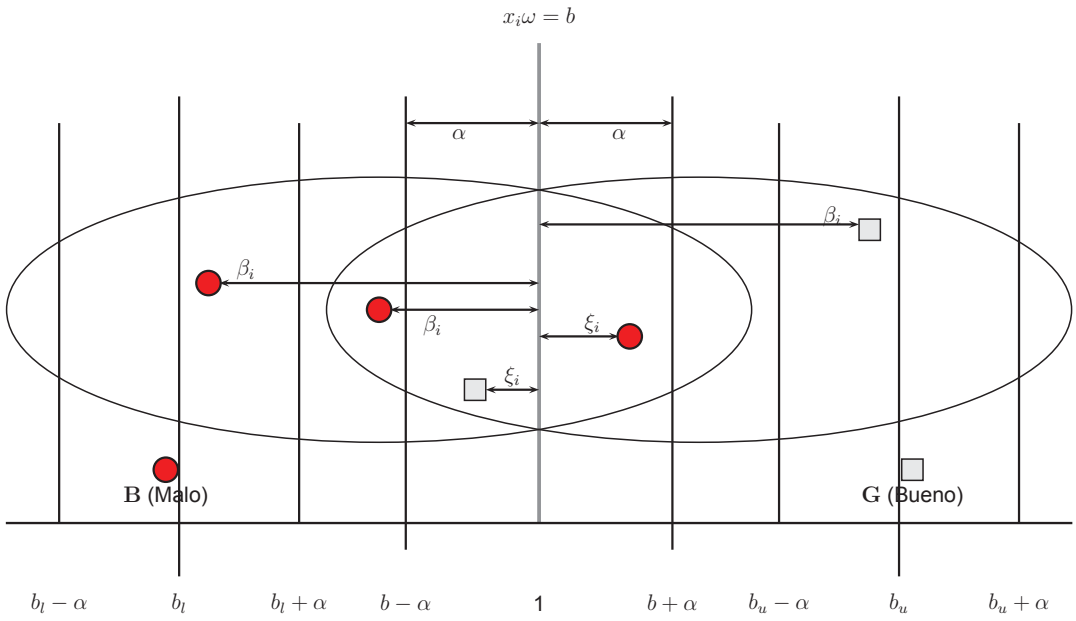
Los dos programas anteriores corresponden a un esquema de programación lineal (PL) con múltiples restricciones. Esta formulación del problema siempre da una solución no trivial.

Un modelo híbrido que combina los modelos de (MSD) y (MMD), con nivel de múltiples restricciones viene dada por:

$$\begin{cases} \text{mín} \sum_i \xi_i \\ \text{máx} \sum_i \beta_i \\ \text{sujeto a:} \\ (x_i \cdot \omega) = \lambda_1 \cdot b_l + \lambda_2 \cdot b_u + \xi_i - \beta_i, \quad x_i \in \mathbf{M}, \\ (x_i \cdot \omega) = \lambda_1 \cdot b_l + \lambda_2 \cdot b_u - \xi_i + \beta_i, \quad x_i \in \mathbf{N}, \\ \lambda_1 + \lambda_2 = 1, \\ 0 \leq \lambda_1, \lambda_2 \leq 1, \end{cases} \tag{5.4}$$

donde x_i, b_u, b_l son dados, ω es libre, $\xi_i \geq 0$ y $\beta_i \geq 0$, $i = 1, 2, \cdots, n$, [He et al., 2010]. Una representación gráfica de este modelo en términos de ξ se muestra en la figura 5.1. Para el modelo (5.4), teóricamente, encontrar la solución ideal que represente simultáneamente los objetivos de maximizar y minimizar es casi imposible. Sin embargo, la

Figura 5.1: Superposición de observaciones para el caso de separación de dos clases



Fuente:Shi Y., Tian Y., Kou G., Peng Y., Li J., Optimization Based Data Mining: Theory and Applications. pag:185. Elaborado por: Autor.

teoría de la Programación Lineal Multicriterio permite estudiar las ventajas y desventajas del espacio de criterios. Para este caso, el espacio de criterios es un plano bidimensional que consta de MSD y MMD. Se utiliza una solución comprometida de múltiples criterios y programación lineal con múltiples restricciones para minimizar la suma de ξ_i y maximizar la suma de β_i simultáneamente. Entonces, el modelo se puede reescribir como:

$$\begin{cases} \text{máx}\, \gamma_1 \sum_i \xi_i + \gamma_2 \sum_i \beta_i \\ \text{sujeto a:} \\ (x_i \cdot \omega) = \lambda_1 \cdot b_l + \lambda_2 \cdot b_u - \xi_i - \beta_i, \quad x_i \in \mathbf{B}, \\ (x_i \cdot \omega) = \lambda_1 \cdot b_l + \lambda_2 \cdot b_u + \xi_i + \beta_i, \quad x_i \in \mathbf{G}, \\ \gamma_1 + \gamma_2 = 1, \\ \lambda_1 + \lambda_2 = 1, \\ 0 \leq \gamma_1, \gamma_2 \geq 1, \\ 0 \leq \lambda_1, \lambda_2 \leq 1. \end{cases} \tag{5.5}$$

Esta formulación del problema siempre da una solución no trivial y es invariante bajo transformaciones lineales de los datos. Tanto γ_1 como γ_2 son los parámetros de los pesos para MSD y MMD. λ_1 y λ_2 son los parámetros de los pesos para b_l y b_u,

[Shi, 2001],[Shi et al., 2005]. Esto sirve para normalizar los niveles de restricción y los parámetros de criterios de nivel.

Observación 2. *. Notar que el punto clave para tratar modelos de clasificación lineal de dos clases es utilizar una combinación lineal de la minimización de la suma de ξ_i y/o la maximización de la suma de β_i para reducir los dos criterios del problema en un solo criterio. La ventaja de esta conversión es para utilizar fácilmente todas las técnicas de la programación lineal para la separación, mientras que la desventaja es que se puede pasar por alto la situación de equilibrio entre estos dos criterios de separación [He et al., 2004].*

5.3 Un nuevo modelo de dos puntos de corte alterables basado en MC2LP

5.3.1 Marco del nuevo modelo MC2LP

Para el modelo original MCLP, un punto de corte se utiliza para predecir la clase de una nueva observación, es decir, sólo existe un único hiperplano. El modelo más formal MC2LP señala que se puede definir dos puntos de corte en lugar de solo el punto de corte original. Y entonces un método sistemático se puede utilizar para resolver este problema. En consecuencia, todas las soluciones posibles en cada nivel de restricción de compensación se pueden adquirir. Sin embargo, un problema es cómo encontrar los puntos de corte, es decir, b_l y b_u.

Por un lado, se utiliza dos puntos de corte para descubrir la solución de mayor precisión. Por otro lado, se espera que los puntos de corte se puedan obtener desde el sistema directamente. Inspirado en la idea anterior, se considera el primer modelo MC2LP, que resuelve el problema de clasificación en dos pasos.

En el primer paso, el modelo de MCLP se utiliza para encontrar el vector de las desviaciones externas ξ_i. Este es una función de λ. Por simplicidad, se fija $b = 1$. Y luego, se fija el parámetro de λ para obtener una posible solución. Ahora se requiere un vector no paramétrico de desviaciones externas ξ. El componente $(\xi_i > 0)$ significa que la observación correspondiente en el conjunto de entrenamiento está mal clasificada. En otras palabras, se producen los errores Tipo I y Tipo II. De acuerdo con la idea de MC2LP, se puede detectar el resultado de cada MCLP fijando el parámetro de γ en cada nivel en el intervalo $[b_l, b_u]$. Ahora, se puede encontrar el componente máximo de ξ:

$$\xi_{\text{máx}} = \text{máx}\{\xi_i,\ 1 \leq i \leq l\} \tag{5.6}$$

De hecho, cuanto menor es el peso de las desviaciones externas, más grande es $\xi_{\text{máx}}$.

Las observaciones mal clasificadas son todas proyectadas en el intervalo $[1-\xi_{\text{máx}}, 1+\xi_{\text{máx}}]$ de acuerdo con el vector de pesos ω obtenido a partir del modelo MCLP. De esta manera, se define b_l y b_u como $1-\xi_{\text{máx}}$ y $1+\xi_{\text{máx}}$, respectivamente. Es fácil ver, si se quiere disminuir el número de los dos tipos de errores, de hecho, sólo se tiene que inspeccionar los puntos de corte mediante la alteración del corte en el intervalo $[1-\xi_{\text{máx}}, 1+\xi_{\text{máx}}]$.

Por otra parte, para la segunda etapa, un nuevo modelo de clasificación MC2LP puede expresarse como sigue:

$$\begin{cases} \text{mín}\, \lambda_1 \sum_i \xi_i - \lambda_2 \sum_i \beta_i \\ \text{sujeto a:} \\ (x_i \cdot \omega) = [1-\xi_{\text{máx}}, 1+\xi_{\text{máx}}] + \xi_i - \beta_i, \quad x_i \in \mathbf{B}, \\ (x_i \cdot \omega) = [1-\xi_{\text{máx}}, 1+\xi_{\text{máx}}] - \xi_i + \beta_i, \quad x_i \in \mathbf{G}, \\ \xi_i, \beta_i \geq 0,\ i = 1, 2, \cdots, l \end{cases} \tag{5.7}$$

donde x_i, $\xi_{\text{máx}}$ son dados, y ω es libre, $[1-\xi_{\text{máx}}, 1+\xi_{\text{máx}}]$ significa cierto equilibrio en el intervalo. Al mismo tiempo, $\lambda = (\lambda_1, \lambda_2)$ es el parámetro elegido en el primer paso.

5.3.2 Discusión del nuevo modelo MC2LP

La modificación más directa del nuevo modelo MC2LP es transferir la única función objetivo a una función objetivo de múltiples criterios. Debido a que el vector de las desviaciones externas es una función de λ, es fácil observar que si el peso entre las desviaciones externas y las desviaciones internas cambia, ξ cambia. En consecuencia, $\xi_{\text{máx}}$ se altera. Y el valor de ξ ideal es el que permite que $\xi_{\text{máx}}$ no sea demasiado grande. En otras palabras, no se espera comprobar que el peso que satisface λ_1 sea demasiado pequeño. En realidad, algunos trabajos han demostrado que sólo si $\lambda_1 > \lambda_2$, entonces $\xi \cdot \beta = 0$, lo que hace significativo al modelo [9]. Como resultado, sólo se tiene que comprobar los parámetros de las funciones objetivas que hacen $\xi_{\text{máx}}$ no demasiado grande, en fin, no demasiado lejos de la original.

Por otro lado, se espera que $\xi_{\text{máx}}$ no sea demasiado pequeño. Es decir, se espera que el modelo tenga alguna generalización. Por lo tanto, dos números pequeños ϵ_1 y ϵ_2 positivos se eligen manualmente. Y luego, el intervalo se ha construido como

$$[[1-\xi_{\text{máx}}-\epsilon_1, 1-\xi_{\text{máx}}+\epsilon_1], [1+\xi_{\text{máx}}-\epsilon_2, 1+\xi_{\text{máx}}+\epsilon_2]]\,.$$

Esto significa que el límite inferior y el límite superior del intervalo debe ser compensado por algunos intervalos, es decir, los niveles de múltiples restricciones son en realidad

intervalos de múltiples restricciones. De hecho, revisando todos los acuerdos de los intervalos es el mismo que comprobar cada acuerdo de $1 - \xi_{\text{máx}} - \epsilon_1$ y $1 + \xi_{\text{máx}} + \epsilon_2$. En este caso, se puede considerar la función objetivo como una función de criterios múltiples, la misma que puede establecerse como sigue:

$$\begin{cases} \text{mín} \sum_i \xi_i \\ \text{máx} \sum_i \beta_i \\ \text{sujeto a:} \\ (x_i \cdot \omega) = [1 - \xi_{\text{máx}} - \epsilon_1, 1 + \xi_{\text{máx}} + \epsilon_2] + \xi_i - \beta_i, \quad x_i \in \mathbf{B}, \\ (x_i \cdot \omega) = [1 - \xi_{\text{máx}} - \epsilon_1, 1 + \xi_{\text{máx}} + \epsilon_2] - \xi_i + \beta_i, \quad x_i \in \mathbf{G}, \\ \xi_i, \beta_i \geq 0, \ i = 1, 2, \cdots, l \end{cases} \tag{5.8}$$

donde x_i, $\xi_{\text{máx}}$, ϵ_1 y ϵ_2 son dados, y ω es libre. Además, ϵ_1 y ϵ_2 son dos números no negativos.

Lema 1. *Para cierto equilibrio entre las funciones objetivo, si b mantiene el mismo signo, entonces los hiperplanos, que se obtienen en el modelo de MCLP, son los mismo. Además, diferentes signos resultan en diferentes hiperplanos.*

Demostración. Se supone que el equilibrio entre las funciones objetivo es $\lambda = (\lambda_1, \lambda_2)$ y ω_1 es la solución obtenida mediante la fijación de b igual a 1. A continuación, sea b_1 un número positivo arbitrario. El modelo MCLP puede transformarse como sigue:

$$\text{mín} \, \lambda_1 \sum_i \xi_i - \lambda_2 \sum_i \beta_i$$

sujeto a:

$$(x_i \cdot \omega) = b_1 + \xi_i - \beta_i, \quad x_i \in \mathbf{B}$$

$$(x_i \cdot \omega) = b_1 - \xi_i + \beta_i, \quad x_i \in \mathbf{G}$$

$$\xi_i, \beta_i \geq 0, \ i = 1, 2, \cdots, l$$

El problema anterior es el mismo que:

$$\text{mín} \, \lambda_1 \frac{\sum_i \xi_i}{b_1} - \lambda_2 \frac{\sum_i \beta_i}{b_1}$$

sujeto a:

$$(x_i \cdot \frac{\omega}{b_1}) = 1 + \frac{\xi_i}{b_1} - \frac{\beta_i}{b_1}, \quad x_i \in \mathbf{B}$$

$$(x_i \cdot \frac{\omega}{b_1}) = 1 - \frac{\xi_i}{b_1} + \frac{\beta_i}{b_1}, \quad x_i \in \mathbf{G}$$

$$\xi_i, \beta_i \geq 0,\ i = 1, 2, \cdots, l$$

y entonces, si se define $\xi_i' = \frac{\xi_i}{b_1}$, $\beta_i' = \frac{\beta_i}{b_1}$, $\omega_i' = \frac{\omega_1}{b_1}$ y el hiperplano $x\omega' = b_1$ es el mismo que $x\omega_1 = 1$.

Del mismo modo, podemos probar que cuando b es un número negativo, la solución es la misma que la que se obtiene de $b = -1$.

Como resultado, sólo se tiene que comparar las soluciones (hiperplanos) resultantes de $b = 1$ y $b = -1$. Para este caso, es fácil ver que los signos delante de ξ_i y de β_i cambian cuando se transforma $b = -1$ en $b = 1$. Si esto sucede, entonces la función objetivo cambia a $-\lambda_1 \sum_i \xi_i + \lambda_2 \sum_i \beta_i$. Esto significa que las soluciones serán diferentes. □

De acuerdo con el lema, tenemos el siguiente teorema:

Teorema 6. *Para el modelo MC2LP (5.8) anterior, de acuerdo con las soluciones (hiperplanos), el espacio γ se divide en dos partes que no se intersecan.*

5.4 Modelo basado en la corrección de los dos tipos de errores

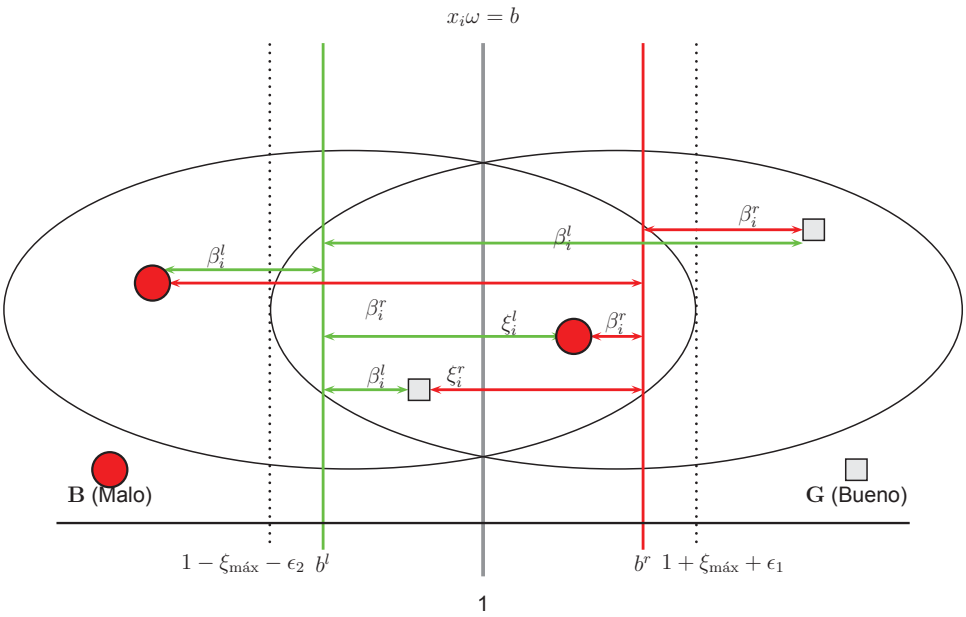
En muchos modelos de clasificación, incluyendo el modelo original MCLP, tratar dos tipos de errores en la clasificación es un gran problema. Por ejemplo, en la clasificación de cuentas de tarjetas de crédito, corregir los dos tipos de errores no sólo puede mejorar la precisión de la clasificación, sino también ayudar a encontrar algunas cuentas importantes. En consecuencia, muchos investigadores se han centrado en este tema. En base a esta consideración, se debe prestar más atención a las muestras que se localicen entre los dos hiperplanos adquiridos por el modelo original MCLP, [10]. Por lo tanto, se definen las desviaciones externas y las desviaciones internas relacionadas con dos hiperplanos diferentes, el de la izquierda y la derecha, es decir, ξ^l, ξ^r, β^l y β^r.

Definición 5. *Las condiciones que las desviaciones deben satisfacer se expresan como sigue:*

$$\xi_i^l = \begin{cases} 0, & (x_i \cdot \omega) < 1 - \xi_{\text{máx}}\ y\ x_i \in \mathbf{B}; \\ (x_i \cdot \omega) - (1 - \xi_{\text{máx}}), & (x_i \cdot \omega) \geq 1 - \xi_{\text{máx}}\ y\ x_i \in \mathbf{B}; \\ 0, & (x_i \cdot \omega) \geq 1 - \xi_{\text{máx}}\ y\ x_i \in \mathbf{G}; \\ (1 - \xi_{\text{máx}}) - (x_i \cdot \omega), & (x_i \cdot \omega) < 1 - \xi_{\text{máx}}\ y\ x_i \in \mathbf{G}; \end{cases}$$

$$\xi_i^r = \begin{cases} 0, & (x_i \cdot \omega) < 1 + \xi_{\text{máx}}\ y\ x_i \in \mathbf{B}; \\ (x_i \cdot \omega) - (1 + \xi_{\text{máx}}), & (x_i \cdot \omega) \geq 1 + \xi_{\text{máx}}\ y\ x_i \in \mathbf{B}; \\ 0, & (x_i \cdot \omega) \geq 1 + \xi_{\text{máx}}\ y\ x_i \in \mathbf{G}; \\ (1 + \xi_{\text{máx}}) - (x_i \cdot \omega), & (x_i \cdot \omega) < 1 + \xi_{\text{máx}}\ y\ x_i \in \mathbf{G}; \end{cases}$$

Figura 5.2: Modelo MC2LP



Fuente: B. Wang, Y. Shi, *Error Correction Method in Classification by Using Multiple-Criteria and Multiple-Constraint Levels Linear Programming*. Elaborado por: Autor.

$$\beta_i^l = \begin{cases} (1 - \xi_{\text{máx}}) - (x_i \cdot \omega), & (x_i \cdot \omega) < 1 - \xi_{\text{máx}} \; y \; x_i \in \mathbf{B}; \\ 0, & (x_i \cdot \omega) \geq 1 - \xi_{\text{máx}} \; y \; x_i \in \mathbf{B}; \\ (x_i \cdot \omega) - (1 - \xi_{\text{máx}}), & (x_i \cdot \omega) \geq 1 - \xi_{\text{máx}} \; y \; x_i \in \mathbf{G}; \\ 0, & (x_i \cdot \omega) < 1 - \xi_{\text{máx}} \; y \; x_i \in \mathbf{G}; \end{cases}$$

$$\beta_i^r = \begin{cases} (1 + \xi_{\text{máx}}) - (x_i \cdot \omega), & (x_i \cdot \omega) < 1 + \xi_{\text{máx}} \; y \; x_i \in \mathbf{B}; \\ 0, & (x_i \cdot \omega) \geq 1 + \xi_{\text{máx}} \; y \; x_i \in \mathbf{B}; \\ (x_i \cdot \omega) - (1 + \xi_{\text{máx}}), & (x_i \cdot \omega) \geq 1 + \xi_{\text{máx}} \; y \; x_i \in \mathbf{G}; \\ 0, & (x_i \cdot \omega) < 1 + \xi_{\text{máx}} \; y \; x_i \in \mathbf{G}; \end{cases}$$

La Figura 5.2 representa un boceto para el modelo en discusión. En la gráfica, las líneas verde y roja son el hiperplano izquierda y derecho, b^l y b^r respectivamente, las cuales representan un equilibrio de los dos intervalos, es decir, $[1-\xi_{\text{máx}}-\epsilon_2, 1]$ y $[1, 1+\xi_{\text{máx}}+\epsilon_1]$. Y todas las desviaciones se miden de acuerdo con estos en diferentes colores. Por ejemplo, si una observación de la clase "bueno" está mal clasificado como en la clase "Malo", esto significa que $\xi_i^r > \beta_i^l \geq 0$ y $\xi_i^l = \beta_i^r = 0$; y si por el contrario una observación de la clase "Malo" está mal clasificado en la clase "Bueno", esto significa que $\xi_i^l > \beta_i^r \geq 0$ y $\xi_i^r = \beta_i^l = 0$. Por lo tanto, para aquellas observaciones mal clasificadas $\xi_i^r + \xi_i^l - \beta_i^r - \beta_i^l$

debe ser minimizado.

Como resultado, un modelo más minucioso podría expresarse como sigue:

$$
\begin{cases}
\text{mín} \sum_i (\xi_i^r + \xi_i^l) \\
\text{mín} \sum_i (\xi_i^l - \beta_i^r) \\
\text{mín} \sum_i (\xi_i^r - \beta_i^l) \\
\text{máx} \sum_i (\beta_i^r + \beta_i^l) \\
\text{sujeto a:} \\
(x_i \cdot \omega) = 1 + [0, \xi_{\text{máx}} + \epsilon_1] + \xi_i^r - \beta_i^r, & x_i \in \mathbf{B}, \\
(x_i \cdot \omega) = 1 - [0, \xi_{\text{máx}} + \epsilon_2] + \xi_i^l - \beta_i^l, & x_i \in \mathbf{B}, \\
(x_i \cdot \omega) = 1 + [0, \xi_{\text{máx}} + \epsilon_1] - \xi_i^l + \beta_i^l, & x_i \in \mathbf{G}, \\
(x_i \cdot \omega) = 1 - [0, \xi_{\text{máx}} + \epsilon_2] - \xi_i^l + \beta_i^l, & x_i \in \mathbf{G}, \\
\xi_i^r, \xi_i^l, \beta_i^r, \beta_i^l \geq 0, & i = 1, 2, \cdots, l.
\end{cases}
\tag{5.9}
$$

donde x_i, $\xi_{\text{máx}}$, $\epsilon_1 > 0$, $\epsilon_2 > 0$ son dados y ω es libre de restricciones. En la Figura 5.2, para cada punto, hay como máximo dos tipos de desviaciones distintas de cero. Las funciones objetivas aparecen para hacer frente a las desviaciones de acuerdo con la posición que se muestra en la Figura 5.2, respectivamente, mientras que tienen su propio significado especial. Esto quiere decir, mide dos tipos de error en algún grado por medio de las funciones objetivos segunda y tercera. Como resultado, en esta nueva versión de MC2LP, no sólo se consideran las desviaciones respectivamente, sino también toma la relación de las desviaciones sobre la base de los dos tipos de error en una cuenta en las funciones objetivo. En virtud del método MC2LP, cada compromiso entre $1 - \xi_{\text{máx}} - \epsilon_2$ y 1 para el hiperplano izquierdo, así como cada compromiso entre 1 y $1 + \xi_{\text{máx}} + \epsilon_1$ para el hiperplano derecho, pueden ser comprobadas.

Después de obtener el vector de pesos ω del hiperplano, $(x \cdot \omega) = 1$ todavía se utiliza como el hiperplano de clasificación. Sin embargo, en el nuevo modelo, se minimiza la distancia entre el hiperplano izquierdo y el derecho.

En realidad, en las estadísticas, los errores del Tipo I y los errores del Tipo II son dos objetivos opuestos. Es decir, es muy difícil de corregir ambos al mismo tiempo. Como resultado, se modifica el modelo anterior en dos modelos diferentes que se centran en dos tipos de error, respectivamente, como sigue:

$$
\begin{cases}
\text{mín} \sum_i (\xi_i^r + \xi_i^l) \\
\text{mín} \sum_i (\xi_i^l - \beta_i^r) \\
\text{máx} \sum_i (\beta_i^r + \beta_i^l) \\
\text{sujeto a:} \\
(x_i \cdot \omega) = 1 + [0, \xi_{\text{máx}} + \epsilon] + \xi_i^r - \beta_i^r, & x_i \in \mathbf{B}, \\
(x_i \cdot \omega) = 1 + \xi_i^l - \beta_i^l, & x_i \in \mathbf{B}, \\
(x_i \cdot \omega) = 1 + [0, \xi_{\text{máx}} + \epsilon] - \xi_i^r + \beta_i^r, & x_i \in \mathbf{G}, \\
(x_i \cdot \omega) = 1 - \xi_i^l + \beta_i^l, & x_i \in \mathbf{G}, \\
\xi_i^r, \xi_i^l, \beta_i^r, \beta_i^l \geq 0, & i = 1, 2, \cdots, l.
\end{cases} \tag{5.10}
$$

donde x_i, $\xi_{\text{máx}}$ y $\epsilon > 0$ son dadas, y ω es libre de restricciones. En este modelo, $\sum_i (\xi_i^r - \beta_i^l)$ no está contenida en las funciones objetivo. Este modelo puede lidiar con el error tipo II, es decir, la clasificación de un punto “bueno” como uno “malo”.
Como el resultado mostrado anteriormente, el modelo (5.10) puede corregir el error de Tipo II en algún grado. Se concluye esto en la siguiente proposición.

Proposición 7. *El Modelo (5.10) puede corregir el error tipo II moviendo el hiperplano derecho a la derecha sobre la base del concepto de múltiples niveles de restricción.*

Observación 3. *La segunda función objetivo en el modelo (5.10) es distinta de cero para las observaciones de la clase “Malo”, y es negativo cuando el hiperplano derecho se mueve hacia la derecha. Es decir, se toleran algunos errores del tipo I. Al mismo tiempo, la primera función objetivo en el modelo (5.10) permite un castigo creciente de los errores de Tipo II con el movimiento del hiperplano derecho hacia la derecha. Como resultado, se puede corregir el error de Tipo II en algún grado.*

Al igual que en el modelo (5.10), se plantea el modelo (5.11) que puede lidiar con el error tipo I de la siguiente manera:

$$
\begin{cases}
\text{mín} \sum_i (\xi_i^r + \xi_i^l) & \\
\text{mín} \sum_i (\xi_i^r - \beta_i^l) & \\
\text{máx} \sum_i \beta_i^r + \beta_i^l & \\
\text{sujeto a:} & \\
(x_i \cdot \omega) = 1 + \xi_i^r - \beta_i^r, & x_i \in \mathbf{B} \\
(x_i \cdot \omega) = 1 - [0, \xi_{\text{máx}} + \epsilon] + \xi_i^l - \beta_i^l, & x_i \in \mathbf{B} \\
(x_i \cdot \omega) = 1 - \xi_i^r + \beta_i^r, & x_i \in \mathbf{G} \\
(x_i \cdot \omega) = 1 - [0, \xi_{\text{máx}} + \epsilon] - \xi_i^l + \beta_i^l, & x_i \in \mathbf{G} \\
\xi_i^r, \xi_i^l, \beta_i^r, \beta_i^l \geq 0, & i = 1, 2, \cdots, l.
\end{cases}
\tag{5.11}
$$

donde x_i,$\xi_{\text{máx}}$ y $\epsilon > 0$, y ω es libre de restricciones. En este modelo, $\sum_i (\xi_i^l - \beta_i^r)$ no está dentro de las funciones objetivo. Este modelo puede lidiar con el error tipo I, es decir, la clasificación de un punto “malo” como uno “bueno”.

Capítulo 6

APLICACIÓN: CALIFICACIÓN DE RIESGO DE CRÉDITO

En el presente capítulo se aplican los resultados teóricos que se presentaron en los capítulos anteriores, utilizando datos reales de una Institución Microfinanciera (IMF). Para comenzar se establecerán las reglas con las que los créditos funcionan en la mayoría de las IMFs de nuestro país, lo que en la práctica se conoce como operación de los créditos. Estos precedentes permitirán definir eventos, entre ellos el incumplimiento, asociados al comportamiento de los créditos, que será modelado como un problema de clasificación.

6.1 Introducción

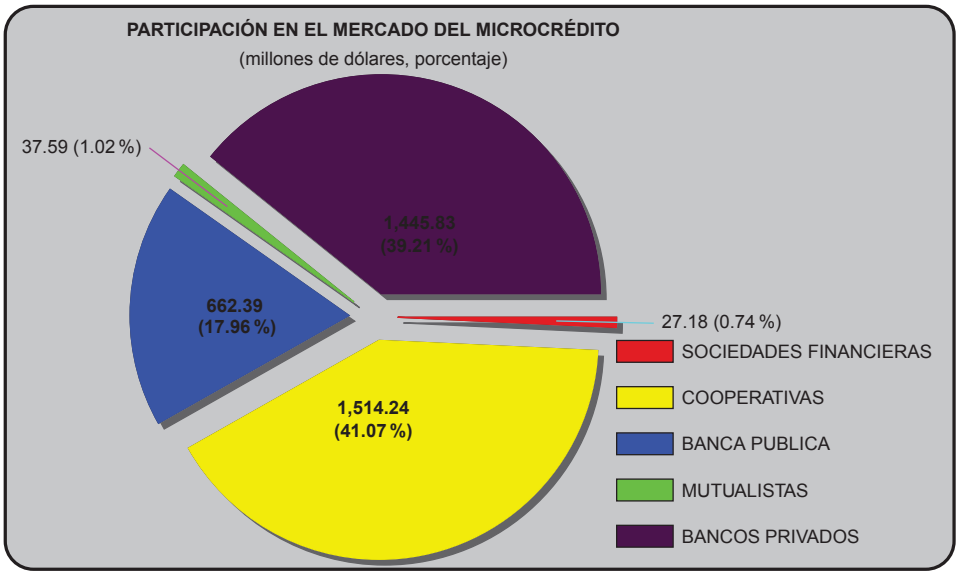
Las microfinanzas, se refieren a la provisión de servicios financieros tales como: préstamos, ahorro, seguros o transferencias de recursos hacia hogares con bajos ingresos o hacia actividades u organizaciones económicas cuya administración se encuentra bajo una persona o grupo de personas emprendedoras, que se han organizado para por medio de la autogestión, lograr objetivos económicos que les permita mejorar su calidad de vida.

Las entidades dedicadas a proveer este servicio lo hacen principalmente a través del denominado microcrédito, es decir, préstamos pequeños que permiten a las personas u organizaciones que no cumplen con el requisito de una garantía real, iniciar o ampliar su propio emprendimiento y por tanto, aumentar sus ingresos.

Esta actividad, que antes era exclusiva del Estado o de instituciones no formales, tiene actualmente la intervención de variadas instituciones especializadas reguladas por la Superintendencia de Bancos (SB), la Superintendencia de Economía Popular y Solidaria (SEPS), de las no reguladas, que están bajo el control del Ministerio de Inclusión Económica y Social (MIES) a través de la Dirección Nacional de Cooperativas.

En el año 2002 existían 14 entidades supervisadas por la Superintendencia de Bancos y Seguros que proporcionaban servicios microfinancieros. A Diciembre de 2014 éstas ascienden a 74 instituciones financieras dedicadas a este negocio, entre las que se encuentran: 23 bancos privados, 39 cooperativas, 9 sociedades financieras, 4 mutualistas y 2 entidades públicas (Banco Nacional de Fomento y Corporación Financiera Nacional), evidenciándose la importancia que el sistema financiero le ha dado a este sector cada vez con más presencia en el mercado. A partir de Enero de 2013 La Superintendencia de la Economía Popular y Solidaria es la entidad reguladora de las Cooperativas de Ahorro y Crédito.

Figura 6.1: Participación en el mercado del microcrédito



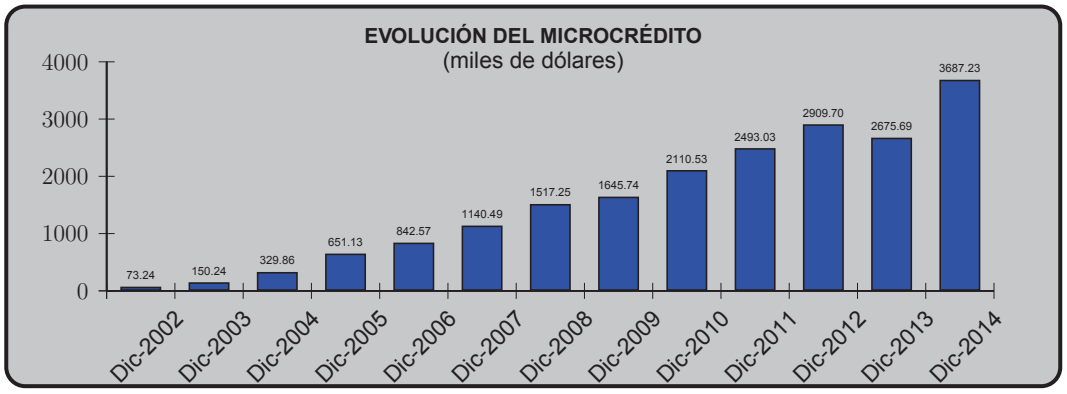
Fuente: Superintendencia de Bancos, Superintendencia de la Economía Popular y Solidaria.
Elaborado por: Autor

El total de la cartera bruta de microempresa a Diciembre de 2014 ascendió a 3,687.23 millones de dólares, superior en 1,011.54 millones de dólares (37.80 %) a lo entregado en Diciembre de 2013. A la banca privada le corresponde USD 1,445.83 millones, es decir el 39.21 % del total entregado para el sector de microempresa, siendo el Banco Pichincha el más representativo con USD 721.09 millones que concentra el 49.87 % en el total concedido por sistema de bancos, le sigue Solidario con una cartera de USD 292.08 millones correspondiente a 20.20 % y Procredit con USD 93.29 millones que constituye el 6.45 %. Por parte de las Cooperativas, éstas aportaron con USD 1,514.24 millones, participando del 41.07 % del total del sector, siendo las más representativas Juventud Ecuatoriana Progresista (JEP) con USD 167.96 millones (11.10 %), San Francisco con USD 122.25 millones (8.08 %) y Mushuc Runa con USD 111.62 millones (7.37 %). En cuanto a la Banca Pública, este subsistema otorgó USD 662.3 millones que equivale

al 17.96 % de participación en el sector, destacándose el Banco Nacional de Fomento con USD 649.24 millones, equivalente al 98.01 % de contribución dentro del subsistema de banca pública. En lo que corresponde a Sociedades Financieras y Mutualistas, estos subsistemas tienen una participación marginal como proveedores de este producto, correspondiéndole al primer subsistema la suma de USD 27.18 millones, es decir un 0.74 % dentro del sector, mientras que Mutualistas concedió un total de USD 37.59 millones participando con apenas el 1.02 %. Dentro del grupo de Sociedades Financieras, Unifinsa abarca el 66.94 %, mientras que en Mutualistas se destaca Mutualista Pichincha con el 83.33 % dentro de su grupo.

Analizando por entidad se desprende que al Banco Pichincha le corresponde el 19.6 % de estos recursos colocados, le sigue el Banco Nacional de Fomento con el 17.6 %, Banco Solidario con el 7.9 %, y finalmente Juventud Ecuatoriana Progresista con el 4.7 %. Es decir que, tomando en cuenta las cuatro principales entidades que tienen el producto de microfinanzas, la banca privada abarca el 27.5 % mientras que el subsistema de Cooperativas tan solo lo hace con el 4.6 %. Por tanto de las 77 entidades que hacen microcrédito, 4 entidades concentran con el 49,6 % de la cartera, en tanto que 73 participan del 50,4 %.

Figura 6.2: Evolución del microcrédito en el Ecuador, en el periodo 2002-2014



Fuente: Superintendencia de Bancos, Superintendencia de la economía popular y solidaria. Elaborado por: Autor

6.1.1 Descripción general del portafolio

La forma en la que la mayoría de las IMFs operan sus créditos, comienza cuando se desembolsa un importe monetario **C** (*monto otorgado*) y le es otorgado a una o varias personas que han sido elegidos como viables (*deudores*). La evaluación de la viabilidad, se realiza con un análisis del flujo de efectivo de su negocio (existente o potencial), por

parte de los promotores de crédito (*oficiales de crédito*). Posteriormente, se establece un periodo de tiempo en el que el cliente adquiere el compromiso de pagar **C**, y otro importe monetario *I* llamado interés. La duración en meses **M** de este periodo de tiempo llamado *plazo*, así como un factor **a** conocido como la tasa, sirven para determinar el importe de intereses

$$I = aMC \tag{6.1}$$

La deuda adquirida por el cliente será saldada a través de **K** pagos (*dividendos*) conocidos como amortizaciones, con valor $C(1 + aM)/K$ cada una. Las fechas calendario correspondientes a cada una de estas fechas son distribuidas de manera homogénea en el plazo, considerando una frecuencia mensual, trimestral, semestral, anual que sea consistente con los flujos de efectivo del cliente, estas son las fechas de exigibilidad del crédito.

Si en alguna fecha calendario posterior a una o más de las de exigibilidad, el dividendo correspondiente no es cubierto, la diferencia en días entre estas dos fechas es llamada *días de atraso del dividendo*. Dado que en algún momento más de un dividendo puede tener días de atraso, el máximo de los mismos es llamado *días de atraso del crédito*. La suma del importe no cubierto de los dividendos exigibles, más las que aún no lo son, es igual a la *exposición al riesgo*. Si más de un dividendo posee días de atraso y el cliente realiza un pago, este servirá para cubrir aquella con la fecha de exigibilidad más antigua, con lo que cada pago puede reducir los días de atraso. Cuando todas las amortizaciones de un crédito son cubiertas, se dice que éste se encuentra *Saldado o Cancelado*. Esto puede ocurrir durante o después del plazo pactado (con sus respectivos días de atraso), con lo que los compromisos del cliente con la IMF terminarían. Existe la posibilidad de que el crédito no sea saldado y acumule días de atraso, si éstos llegan a superar los 180 (para la mayoría de las IMFs), se dice que el crédito ha caído en incumplimiento y se encuentra en un *estatus de Castigado*. En ese momento la IMF asume el valor EADn como una pérdida, debido principalmente a que; la responsabilidad de cobro de los créditos no castigados corre a cargo de los mismos promotores que lo originaron, pero cuando el crédito se castiga resulta poco viable que esta gestión siga siendo llevada a cabo por los mismos, pues si no hubo voluntad de pago durante seis meses, muy probablemente no ocurra en lo posterior y puede mejor dedicarse a originar un nuevo crédito que si genere ingresos. La responsabilidad de cobro de créditos castigados se delega al área legal.

De acuerdo al funcionamiento descrito, el tiempo que las IMFs dedican (sin llegar a instancias legales) a cada uno de sus créditos, también conocido como *horizonte o vida del crédito*, estará sujeto al tiempo que tarde en saldarse o bien en castigarse, lo que

Tabla 6.1: Categorías de Calificación de los Microcréditos

CATEGORÍAS DE RIESGO	DÍAS DE MOROSIDAD
A1	0
A2	1-8
A3	9-15
B1	16-30
B2	31-45
C1	46-70
C2	71-90
D	91-120
E	mayor a 120

Fuente: Superintendencia de Bancos
Elaborado por: Autor

pase primero. Mientras esto no ocurra diremos que el crédito se encuentra activo.

Los plazos a los que las IMFs otorgan sus créditos varían entre los 4 y 12 meses, por lo que los horizontes de los mismos varían entre los 11 y los 19, que en comparación con los créditos que se otorgan en otro tipo de instituciones financieras, estos periodos de tiempo son cortos. Por ejemplo, las tarjetas de crédito poseen un período de vida de 3 o más años, o los créditos hipotecarios que duran desde 5 hasta 30 años; incluso los créditos comerciales, que en general son superiores a los 18 meses. Estos plazos cortos de las microfinanzas son explicados principalmente por los montos tan pequeños que se otorgan, así como, para reducir el riesgo de la falta de flujo de efectivo del cliente en un plazo mayor. La distribución de los plazos tanto en número de créditos como en exposición al riesgo en el portafolio de la IMF, con cuyos datos se trabajó esta aplicación, varían entre los 12 y 63 meses. Los plazos mayores a 66 meses y menores a 12 meses cuentan con una participación significativamente menor.

Debido a la amplia variedad que un portafolio posee en cuanto a los días de atraso de cada uno de sus créditos, es común agruparlos en categorías de riesgo, con lo que además puede afinarse el conocimiento del porcentaje de operaciones en atraso. Las categorías de riesgo son una propiedad para cada una de las operaciones en alguna fecha de evaluación, y son asignados de acuerdo a la tabla (6.1)

6.2 Metodología

En esta sección se detallan las especificaciones técnicas utilizadas, así como el desarrollo y resultados de los modelos de Clasificación estadístico-matemático para el portafolio de Microcréditos en la IMF.

En la presente sección se expondrán los resultados para los modelos de clasificación desarrollados, pero primeramente se partirá con la descripción de la base de datos requerida.

6.2.1 Base de Datos

Para la presente aplicación se considera como referente la base del historial crediticio de la IMF con fecha corte al 31/12/2014, cuya estructura es la siguiente: cartera comercial 21.00 % que corresponde a 62,604 operaciones, cartera microcrédito con 78.26 % que equivale a 276,090 operaciones y la cartera de consumo con el 0.74 % de la cratera total, que corresponde a 2,625 operaciones. Por lo tanto, en adelante nuestro estudio se centra en la cartera de microcrédito que es la más representativa de la cartera total de la IMF, con 272,539 operaciones originales, de la misma que el 52.52 % (143,151 operaciones) corresponden al segmento del bono de desarrollo humano. De esta manera la población de estudio se reduce al 47.48 % que equivale a 129,388 operaciones.

Para el caso de microcréditos, se han venido utilizando dos tipos de formularios de evaluación financiera para la unidad familiar y productiva: uno para el sector Agropecuario y, otro para el resto de sectores. Así, en función de este concepto la base se ha dividido en dos segmentos:

- Microagropecuario.- Operaciones cuyo sector es: pecuario, agrícola, agroindustrial, acuícola, piscicultura y forestal. (Base con 70,374 registros, ver figura 6.3)
- Microempresa.- resto de sectores. (base con 59,014 registros).

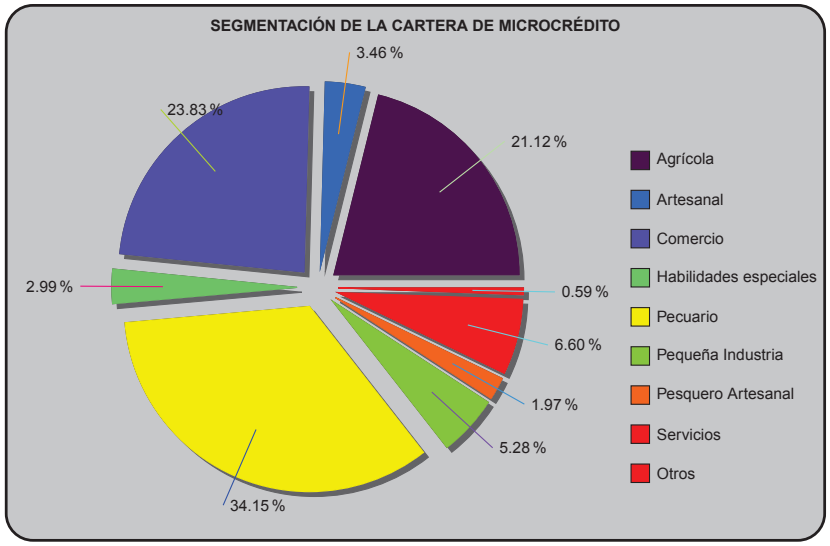
El periodo analizado considera datos de operaciones de microcrédito del IMF concedidas desde enero de 2012 hasta diciembre 2014. La fecha de corte es 31 de diciembre 2014, misma que se considera para el análisis de la morosidad. El número de registros es de 59,382 registros que corresponden a operaciones originales del sector microagropecuario.

En el desarrollo de la presente aplicación, se utilizó información sociodemográfica del cliente, información de las operaciones y del comportamiento crediticio que el cliente mantiene con la IMF y con las instituciones financieras reguladas por la SB. Este requerimiento genera una dicotomía entre dos elementos de decisión:

a) la muestra debe ser representativa de aquellos clientes (clientes potenciales) que deseen aplicar a una línea de crédito en el futuro (throught-the-door-population),

b) la muestra debe incorporar información suficiente acerca de las diferentes conductas de pago de los clientes.

Figura 6.3: Portafolio de Microcrédito



Fuente: Registro de operaciones de la IMF al 31 de Diciembre de 2014. Elaborado por: Autor

Es aquí justamente donde se genera el conflicto, puesto que con esta muestra se necesita definir el criterio de cliente bueno y cliente malo a emplear para encontrar las características relevantes en el modelo de clasificación. El periodo de observación es el tiempo t en el que el investigador decide situarse y observar el desempeño del cliente. Es este periodo de performance (desempeño) el que se emplea para predecir el comportamiento futuro de los potenciales clientes. En el punto del resultado se asigna una calificación (bueno o malo) al cliente con base en el resumen del comportamiento en el periodo de desempeño. Por tanto, la importancia de la madurez de la cartera para no calificar como bueno a un cliente que es malo, pero que no logra denotar un comportamiento porque inicia a pagar su obligación. Para lograr tal propósito es útil definir un indicador de mora o tasa de morosidad:

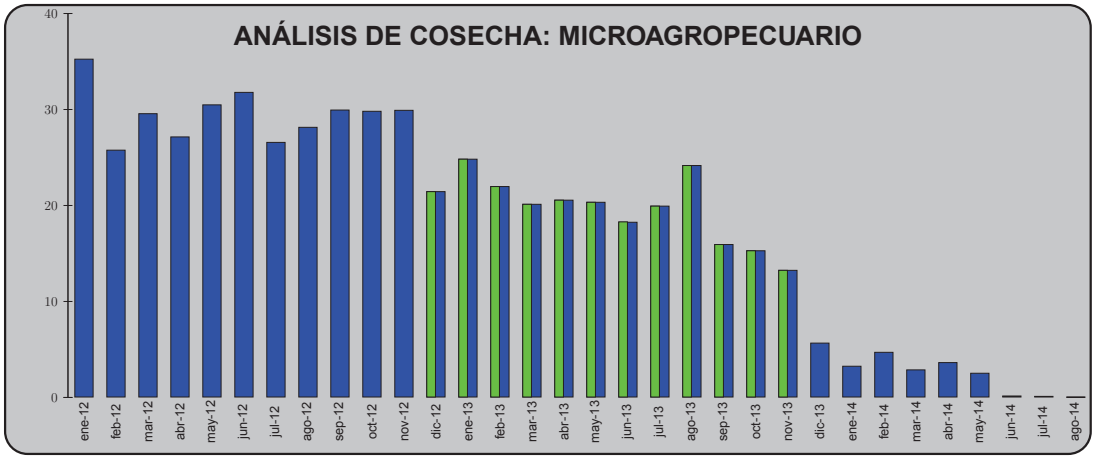
$$\text{Tasa de morosidad}_t = \frac{\text{número de clientes malos del mes de desembolso t}}{\text{número total de clientes del mes de desembolso t}}$$

$$= \frac{\text{número de clientes del mes de desembolso t con atraso máximo mayor a 30 días}}{\text{número total de clientes del mes de desembolso t}}$$

La tasa de morosidad se construye por periodo de cosecha (fecha de venta de las operaciones de crédito) y tiene por objeto mostrar en forma gráfica y por mes de colocación la relación clientes malos sobre total de clientes, con el fin de señalar los periodos en que esta razón se estabiliza como equivalentes a un comportamiento estable de la cartera. Un periodo se considera como estable si la razón de la cosecha t presente pequeñas

variaciones en relación a la cosecha $t-1$ y $t+1$. En términos estadísticos se puede argumentar que se necesita, durante la ventana temporal elegida, que la razón tasa de morosidad siga una distribución uniforme. La elección de periodos con tasas de morosidad decrecientes no implica necesariamente una mejora en el comportamiento de la cartera, puede ser que estas cosechas por ser cercanas a la fecha actual estén reflejando carteras poco maduras y por ende no comparables con otros periodos de análisis. A este tipo de estudios se los conoce como análisis de cosechas.

Figura 6.4: Análisis de Cosechas del sector microagropecuario: ene12 - dic14



Fuente: Registro de operaciones de la IMF al 31 de Diciembre de 2014. Elaborado por: Autor

6.2.2 Selección de la ventana de muestreo.

A modo de ejemplo se puede observar en la figura 6.4 del análisis de cosecha que para enero de 2013 el indicador de mora es del 24.87 % que corresponde al desembolso de alrededor de 2199 créditos, de los cuales 547 han presentado morosidad máxima; así se puede leer para cada fecha de contabilización la tasa de morosidad del total de créditos colocados, sin embargo al acercarnos a fechas más recientes se aprecia una mejora en el comportamiento de los clientes, pues la tasa de morosidad decrece a niveles del 10 %, el error consiste en creer que esto se debe a mejoras en las cosechas, cuando lo que realmente está ocurriendo es que falta maduración o tiempo de observación y por ende de desempeño que permita comparar con otros periodos de análisis. A continuación se presenta el gráfico y prueba de distribución uniforme:
La muestra para el caso del sector Agropecuario corresponde a las operaciones concedidas entre diciembre del 2012 y noviembre del 2013, lo cual representa 25,085 registros; después de un análisis de la completitud de la data, la base depurada contiene 23,468 registros.

Resumen de prueba de hipótesis

	Hipótesis nula	Test	Sig.	Decisión
1	La distribución de tasa_de_mora es uniforme con el mínimo 0.13 y el máximo 0.25.	Prueba Kolmogorov-Smirnov de una muestra	.472	Retener la hipótesis nula.

Se muestran las significancias asintóticas. El nivel de significancia es .05.

6.2.3 Definición de la variable dependiente

En esta sección el objetivo es definir una partición de la cartera de clientes, con un criterio adecuado que permita identificar a los clientes buenos y malos, a partir de la definición de la variable dependiente; por lo tanto, una correcta clasificación a priori dará una correcta clasificación a posteriori.

Para la construcción del indicador buenos/malos (Indicador B/M), se utiliza los indicadores de altura de mora: atraso promedio y atraso máximo, en el que han incurrido los clientes, previamente se eliminan las operaciones castigadas y condonadas resultando en una base de 22,882 registros para el presente estudio. A partir de lo cual se obtiene la matriz de confusión de las variables atraso promedio y atraso máximo como se observa en la tabla 6.2:

Tabla 6.2: Matriz Atraso Promedio/ Atraso Máximo

ATRASO PROMEDIO	ATRASO MÁXIMO										TOTAL
	RANGOS	0	1-8	9-15	16-30	31-45	46-70	71-90	91-120	+120	
	0	4208									4208
	1-8		9293	1013	626	100	6				11038
	9-15			621	734	250	82	8	1		1696
	16-30				814	411	408	101	35	4	1773
	31-45					369	177	162	94	39	841
	46-70						464	98	181	192	935
	71-90							208	46	190	444
	91-120								240	209	449
	+120									1498	1498
TOTAL		4208	9293	1634	2174	1130	1137	577	597	2132	22882

Fuente: Registro Crediticio de la IMF. Elaborado por: Autor

En la tabla 6.2 se distribuyen los clientes según el cruce que les corresponde por rango de atraso promedio (filas) y rango de atraso máximo (columnas), por ejemplo; existen 1013 clientes cuyo atraso promedio está entre 1 y 8 días y con atraso máximo entre 9 y 15 días. La entidad financiera debe establecer un criterio del valor de pérdida que es bueno o malo para la institución. Considerando los niveles de pérdida establecidos por la Superintendencia de Bancos se concluyó que los clientes con atraso promedio menor o igual a 15 días y atraso máximo de hasta 15 días son clientes buenos, casillas

Tabla 6.3: Matriz Atraso Promedio/ Atraso Máximo (Porcentaje)

ATRASO PROMEDIO	ATRASO MÁXIMO									TOTAL
	RANGOS	0	1-8	9-15	16-30	31-45	46-70	71-90	91-120	+120	
	0	18.39 %									18.39 %
	1-8		40.61 %	4.43 %	2.74 %	0.44 %	0.03 %				48.24 %
	9-15			2.71 %	3.21 %	1.09 %	0.36 %	0.03 %			7.41 %
	16-30				3.56 %	1.80 %	1.78 %	0.44 %	0.15 %	0.02 %	7.75 %
	31-45					1.61 %	0.77 %	0.71 %	0.41 %	0.17 %	3.68 %
	46-70						2.03 %	0.43 %	0.79 %	0.84 %	4.09 %
	71-90							0.91 %	0.20 %	0.83 %	1.94 %
	91-120								1.05 %	0.91 %	1.96 %
	120									6.55 %	6.55 %
TOTAL		18.39 %	40.61 %	7.14 %	9.50 %	4.94 %	4.97 %	2.52 %	2.61 %	9.32 %	100.00 %

Fuente: Registro Crediticio de la IMF. Elaborado por: Autor

en color verde. Sin embargo los clientes con atraso promedio entre 1 y 15 días, pero con atrasos máximos entre 16 y 120 días no pueden ser determinados como buenos o malos; a este grupo se le denominará clientes indeterminados, puesto que no se tiene la información suficiente a cerca de ellos como para ubicarlos en un grupo definido. los clientes indeterminados están representados por la casillas en color amarillo. Los clientes restantes son los clientes clasificados por el criterio como malos, casillas rojas. La tabla 6.4 a continuación muestra los porcentajes de clientes en cada categoría:

Tabla 6.4: Distribución de clientes

	NÚMERO	PORCENTAJES (%)
BUENOS	15135	66.94 %
INDETERMINADOS	1844	7.90 %
MALOS	5940	25.96 %
TOTAL	22882	100.00 %

Fuente: Registro Crediticio de la IMF. Elaborado por: Autor

Se aprecia que hay una alta incidencia de clientes malos, que representan un 25.96 % de la cartera, los clientes indeterminados constituyen el 7.90 % y el 66.14 % restante son los clientes buenos. A nivel de consultoras se ha definido un estándar para la definición de buenos y malos, en el que se indica que la distribución es adecuada si mantiene un porcentaje de clientes indeterminados no mayor a 12 %. Adicionalmente, la definición de buenos y malos dependerá de la acidez con que se quiera realizar el modelo.

La relación de los clientes buenos y malos es de aproximadamente 2 a 1, al excluir el grupo de clientes de la categoría indeterminados.

La distribución de buenos y malos en la muestra de construcción se mantiene, como se puede observar en la tabla 6.5:

Tabla 6.5: Distribución de clientes en la muestra

	NÚMERO	PORCENTAJES (%)
BUENOS	12081	66.00 %
INDETERMINADOS	1445	7.89 %
MALOS	4779	26.11 %
TOTAL	18305	100.00 %

Fuente: Registro Crediticio de la IMF. Elaborado por: Autor

Pues la representatividad que tienen las muestras de construcción y de back testing debe mantenerse en las variables, por ejemplo, en la tabla 6.6, se observa la distribución según la variable zonal.

Tabla 6.6: Distribución clientes por zonal

ZONAL	Distribución de operaciones en la muestra total (%)	Distribución de operaciones en la muestra de construcción (%)	Distribución de operaciones en la muestra de back testing (%)
ZONAL CUENCA	5.06	5.05	5.10
ZONAL EL PUYO	16.32	16.26	16.56
ZONAL GUAYAQUIL	5.41	5.42	5.36
ZONAL LOJA	13.69	13.57	14.17
ZONAL MACHALA	3.82	3.80	3.87
ZONAL PORTOVIEJO	14.25	14.30	14.06
ZONAL QUITO	7.75	7.90	7.14
ZONAL RIOBAMBA	25.41	25.45	25.21
ZONAL STO. DOMINGO	8.29	8.24	8.51
TOTAL	100.00	100.00	100.00

Fuente: Registro Crediticio de la IMF. Elaborado por: Autor

Sin embargo considerando que el objetivo final del modelo es diferenciar a los clientes buenos de los malos, se excluirán de la construcción a los clientes indeterminados, ya que estos sesgarían el modelo. Una vez obtenido el indicador de buenos y malos (indicador B/M), este se constituye como la variable dependiente del modelo. La variable dependiente es entonces dicotómica, donde los buenos se representarán con 1 (uno) y los malos con 0 (cero).

6.2.4 Análisis descriptivo de la base de datos

En esta sección se escogen las variables que aportan información al modelo, como una primera aproximación se realizará el análisis exploratorio de la data.

La base de datos utilizada en este proyecto está conformada por información sociodemográfica del cliente, información de comportamiento crediticio interna y la información externa como es la información del buró de crédito, esta última obtenida de la central de riesgo, que corresponde exclusivamente al comportamiento del cliente en instituciones diferentes a la entidad para la que se realiza el modelo. En la tabla 6.7 se muestra una descripción de las variables con las que se cuenta inicialmente.

Tabla 6.7: Descripción de las variables en la base de datos

Código/Alias	Descripción	Tipo	Medida
nro_tramite	Código del documento	numérico	nominal
nro_operacion	Código de la operación crediticia	numérico	nominal
ente	Código único del cliente	numérico	nominal
fecha_contabilizacion	Fecha de desembolso del crédito	fecha	escala
mes_contabilizacion	Mes de desembolso del crédito	numérico	escala
fecha_vencimiento	Fecha de vencimiento del crédito	fecha	escala
cod_forma_pago	Código de la forma de pago del crédito	texto	nominal
des_forma_pago	Forma de pago del crédito	texto	nominal
monto_aprobado	Monto del crédito	numérico	escala
cod_destino_final	Código del destino final de la inversión	texto	nominal
des_destino_final	Destino final de la inversión	texto	nominal
cod_oficina	Código de la oficina	numérico	nominal
des_oficina	Ofician donde se recepto la solicitud de crédito	texto	nominal
div_pendientes	número de dividendos pendientes de pago	numérico	escala
div_pagados	número de dividendos pagados	numérico	escala
atraso_maximo	el mayor número de días de vencimiento de una cuota que el cliente registra en la IMF	numérico	escala
atraso_prom	el número promedio de días de vencimiento de una cuota que el cliente registra en la IMF	numérico	escala
op_sector_bnf	Código del Sector al cual se destina el crédito	numérico	escala
des_op_sector_bnf	Sector al cual se destina el crédito	texto	nominal
cod_provincia_ubi_inv	Código de la provincia donde se encuentra la inversión	numérico	nominal
des_provincia_ubi_inv	Provincia de la ubicación de la inversión	texto	nominal
edad	Edad del cliente a la fecha corte	numérico	escala
total_ingresos	Total ingresos del cliente	numérico	escala
total_gastos	Total gastos del cliente	numérico	escala
total_activos	Total activos del cliente	numérico	escala
total_pasivos	Total pasivos del cliente	numérico	escala
plazo_meses	plazo en meses del créditp	numérico	escala
telefonos_registrados	número de teléfonos registrados del cliente	numérico	escala
cod_sexo	código del sexo del cliente	texto	nominal
des_sexo	sexo del cliente	texto	nominal
num_cargas	Indica el número de cargas del cliente	numérico	escala
num_cargas_estudiando	Indica el número de cargas del cliente que se encuentran estudiando	numérico	escala
cod_estado_civil	código del estado civil del cliente	texto	nominal
des_estado_civil	estado civil del cliente	texto	nominal
cod_nivel_instruccion	código del nivel de educación del cliente	texto	ordinal
des_nivel_instruccion	nivel de educación del cliente	texto	ordinal
cod_profesion	código de la profesión del cliente	texto	nominal
des_profesion	profesión del cliente	texto	nominal
cod_actividad	código de la actividad a la que se dedica el cliente	texto	nominal
des_actividad	actividad económica del cliente	texto	nominal
fecha_nacimiento	Indica la fecha de nacimiento del cliente	fecha	escala
acreedores_anteriores_36_meses	número de acreedores anteriores de 36 meses	numérico	nominal
acreedores_anteriores_consumo_36_meses	número de acreedores anteriores en consumo de 36 meses	numérico	nominal
acreedores_anteriores_comercial_36_meses	número de acreedores anteriores en comercial de 36 meses	numérico	nominal
acreedores_anteriores_microcrédito_36_meses	número de acreedores anteriores en microcrédito de 36 meses	numérico	nominal
acreedores_vigentes	número de acreedores vigentes	numérico	nominal
acreedores_vigentes_microcrédito	número de acreedores vigentes en microcrédito	numérico	nominal
acreedores_vigentes_comercial	número de acreedores vigentes en comercial	numérico	nominal
acreedores_vigentes_consumo	número de acreedores vigentes en consumo	numérico	nominal
saldo_total_deuda_actual_ifis	saldo total de deuda actual que el cliente tiene con otras instituciones financieras	numérico	escala
endeudamiento_promedio	endeudamiento promedio del cliente en el sistema financiero	numérico	escala
endeudamiento_promedio_microcrédito	endeudamiento promedio en microcrédito del cliente en el sistema financiero	numérico	escala
endeudamiento_promedio_consumo	endeudamiento promedio en consumo del cliente en el sistema financiero	numérico	escala
endeudamiento_promedio_comercial	endeudamiento promedio en comercial del cliente en el sistema financiero	numérico	escala
saldo_demanda_judicial	saldo en demanda judicial	numérico	escala
saldo_cartera_castigada	saldo en cartera castigada	numérico	escala
mayorplazovencidoactualcons	Indica el mayor plazo vencido en tipo consumo que el cliente registra en la Central de Riesgo	numérico	nominal
mayorplazovencidoactualcom	Indica el mayor plazo vencido en tipo comercial que el cliente registra en la Central de Riesgo	numérico	nominal
mayorplazovencidoactualmic	Indica el mayor plazo vencido en tipo microcrédito que el cliente registra en la Central de Riesgo	numérico	nominal
mayorplazovencidoactualtotal	Indica el mayor plazo vencido total que el cliente registra en la Central de Riesgo	numérico	nominal
mayorvalorvencidoactualcon	Indica la deuda de tipo consumo que el cliente registra en la Central de Riesgo	numérico	escala
mayorvalorvencidoactualcom	Indica la deuda de tipo comercial que el cliente registra en la Central de Riesgo	numérico	escala
mayorvalorvencidoactualmic	Indica la deuda de tipo microcrédito que el cliente registra en la Central de Riesgo	numérico	escala
mayorvalorvencidoactualtotal	Indica la deuda total que el cliente registra en la Central de Riesgo	numérico	escala
mayorplazovencido6mesescon	Indica el mayor plazo vencido en tipo consumo que el cliente registra en la Central de Riesgo hace 6 meses	numérico	nominal
mayorplazovencido6mesescom	Indica el mayor plazo vencido en tipo comercial que el cliente registra en la Central de Riesgo hace 6 meses	numérico	nominal
mayorplazovencido6mesesmic	Indica el mayor plazo vencido en tipo microcrédito que el cliente registra en la Central de Riesgo hace 6 meses	numérico	nominal
mayorplazovencido6mesestotal	Indica el mayor plazo vencido total que el cliente registra en la Central de Riesgo hace 6 meses	numérico	nominal
mayorvalorvencido6mesescon	Indica la deuda de tipo consumo que el cliente registra en la Central de Riesgo hace 6 meses	numérico	escala
mayorvalorvencido6mesescom	Indica la deuda de tipo comercial que el cliente registra en la Central de Riesgo hace 6 meses	numérico	escala
mayorvalorvencido6mesesmic	Indica la deuda de tipo microcrédito que el cliente registra en la Central de Riesgo hace 6 meses	numérico	escala
mayorvalorvencido6mesestotal	Indica la deuda total que el cliente registra en la Central de Riesgo hace 6 meses	numérico	escala
mayorplazovencidohistóricotota	Indica el mayor plazo vencido total que el cliente registra en la Central de Riesgo hace 36 meses	numérico	nominal
score_buro	Indica la puntuación que el cliente registra en la central de riesgos	numérico	escala
cuota_estimada_buro	Indica la cuota promedio que el cliente tiene en la central de riesgos	numérico	escala

Fuente: Registro Crediticio de la IMF. Elaborado por: Autor

Análisis descriptivo

El análisis descriptivo de una base de datos consiste en la aplicación de una serie de herramientas para obtener una información inicial de los mismos permitiendo cierta familiarización con las bases de datos. Entre los objetivos de este análisis se tiene:

- Investigar la calidad de los datos y descripción de los mismos.

- Obtener un conocimiento básico de los datos y de las relaciones entre las variables.

El análisis descriptivo requiere la identificación de los tipos de variables con las que se dispone para trabajar, como por ejemplo:

Clasificación por el nivel de medición

- **Variables Nominales:** son aquellas que indican diferencia en categoría, clase, calidad o tipo, por lo que poseen categorías representadas por nombres. Por ejemplo lugar de nacimiento, carrera académica, género, raza, etc
- **Variables Ordinales:** designan categorías, que pueden ser clasificadas desde la menor hasta la mayor; es decir, que existe un orden natural entre las categorías. Las variables ordinales comunes incluyen, entre otras clasificación de clase social (alta, media, baja), calidad de vivienda (estándar, insuficiente, en ruinas).
- **Variables de Intervalo (Escala):** estas variables identifican las diferencias en monto, cantidad, grado o distancia, y se les asignan puntuaciones numéricas más útiles.
- **Variables de Razón:** estas variables poseen las características de las variables de Intervalo y un punto cero verdadero, donde una puntuación cero significa ninguno: Peso, altura, edad, ingreso, distancia, duración de tiempo y promedio son variables de razón.

Clasificación por la periodicidad de medición:

- **Variables Longitudinales: se observan a lo largo del tiempo (series temporales)**
- **Variables Transversales o de Corte Transversal:** se observan en un instante de tiempo dado.
- **Variables de tipo Panel:** caso mixto, considera variables longitudinales y variables transversales.

Clasificación por la posibilidad de ser cuantificadas numéricamente

- **Variables Cualitativas o no métricas:** describen cualidades de un objeto.
- **variables Cuantitativas o métricas:** utilizan unidades de medida.

Para la determinación de la estructura y la calidad de la información se debe investigar la posible presencia de errores, puntos atípicos y datos perdidos. A continuación se presentarán, a manera de ejemplo, los resultados obtenidos para algunas de las variables, la totalidad del análisis univariante se presentará en el Anexo II.

Para el análisis descriptivo de los datos se utilizará el programa estadístico R 3.0.2, que ofrece el análisis de frecuencias, con sus respectivos gráficos, para las variables nominales y el análisis descriptivo para las variables cuantitativas.

En la tabla 6.8 muestra el análisis de frecuencias de la variable provincia de ubicación de la inversión, en el que se aprecia que las provincias de mayor concentración son Manabí, Riobamba y Loja, mientras que las provincias de Galápagos y Santa Elena tienen muy poca representatividad. No se encuentran valores perdidos.

Tabla 6.8: Análisis de frecuencias variable provincia de inversión.

Provincia	Frecuencia	Frecuencia_relativa	Porcentaje_Valido	Frecuencia_relativa_acumulada
AZUAY	593.00	0.04	0.04	0.04
BOLIVAR	639.00	0.04	0.04	0.07
CAÑAR	276.00	0.02	0.02	0.09
CARCHI	441.00	0.03	0.03	0.12
CHIMBORAZO	1910.00	0.11	0.11	0.23
COTOPAXI	1203.00	0.07	0.07	0.30
EL ORO	643.00	0.04	0.04	0.34
ESMERALDAS	585.00	0.03	0.03	0.37
GALAPAGOS	14.00	0.00	0.00	0.37
GUAYAS	886.00	0.05	0.05	0.43
IMBABURA	357.00	0.02	0.02	0.45
LOJA	1611.00	0.10	0.10	0.54
LOS RIOS	605.00	0.04	0.04	0.58
MANABI	2399.00	0.14	0.14	0.72
MORONA SANTIAGO	989.00	0.06	0.06	0.78
NAPO	483.00	0.03	0.03	0.81
ORELLANA	468.00	0.03	0.03	0.84
PASTAZA	278.00	0.02	0.02	0.85
PICHINCHA	485.00	0.03	0.03	0.88
SANTA ELENA	20.00	0.00	0.00	0.88
SANTO DOMINGO DE LOS TSACHILAS	189.00	0.01	0.01	0.89
SUCUMBIOS	570.00	0.03	0.03	0.93
TUNGURAHUA	511.00	0.03	0.03	0.96
ZAMORA CHINCHIPE	705.00	0.04	0.04	1.00
TOTAL	16860.00	1.00	1.00	

Fuente: Registro Crediticio de la IMF. Elaborado por: Autor

La variable estado civil presenta una concentración en las categorías estado civil casado y soltero, mientras que en las categorías viudo y separado se encuentran pocos casos en comparación al tamaño de la muestra, el detalle se puede apreciar en la tabla 6.9.

Para la variable nivel de instrucción, obtenida de la base de la información de clientes de la IMF, se encontró que un 58 % de la muestra de construcción tienen educación primaria y el 33 % tienen educación secundaria, mientras que en las categorías restantes, los casos se distribuyen con poca representatividad.

Para el análisis de las variables cuantitativas se encontraron los estadísticos descriptivos mínimo, máximo, media, cuartiles y desviación típica. La tabla 6.11 resume los

Tabla 6.9: Análisis de frecuencias variable estado civil.

	Frecuencia	Frecuencia_relativa	Porcentaje_Valido	Frecuencia_relativa_acumulada
CASADO	8266.00	0.49	0.49	0.49
DIVORCIADO	614.00	0.04	0.04	0.53
SOLTERO	7238.00	0.43	0.43	0.96
UNION LIBRE	305.00	0.02	0.02	0.97
VIUDO	437.00	0.03	0.03	1.00
TOTAL	16860.00	1.00	1.00	

Fuente: Registro Crediticio de la IMF. Elaborado por: Autor

Tabla 6.10: Análisis de frecuencias variable nivel de instrucción.

	Frecuencia	Frecuencia_relativa	Porcentaje_Valido	Porcentaje_acumulado
FORMACION INTERMEDIA (TECNICA)	8.00	0.00	0.00	0.00
POSTGRADO	3.00	0.00	0.00	0.00
PRIMARIA	9745.00	0.58	0.58	0.58
SECUNDARIA	5523.00	0.33	0.33	0.91
SIN ESTUDIOS	789.00	0.05	0.05	0.95
UNIVERSITARIA	792.00	0.05	0.05	1.00
Sum	16860.00	1.00	1.00	

Fuente: Registro Crediticio de la IMF. Elaborado por: Autor

estadísticos calculados en el programa R.

Tabla 6.11: Resumen del Análisis Univariante para las Variables Cuantitativas

VARIABLE	N	MIN	Q1	MEDIANA	PROMEDIO	Q3	MAX	DESVEST	R.I.
edad	16860.00	19.00	31.00	41.00	42.39	52.00	91.00	13.74	21.00
monto_aprobado	16860.00	512.80	4000.00	6400.00	7650.83	10000.00	20000.00	4442.88	6000.00
div_pendientes	16860.00	1.00	3.00	4.00	5.92	7.00	83.00	6.41	4.00
div_pagados	16860.00	0.00	1.00	2.00	2.61	3.00	25.00	3.51	2.00
dividendos	16860.00	1.00	4.00	6.00	8.54	10.00	84.00	9.27	6.00
plazo_meses	16860.00	2.00	36.00	48.00	49.45	60.00	288.00	19.31	24.00
total_ingresos	16860.00	100.00	1150.00	1780.00	4751.47	3481.50	433510.42	11216.59	2331.50
total_gastos	16860.00	8.00	420.48	715.00	2300.54	1440.00	342588.00	7445.36	1019.52
total_activos	16860.00	100.00	19903.20	34598.05	53763.29	62850.00	30189527.00	243371.81	42946.80
total_pasivos	16860.00	0.00	10.00	1648.89	5158.10	7000.00	200000.00	8314.96	6990.00
num_cargas	16860.00	0.00	0.00	2.00	1.93	3.00	16.00	1.74	3.00
num_cargas_estudiando	16860.00	0.00	0.00	1.00	1.29	2.00	12.00	1.40	2.00
saldo_total_deuda_actual_ifis	16860.00	0.00	0.00	0.00	2105.95	1833.41	161579.72	5344.92	1833.41
endeudamiento_promedio	16860.00	0.00	0.00	17.15	1393.56	1456.43	62499.30	2975.01	1456.43
endeudamiento_promedio_microcrédito	16860.00	0.00	0.00	0.00	840.04	693.68	24268.35	1969.82	693.68
endeudamiento_promedio_consumo	16860.00	0.00	0.00	0.00	458.16	0.00	33355.16	1690.31	0.00
endeudamiento_promedio_comercial	16860.00	0.00	0.00	0.00	50.69	0.00	60652.31	953.31	0.00
saldo_demanda_judicial	16860.00	0.00	0.00	0.00	15.78	0.00	21521.38	341.26	0.00
saldo_cartera_castigada	16860.00	0.00	0.00	0.00	6.19	0.00	7796.81	153.58	0.00
mayorvalorvencidoactualcon	16860.00	0.00	0.00	0.00	9.46	0.00	13768.10	200.79	0.00
mayorvalorvencidoactualcom	16860.00	0.00	0.00	0.00	1.68	0.00	17001.67	140.49	0.00
mayorvalorvencidoactualmic	16860.00	0.00	0.00	0.00	13.67	0.00	5293.88	136.45	0.00
mayorvalorvencidoactualtotal	16860.00	0.00	0.00	0.00	23.71	0.00	17001.67	277.18	0.00
mayorvalorvencido6mesescon	16860.00	0.00	0.00	0.00	17.48	0.00	17677.43	282.36	0.00
mayorvalorvencido6mesescom	16860.00	0.00	0.00	0.00	1.93	0.00	17001.67	144.08	0.00
mayorvalorvencido6mesesmic	16860.00	0.00	0.00	0.00	21.87	0.00	5569.70	186.58	0.00
mayorvalorvencido6mesestotal	16860.00	0.00	0.00	0.00	39.63	0.00	17677.43	363.94	0.00
score_buro	16860.00	45.00	519.00	764.00	642.43	813.00	970.00	246.15	294.00

Fuente: Registro Crediticio de la IMF. Elaborado por: Autor

6.2.5 Selección de las variables explicativas del modelo

Compete ahora seleccionar las variables que explican a la variable dependiente, es decir aquellas que definen a un cliente como bueno o malo. Es entonces parte crucial en la modelización, dado que las variables escogidas deben ser suficientes para explicar a la variable dependiente, pero no deben ser demasiadas tal que compliquen el modelo, es decir se aplica el principio de parsimonia.

Análisis de Correlaciones

Luego del estudio descriptivo de las variables, se debe realizar un análisis de las variables en conjunto para conocer cada uno de sus atributos, el objeto es identificar el grado con el que pueden contribuir en el modelo para discriminar entre buenos y malos. Además en esta etapa se reducirá el número de categorías por variable, esto último se explica por dos razones, cada una asociada al tipo de variable (categórica o continua):

- En el caso de las variables categóricas, contar con demasiados atributos puede agotar la muestra para cada respuesta y quitarle robustez al análisis; además pueden existir atributos que no contienen información necesaria para suponer que su razón buenos/malos se reproduce en el total poblacional. Los atributos deben ser representativos en la variable, según se indicó anteriormente se asumirá como no significativa a una representatividad del 5 %.

- Dado que el objetivo del credit scoring es predecir el riesgo en lugar de explicarlo, para las variables continuas es preferible contar con un sistema en el cual el riesgo no sea monótono en estas variables, siempre que esto permita una mejor predicción del mismo. Puesto que no siempre es posible decir que al crecer una variable continua el riesgo es mayor o menor, muchas veces el riesgo se relaciona a agrupaciones de la variable, por ejemplo, en la variable edad, se puede caracterizar a los clientes menores a 25 años como malos, a los clientes entre 25 y 55 años como buenos, y a los mayores a 65 nuevamente como malos; en tal caso no existe una relación creciente o decreciente definida. Las relaciones que permiten predecir el riesgo, en base a variables continuas, son en su mayoría como la explicada en el ejemplo anterior. Por ello es más eficiente identificar las posibles agrupaciones de la variable y la asignación bueno/malo esperada, a incluir directamente la variable continua.

Las variables explicativas a ser incluidas en el modelo discriminante deben estar altamente correlacionadas con la variable dependiente del modelo (Indicador B/M); por tal razón es necesario determinar si la variable indicador buenos/Malos depende o no de

las variables explicativas, para tal fin se utilizarán árboles de decisión, conocidos también como algoritmos de partición recursiva.

Un árbol de decisión no utiliza un modelo estadístico formal y es más bien un algoritmo para clasificar utilizando particiones binarias sucesivas de los valores de una variable cada vez, esta partición recursiva se traduce en una organización jerárquica del espacio de representación que puede modelarse mediante una estructura de tipo árbol. Cada nodo interior contiene una pregunta sobre un atributo concreto (con un hijo por cada posible respuesta) y cada nodo hoja se refiere a una decisión (clasificación). Las decisiones a tomar en la construcción de los árboles de decisión, para este caso en específico son:

- La selección de las variables y de sus puntos de corte para hacer las divisiones, regla de partición.
- La asignación de los nodos terminales a categoría Bueno o Malo, decisión de asignación. Con esto se determina cual es el signo esperado de la variable.

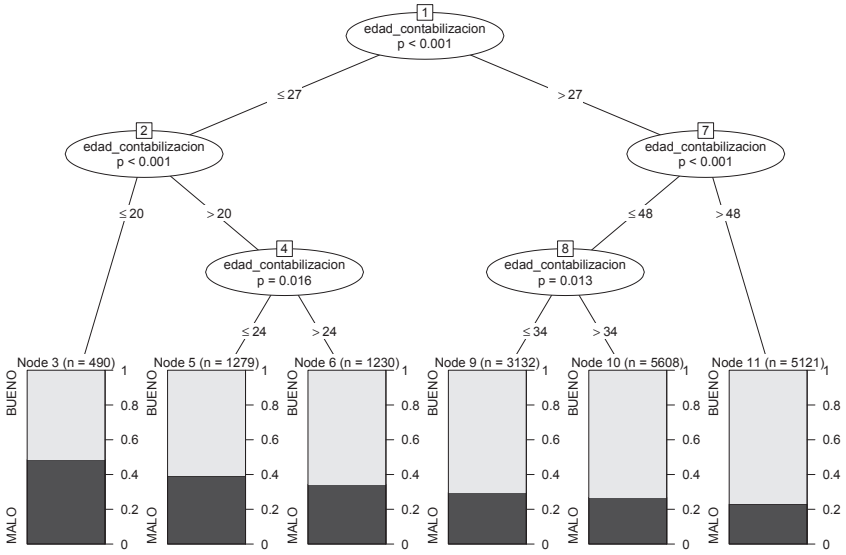
La decisión de asignación se toma en función de la proporcionalidad de las categorías Bueno-Malo en el nodo inicial, normalmente se asigna el nodo como bueno, si la proporcionalidad se mantiene o se concentra en la categoría de Buenos; por ejemplo si la proporcionalidad inicial es 70/30, un nodo será considerado como bueno si existe una concentración en la categoría de Buenos mayor a la original.

En la actualidad existen varios modelos y programas para la construcción de árboles de decisión, para este estudio se utiliza árboles de inferencia condicional (ctree) disponible en el software **R** en la librería **party**. En el gráfico 6.5 se presentan los resultados obtenidos para la variable Edad.

En el gráfico 6.5, se aprecia que el valor p es 0, menor que 5 %, por lo tanto se rechaza la hipótesis de independencia a un nivel del 95 % de confianza. Se observa que el número de clientes buenos va incrementándose a medida que se incrementa la edad en los grupos obtenidos; es decir, por ejemplo, los clientes mayores de 48 son mejores clientes que aquellos con edades entre 34 y 48 años, la relación encontrada indica que la variable edad es candidata para ser ingresada al modelo como variable continua, aunque también se podría utilizar los rangos de edades encontrados para cruzar con otras variables y encontrar nichos de clientes bueno o malos.

La partición incluso ha permitido discretizar la variable continua edad en 4 categorías, de tal manera que esté correlacionada con la variable Indicador B/M. Los grupos obtenidos son los descritos en la tabla 6.12.

Figura 6.5: Árbol de decisión: Variables edad a la fecha de contabilización



Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla 6.12: Discretización Variable Edad

	BUENO	MALO	BUENO (%)	MALO(%)
≤ 27	1838.00	1161.00	0.61	0.39
27-34	2202.00	930.00	0.70	0.30
34-48	4104.00	1504.00	0.73	0.27
≥ 48	3945.00	1176.00	0.77	0.23

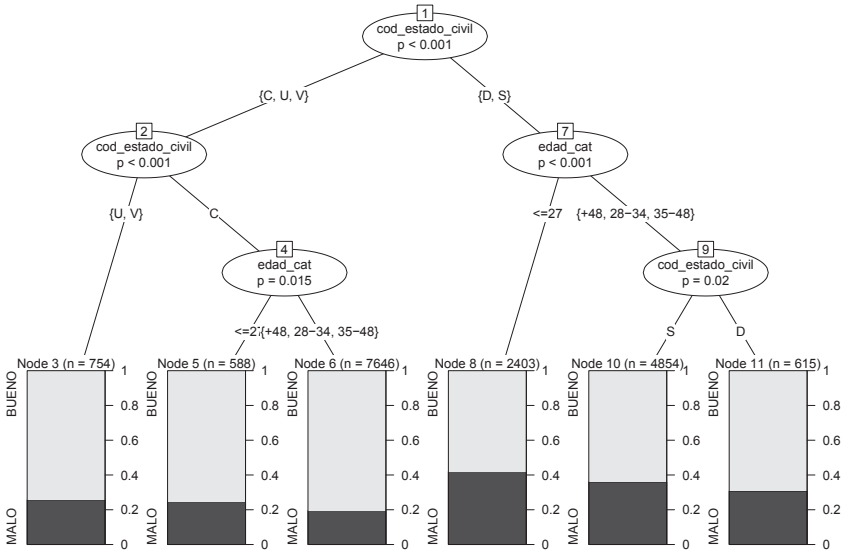
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

La edad debe ser parte de las variables explicativas del modelo, debido a que el Indicador B/M es dependiente de la misma.

La influencia de una de las variables sobre la probabilidad de ocurrencia de un suceso o evento se modifica en función del valor de otra de las variables y es necesario incluir en el modelo una tercera que sea el producto de las anteriores. Estos son los conocidos como términos de interacción o cruces de variables que pueden incluir 2 o más variables. Los cruces de variables se introducen cuando existen razones para suponer que la influencia de una de las variables varía en función del valor que asume otra de las variables incluidas en el modelo; o sea, si la influencia de X_1 varía en función del valor que toma X_2, incluimos en el modelo un término que represente la interacción de X_1 y X_2.

Para el análisis se deben cruzar las variables, de tal manera que se identifiquen características de discriminación conjuntas, el cruce de variables permite no caer en la paradoja de Simpson [1]. Los árboles de decisión permiten conseguir los cruces de variables que estén más correlacionados con la variable indicador B/M.

Figura 6.6: Árbol de Decisión Variable Cruzada Estado Civil, Edad



Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

El gráfico 6.6 muestra el árbol de decisión para el cruce de las variables Estado Civil y Edad, el primer nivel de partición se lo hace con la variable Estado civil; mientras que en el segundo nivel de partición se cruzan cada uno de los rangos de la variable Edad. Un cruce dependiente y significativo debe ser incluido en el modelo. Las variables resultantes del análisis de árboles (B) se presentan en la tabla 6.13:

Seleccionadas y construidas las variables se procedió a correr los modelos expuestos en este estudio en búsqueda del modelo que logre un mejor poder de discriminación. Los modelos se construyeron empleando el software estadístico **R** usando la librería `library(MASS)` con la función `lda` para el análisis discriminante lineal y la librería `e1071` con la función `svm` para las máquinas de soporte vectorial; y el software `zimpl` para la corrida de los modelos de programación lineal. las variables que mejor discriminaron se muestran a continuación junto con las matrices de confusión o clasificación para cada uno de los modelos construidos.

[1] Se denomina paradoja de Simpson al cambio en el sentido de una asociación entre dos variables (numéricas o cualitativas) cuando se controla el efecto de una tercera variable.

Tabla 6.13: Variables Construidas

Alias	Variables que intervienen	Descripción	Signo Esperado
V_1	Provincia ubicación de inversión	Cañar, Chimborazo, El Oro, Imbabura, Loja, Pichincha, Tungurahua	Buenos
V_2	Edad	Variable continua	Buenos
V_3	Estado civil Nivel instrucción	Soltero Analfabeto	Malos
V_4	Estado civil Edad	Casado mayor de 27 años	Buenos
V_5	Nivel Educación Edad	Analfabeto, Primaria o Secundaria, Menor a 27 años	Malos
V_6	Nivel Educación Edad	Analfabeto, Primaria o Secundaria, Mayor a 28 años	Buenos
V_7	Género - Edad	Femenino, 27 años o menos	Malos
V_8	Género - Edad	Masculino, 27 años o menos	Malos
V_9	mayorplazovencidohistoricototal	cero plazos vencidos	Buenos
V_10	mayorplazovencido6mesestotal	cero plazos vencidos	Buenos
V_11	acreedores_anteriores_36_meses	2 acreedores o menos	Buenos
V_12	mayorplazovencidohistoricototal - score_buro	cero plazos vencidos y un score de 694 o más	Buenos
V_13	acreedores_anteriores_36_meses-score_buro	1 acreedor o menos y un score de 692 o más	Buenos
V_14	mayorplazovencidohistoricototal - endeudamiento_promedio	cero plazos vencidos y un endeudamiento promedio menor o igual a 830,84	Buenos

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

6.3 Aplicación de los modelos

Para comparar los modelos estadísticos y los métodos no paramétricos, se considera el conjunto de datos o muestra de entrenamiento (training) de 16860 solicitantes de crédito. Este conjunto de datos se divide en dos grupos de solicitantes como se ha expuesto en las secciones anteriores: los solicitantes que fueron aceptados, y han registrado un buen comportamiento de crédito (Grupo 1 “buenos préstamos”); los solicitantes que fueron aceptados, pero han incumplido sus obligaciones crediticias (Grupo 2 “malos préstamos”). Además, los dos grupos se caracterizan por variables cualitativas y cuantitativas (forma de pago (x_1), monto solicitado (x_2), destino de la inversión (x_3), zonal (x_4), número de dividendos (x_5), acreedores anteriores de 36 meses (x_6), mayor plazo vencido total 6 meses (x_7), mayor plazo vencido total 36 meses (x_8), edad a la fecha de contabilización (x_9), provincia de ubicación de la inversión (x_{10}), score del buró de crédito (x_{11}), número de cargas familiares (x_{12}), número de cargas familiares estudiando (x_{13}), estado civil (x_{14}), nivel de instrucción (x_{15}), tipo de vivienda (x_{16}), meses de residencia (x_{17}), actividad económica (x_{18})).

6.3.1 Resultados del planteamiento estadístico

Resultados del procedimiento de la función discriminante (LDF)

El objetivo de este método es seleccionar, entre el conjunto de variables, las variables más significativas para determinar la función discriminante lineal de Fisher. La tabla 6.14 muestra las variables más significativas dadas por el procedimiento paso a paso utilizando el test Lambda de Wilks.

Esta tabla muestra que las variables más importantes son score del buró de crédito, acreedores anteriores de 36 meses, mayor plazo vencido total 36 meses, mayor plazo

Tabla 6.14: Procedimiento paso a paso

Paso	Introducir	Estadístico	df_1	df_2	df_3	Estadístico	df_1	df_2	Significancia
1	score_buro	0.543	1	1	16858	14150.841	1	16858	0
2	acreedores_anteriores_36_meses	0.506	2	1	16858	8224.858	2	16857	0
3	mpvht	0.499	3	1	16858	5637.744	3	16856	0
4	mayorplazovencido6mesestotal	0.490	4	1	16858	4374.788	4	16855	0
5	cod_estado_civil	0.487	5	1	16858	3541.572	5	16854	0
6	act_cat	0.485	6	1	16858	2973.303	6	16853	0
7	dividendos	0.484	7	1	16858	2560.503	7	16852	0
8	cod_tipo_vivienda	0.484	8	1	16858	2244.568	8	16851	0
9	cod_provincia_ubi_inv	0.483	9	1	16858	1998.321	9	16850	0
10	cod_destino_cat	0.483	10	1	16858	1800.779	10	16849	0
11	cod_nivel_instruccion	0.483	11	1	16858	1638.608	11	16848	0
12	cod_forma_pago	0.482	12	1	16858	1503.241	12	16847	0
13	edad_contabilizacion	0.482	13	1	16858	1388.614	13	16846	0

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

vencido total 6 meses, estado civil, actividad económica categorizada, dividendos, tipo de vivienda, provincia de ubicación de la inversión, destino final de la inversión, nivel de instrucción, forma de pago, edad a la fecha de otorgación del crédito. Ahora, se debe determinar los coeficientes de estas variables y el término constante para la función discriminante lineal de Fisher. La tabla 6.15 a continuación proporciona los coeficientes de la función lineal dicriminante:

Tabla 6.15: Coeficientes de la función discriminante lineal

Variables	LD1
cod_forma_pago	-0.012129534
cod_destino_cat	-0.040603846
dividendos	-0.005813963
acreedores_anteriores_36_meses	-0.280963626
mayorplazovencido6mesestotal	-0.219606946
mpvht	0.170840658
edad_contabilizacion	-0.002342465
cod_provincia_ubi_inv	0.007122041
score_buro	-0.005341731
cod_estado_civil	0.076595358
cod_nivel_instruccion	-0.066034638
cod_tipo_vivienda	-0.058865276
act_cat	-0.089233033

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Por lo tanto, la función discriminante lineal está dada por la siguiente ecuación

$$Z = -0{,}012x_1 - 0{,}040x_2 - 0{,}005x_3 - 0{,}280x_4 - 0{,}219x_5 + 0{,}170x_6 - 0{,}002x_7 + 0{,}007x_8 \quad (6.2)$$

$$-0{,}005x_9 + 0{,}076x_{10} - 0{,}066x_{11} - 0{,}058x_{12} - 0{,}089x_{13}$$

En consecuencia utilizando la ecuación (6.2) se puede clasificar futuras observaciones en el grupo 1 (bueno) o en el grupo 2 (malo), también se puede determinar la proporción de aciertos de la clasificación. Los resultados de clasificación se presentan en la tabla 6.16:

Tabla 6.16: Resultado de Clasificación de la función lineal discriminante

		Clases Asignadas		
	GRUPO	BUENO	MALO	TOTAL
ORIGINAL	BUENO	11479.00	602.00	12081.00
	MALO	1288.00	3491.00	4779.00
PROPORCIÓN	BUENO	0.95	0.05	1.00
	MALO	0.27	0.73	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla 6.17: Resultado del Back testing de la función lineal discriminante

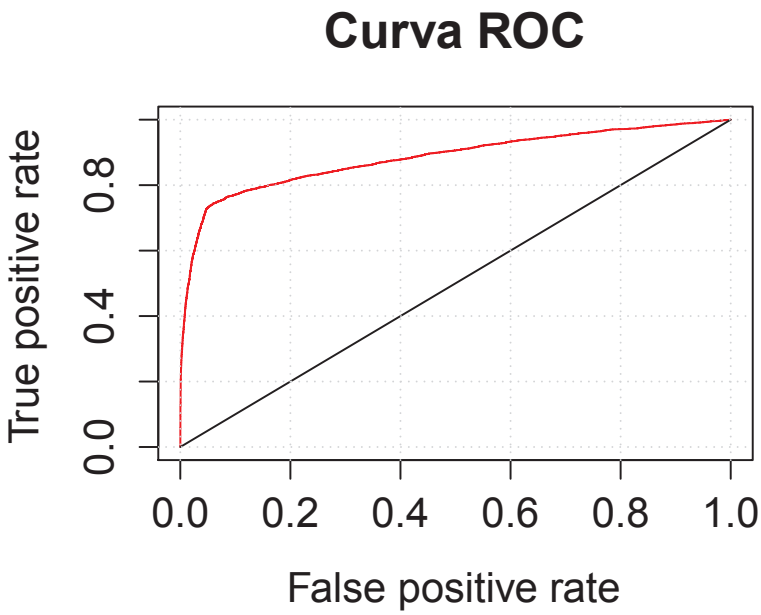
		Clases Asignadas		
	GRUPO	BUENO	MALO	TOTAL
ORIGINAL	BUENO	2901.00	153.00	3054.00
	MALO	324.00	837.00	1161.00
PROPORCIÓN	BUENO	0.949	0.050	1.00
	MALO	0.279	0.720	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

De acuerdo con la tabla 6.16, podemos notar que sólo hay 1890 observaciones que son mal clasificados. Por lo tanto, la proporción de aciertos dada por el procedimiento LDF en la muestra original (training) es del 88,79 %, mientras que el error de clasificación es del 11,21 %.

Este valor indica que la ecuación (6.2) da una buena discriminación entre los dos grupos demandantes y se pueden utilizar para clasificar las nuevas observaciones.
El área bajo la curva es 0.8837245. La figura 6.7 representa la curva ROC del análisis discriminante. La curva no está cercana a la diagonal principal, por lo que se puede afirmar que el método que se está empleando tiene una calidad de predicción buena.

Figura 6.7: Curva ROC: Función Discriminante Lineal



Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

6.3.2 Resultados del modelo de Máquinas de Vectores de Soporte

El resultado del modelo SVM implementado en R proporciona los siguientes resultados:

Tabla 6.18: Pesos de las variables con el modelo SVM

cod_forma_pago	ω_1	18.58261
cod_destino_cat	ω_2	112.60166
dividendos	ω_3	204.50406
acreedores_anteriores_36_meses	ω_4	1582.62908
mayorplazovencido6mesestotal	ω_5	827.80883
mpvht	ω_6	-1025.15709
edad_contabilizacion	ω_7	108.77614
cod_provincia_ubi_inv	ω_8	-66.70024
score_buro	ω_9	5652.6518
cod_estado_civil	ω_{10}	-253.58522
cod_nivel_instruccion	ω_{11}	141.74398
cod_tipo_vivienda	ω_{12}	134.67654
act_cat	ω_{13}	243.18115
término constante	b	16.62713

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Por lo tanto, la regla de clasificación es:

$$d = \sum_{j=1}^{13} \omega_j x_{ij} + b$$

$$\begin{cases} \text{si } d > 0 & x_i \in G_1 \\ \text{caso contrario} & x_i \in G_2 \end{cases}$$

Tabla 6.19: Resultado de Clasificación de la Máquina de Vectores de Soporte

		Clases Asignadas		
	GRUPO	BUENO	MALO	TOTAL
ORIGINAL	BUENO	11467.00	614.00	12081.00
	MALO	1274.00	3505.00	4779.00
PROPORCIÓN	BUENO	0.9491	0.0508	1.00
	MALO	0.2665	0.7334	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

De la tabla 6.19 (6.20) se puede deducir que el porcentaje de clientes bien clasificados es del 88.80 %, es importante considerar que 73,33 % de los clientes que son pronosticados malos en realidad son malos; esto no debe preocupar pues no se va a construir un scoring restrictivo, lo que se va a construir son perfiles de clientes, y probablemente la gran mayoría de estos clientes van a estar dentro de perfiles donde se les va a pedir más garantías y se dará mayor importancia a la información de la solicitud, previa verificación de la veracidad de la misma.

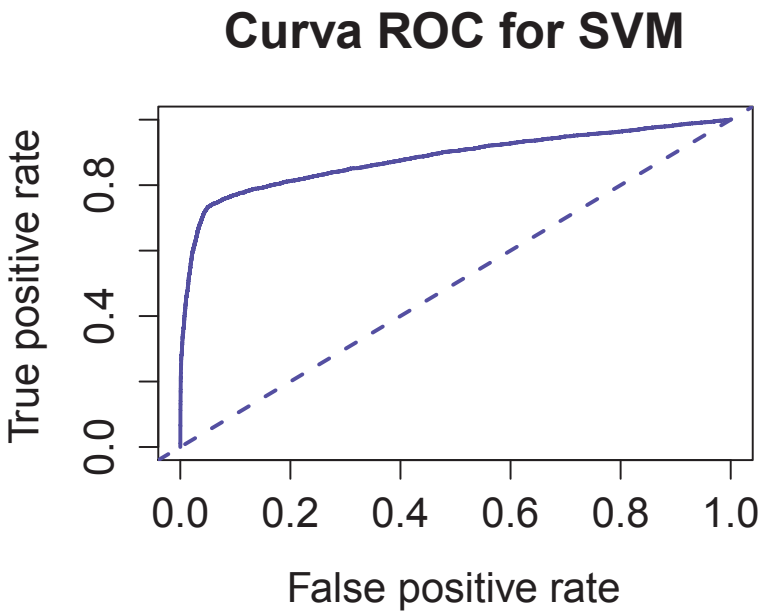
Tabla 6.20: Resultado del Back testing de la Máquina de Vectores de Soporte

		Clases Asignadas		
	GRUPO	BUENO	MALO	TOTAL
ORIGINAL	BUENO	2907.00	147.00	3054.00
	MALO	320.00	841.00	1161.00
PROPORCIÓN	BUENO	0.9519	0.0481	1.00
	MALO	0.2756	0.7224	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

En consecuencia la clasificación por máquinas de vectores de soporte es buena lo cual se rectifica con el coeficiente de gini (área bajo la curva 6.8) de 0.88. La figura 6.8 representa la curva ROC de la aplicación de máquinas de vectores de soporte. La curva no está cercana a la diagonal principal, por lo que se puede afirmar que el método que se está empleando tiene una calidad de predicción buena.

Figura 6.8: Curva ROC: Máquina de Vectores de Soporte



Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

6.4 Resultados de los modelos basados en programación lineal

En esta sección se describen los resultados de los modelos implementados en base a la teoría de la programación lineal, pero antes de esto es importante mencionar que cada uno de los algoritmos de programación lineal, tanto para los ejemplos de las secciones anteriores como para la aplicación real, fueron realizados en un ordenador (TOSHIBA) de las siguientes características: Sistema operativo Windows 8.1 de 64 bits, procesador Intel(R) Core (TM) i7-4700MQ CPU @ 2.40GHz, 8.00 GB de memoria RAM.

Los diferentes algoritmos fueron implementados inicialmente en el solver GAMS 24.4.1 por la facilidad de su lenguaje para la programación, pero su principal inconveniente es que al tratarse de un solver comercial presenta limitaciones para el número de restricciones y de variables, además de limitar la utilización de datos nulos, lo cual representó un obstáculo para los algoritmos expuestos en este proyecto pues para trabajar en la corrección del error en la clasificación se utiliza matrices dispersas (sparse matrix). Por otra parte también se implementó el algoritmo en el software R, a través de la librería `lpSolveAPI` pero a pesar de ser un software libre se presenta la dificultad de procesar las matrices dispersas que se involucran en el planteamiento de los algoritmos.

Finalmente los algoritmos fueron implementados en el solver `ZIMPL 1.2` el mismo que permitió solventar las dificultades iniciales, y los resultados se exponen a continuación:

6.4.1 Resultados del modelo MSD

La resolución del modelo MSD en ZIMPL con el punto de corte $b = -3{,}29104789$ proporciona los siguientes pesos:

Tabla 6.21: Pesos de las variables con el modelo MSD

variable	pesos	
cod_forma_pago	ω_1	-0.00272
monto_aprobado	ω_2	0.0000021
cod_destino_cat	ω_3	-0.0264678
cod_zonal	ω_4	0.0000608
dividendos	ω_5	-0.0046945
acreedores_anteriores_36_meses	ω_6	-0.3097683
mayorplazovencido6mesestotal	ω_7	-0.1588596
mpvht	ω_8	0.1068601
edad_contabilizacion	ω_9	-0.0013537
cod_provincia_ubi_inv	ω_{10}	0.0035573
score_buro	ω_{11}	-0.0046247
num_cargas	ω_{12}	-0.0038693
num_cargas_estudiando	ω_{13}	0.0002456
cod_estado_civil	ω_{14}	0.0489617
cod_nivel_instruccion	ω_{15}	-0.0557967
cod_tipo_vivienda	ω_{16}	-0.0380747
meses_residencia	ω_{17}	-0.000007
act_cat	ω_{18}	-0.048702

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Por lo tanto, la regla de clasificación es:

$$y_i = 0 \qquad \text{si el cliente } i \text{ es clasificado correctamente}$$

$$y_i = b \pm 1 - \sum_{j=1}^{18} \omega_j x_{ij} \qquad \text{si el cliente } i \text{ es clasificado erróneamente}$$

La tabla 6.22 a continuación muestra el resultado de clasificación del modelo MSD:

Tabla 6.22: Resultado de Clasificación del modelo MSD

	Clases Asignadas			
	GRUPO	BUENO	MALO	TOTAL
ORIGINAL	BUENO	11466.00	615.00	12081.00
	MALO	1281.00	3498.00	4779.00
PROPORCIÓN	BUENO	0.9491	0.0519	1.00
	MALO	0.2680	0.7320	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla 6.23: Resultado del Back testing del Modelo MSD

	Clases Asignadas			
	GRUPO	BUENO	MALO	TOTAL
ORIGINAL	BUENO	2907.00	147.00	3054.00
	MALO	320.00	841.00	1161.00
PROPORCIÓN	BUENO	0.9519	0.0481	1.00
	MALO	0.2748	0.7252	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

De acuerdo a la tabla 6.22, se puede notar que la proporción de aciertos del modelo de MSD es igual a 88.75 %. Por lo tanto, el porcentaje de error de clasificación es igual a 11.25 %. De la tabla 6.23 se deduce que el modelo no es subestimado ni sobrestimado en consecuencia el modelo MSD es un procedimiento de discriminación muy eficaz para resolver los problemas de clasificación.

6.4.2 Resultado del modelo MMD

Los factores de ponderación y el valor de corte determinado por el modelo MMD son iguales a:
y b = 2.577638751. Por lo tanto, la regla de clasificación para el modelo MMD es:

$$\text{Si } \sum_{j=1}^{18} \omega_j x_{ij} > 2{,}577638751 \text{ el cliente pertenece al grupo } G_2\text{(malo)}$$

$$\text{Si } \sum_{j=1}^{18} \omega_j x_{ij} < 2{,}577638751 \text{ el cliente pertenece al grupo } G_1\text{(bueno)}$$

De acuerdo con las reglas citadas, de la tabla 6.25 se puede deducir que el 36.31 % de los clientes son mal clasificados De hecho, sus puntuaciones de clasificación superan el

Tabla 6.24: Pesos de las variables con el modelo MMD

variable	pesos	
cod_forma_pago	ω_1	0.0145947
monto_aprobado	ω_2	-0.0000089
cod_destino_cat	ω_3	0.0667513
cod_zonal	ω_4	-0.0002435
dividendos	ω_5	0.0049109
acreedores_anteriores_36_meses	ω_6	0.0153041
mayorplazovencido6mesestotal	ω_7	0.0878768
mpvht	ω_8	0.0252406
edad_contabilizacion	ω_9	0.0055034
cod_provincia_ubi_inv	ω_{10}	0.0043363
score_buro	ω_{11}	-0.0000424
num_cargas	ω_{12}	0.0118599
num_cargas_estudiando	ω_{13}	0.0545948
cod_estado_civil	ω_{14}	0.145397
cod_nivel_instruccion	ω_{15}	0.2241634
cod_tipo_vivienda	ω_{16}	0.2037987
meses_residencia	ω_{17}	-0.0003098
act_cat	ω_{18}	0.1362728

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla 6.25: Resultado de Clasificación del modelo MMD

	Clases Asignadas			
	GRUPO	BUENO	MALO	TOTAL
ORIGINAL	BUENO	8970.00	3111.00	12081.00
	MALO	3011.00	1768.00	4779.00
PROPORCIÓN	BUENO	0.7425	0.2575	1.00
	MALO	0.63	0.37	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

valor de corte (b = 2.5776). Por otra parte se tiene que solo el 37 % de los clientes clasificados inicialmente como malos son pronosticados como malos, esto no debe preocupar pues no se construye un scoring restrictivo, lo que se construye son perfiles de clientes, y probablemente la gran mayoría de estos clientes van a estar dentro de perfiles donde se les va a pedir más garantías y se dará mayor importancia a la información de la solicitud.

Tabla 6.26: Resultado del Back testing del Modelo MMD

		Clases Asignadas		
	GRUPO	BUENO	MALO	TOTAL
ORIGINAL	BUENO	2279.00	775.00	3054.00
	MALO	747.00	414.00	1161.00
PROPORCIÓN	BUENO	0.7462	0.2538	1.00
	MALO	0.6434	0.3566	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

6.4.3 Resultado del modelo MCLP

Los factores de ponderación y el valor de corte determinado por el modelo MCLP son iguales a

Tabla 6.27: Pesos de las variables con el modelo MCLP

variable	pesos	
cod_forma_pago	ω_1	31.72812
monto_aprobado	ω_2	-0.02978
cod_destino_cat	ω_3	-53.34128
cod_zonal	ω_4	3.90159
dividendos	ω_5	-8.18659
acreedores_anteriores_36_meses	ω_6	-100
mayorplazovencido6mesestotal	ω_7	14.36259
mpvht	ω_8	200
edad_contabilizacion	ω_9	-2.92537
cod_provincia_ubi_inv	ω_{10}	-6.46654
score_buro	ω_{11}	-6.0352
num_cargas	ω_{12}	-33.89111
num_cargas_estudiando	ω_{13}	-9.39655
cod_estado_civil	ω_{14}	-13.36624
cod_nivel_instruccion	ω_{15}	-52.98987
cod_tipo_vivienda	ω_{16}	66.05319
meses_residencia	ω_{17}	-0.23744
act_cat	ω_{18}	-28.17952

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

y b = -3168.86316. Por lo tanto, la regla de clasificación de modelo MCLP es:

Si $\sum_{j=1}^{18} \omega_j x_{ij} > -3168{,}86316$ el cliente pertenece al grupo G_2(malo)

Si $\sum_{j=1}^{18} \omega_j x_{ij} < -3168{,}86316$ el cliente pertenece al grupo G_1(bueno)

De acuerdo con las reglas citadas, la proporción de aciertos es igual a 87.88 % y el resultado de la clasificación del modelo MCLP está dada por la siguiente tabla:

Tabla 6.28: Resultado de Clasificación del modelo MCLP

	Clases Asignadas			
	GRUPO	BUENO	MALO	TOTAL
ORIGINAL	BUENO	11498.00	583.00	12081.00
	MALO	1460.00	3319.00	4779.00
PROPORCIÓN	BUENO	0.9517	0.0483	1.00
	MALO	0.3055	0.6945	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla 6.29: Resultado del Back testing del Modelo MCLP

	Clases Asignadas			
	GRUPO	BUENO	MALO	TOTAL
ORIGINAL	BUENO	2908.00	146.00	3054.00
	MALO	362.00	799.00	1161.00
PROPORCIÓN	BUENO	0.9522	0.0478	1.00
	MALO	0.3118	0.6882	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

En la tabla 6.28 se puede apreciar que el porcentaje de clientes bien clasificados es del 87.88 %, además el 69,45 % de los clientes que son pronosticados malos en realidad son malos; esto no debe preocupar pues no se va a construir un scoring restrictivo, lo que se va a construir son perfiles de clientes, y probablemente la gran mayoría de estos clientes van a estar dentro de perfiles donde se les va a pedir más garantías y se dará mayor importancia a la información de la solicitud. El error de clasificación del modelo es 12.12 %. El modelo es bueno puesto que de los resultados del back testing, tabla 6.29, se deduce que el modelo no es ni subestimado ni sobrestimado. (otros resultados se pueden observar en el apéndice C.1).

6.4.4 Resultado del modelo MC2LP

Los factores de ponderación y el valor de corte determinado por el modelo MC2LP están dados en la tabla 6.30, y b = -2556.37791. Por lo tanto, la regla de clasificación de modelo MC2LP es:

Si $\sum_{j=1}^{18} \omega_j x_{ij} > -2556{,}37791$ el cliente pertenece al grupo G_2(malo)

Si $\sum_{j=1}^{18} \omega_j x_{ij} < -2556{,}37791$ el cliente pertenece al grupo G_1(bueno)

Tabla 6.30: Pesos de las variables con el modelo MC2LP

variable	pesos	
cod_forma_pago	ω_1	-9.45269
monto_aprobado	ω_2	-0.03786
cod_destino_cat	ω_3	-19.371
cod_zonal	ω_4	3.00802
dividendos	ω_5	-4.77158
acreedores_anteriores_36_meses	ω_6	-135.44622
mayorplazovencido6mesestotal	ω_7	-129.56497
mpvht	ω_8	252.24513
edad_contabilizacion	ω_9	-2.9418
cod_provincia_ubi_inv	ω_{10}	0.63589
score_buro	ω_{11}	-5.4423
num_cargas	ω_{12}	-19.55888
num_cargas_estudiando	ω_{13}	-38.4463
cod_estado_civil	ω_{14}	12.77529
cod_nivel_instruccion	ω_{15}	26.23756
cod_tipo_vivienda	ω_{16}	126.08741
meses_residencia	ω_{17}	-0.34889
act_cat	ω_{18}	-121.90193

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

De acuerdo con las reglas citadas el resultado de clasificación del modelo MC2LP está dado por la siguiente tabla:

Tabla 6.31: Resultado de Clasificación del modelo MC2LP

		Clases Asignadas		
	GRUPO	BUENO	MALO	TOTAL
ORIGINAL	BUENO	11469.00	612.00	12081.00
	MALO	1418.00	3361.00	4779.00
PROPORCIÓN	BUENO	0.9493	0.0517	1.00
	MALO	0.2967	0.7033	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla 6.32: Resultado del Back testing del Modelo MC2LP

		Clases Asignadas		
	GRUPO	BUENO	MALO	TOTAL
ORIGINAL	BUENO	2899.00	155.00	3054.00
	MALO	355.00	806.00	1161.00
PROPORCIÓN	BUENO	0.9492	0.0508	1.00
	MALO	0.3058	0.6492	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

En la tabla 6.31 se puede apreciar que el porcentaje de clientes bien clasificados es del 87.90 %, además el 70,33 % de los clientes que son pronosticados malos en realidad son malos; esto no debe preocupar pues no se va a construir un scoring restrictivo, lo que se va a construir son perfiles de clientes, y probablemente la gran mayoría de estos clientes van a estar dentro de perfiles donde se les va a pedir más garantías y se dará mayor importancia a la información de la solicitud. El error de clasificación del modelo es 12.10 %. El modelo es bueno puesto que de los resultados del back testing, tabla 6.32, se deduce que el modelo no es ni subestimado ni sobrestimado. (otros resultados se pueden observar en el apéndice C.2).

6.4.5 Análisis de los resultados

El proceso de clasificación general que se menciona a continuación resume los principales pasos en nuestro experimento de clasificación. Aunque este proceso sólo se refiere a un problema de dos clases, se puede extender a problemas multiclase cambiando la entrada y la función de decisión (Paso 2). Todas las aplicaciones analizadas en este capítulo siguen este proceso general.

Proceso General de Clasificación

Entrada: El conjunto de datos $\mathbb{A} = \{A_1, A_2, A_3 \cdots, A_n\}$, una matriz diagonal $\mathbb{Y}_{n\times n}$ donde

$$\mathbb{Y}_{i,j} = \begin{cases} 1, & i \in \{Malos\} \\ -1, & i \in \{Buenos\} \end{cases}$$

Salida: precisión de clasificación promedio de la validación cruzada para malos y buenos; puntajes de decisión para todos los registros; función de decisión.

Paso 1. Aplicar los métodos de clasificación LDA, SVM, MSD, MMD, MCLP, MC2LP

a $\mathbb{A}$. Las salidas son un conjunto de funciones de decisión, uno para cada método de clasificación.

Paso 2. Calcular la precisión de clasificación utilizando las funciones de decisión.

FIN

Las Tablas 6.33, 6.34 y 6.35 resumen los resultados de validación cruzada de los métodos de clasificación de conjunto de registros de la información crediticia de la IMF. Desde diferentes métricas de desempeño se miden diferentes aspectos de clasificación, se usan cinco criterios: precisión, score KS , errores tipo I y tipo II, y el coeficiente de correlación, para evaluar el desempeño del modelo. La precisión es una de las métricas de rendimiento de clasificación más utilizada. Es la relación entre los registros predicho correctamente y los registros completos o los registros de una clase particular.

$$\textbf{Precisión Total} = \frac{TN + TP}{TP + FP + FN + TN}$$

$$\textbf{Precisión Malos} = \frac{TN}{TN + FP}$$

$$\textbf{Precisión Buenos} = \frac{TP}{TP + FN}$$

donde:

TP (verdaderos positivos) El número de clientes buenos correctamente clasificados.

FP (falsos positivos) El número de clientes malos incorrectamente clasificados como clientes buenos.

TN (verdaderos negativos) El número de clientes malos correctamente clasificados.

FN (falsos negativos) El número de clientes buenos incorrectamente clasificados como clientes malos.

El error de tipo I se define como el porcentaje de sujetos buenos predichos que son realmente sujetos malos y el error tipo II se define como el porcentaje de sujetos malos predichos que son en realidad sujetos buenos. En las tres aplicaciones (riesgo de crédito, aprobación de crédito, y de predicción de quiebra), errores de tipo I, clasificar un cliente como bueno cuando es malo, tiene un impacto más grave que los errores de tipo II, clasificar a un cliente como malo cuando son buenos.

$$\text{Error Tipo I} = \frac{FN}{FN + TN}$$

$$\text{Error Tipo I} = \frac{FP}{FP + TP}$$

Además, una medida popular en el análisis de riesgo de crédito, el score KS, es calculado. El valor de Kolmogorov-Sminrnov (KS) mide la mayor separación entre la distribución de los clientes buenos y la distribución de los clientes malos y se define como:

$$K - S = \text{máx} \left| \text{distribución acumulada malos} - \text{distribución acumulada buenos} \right|$$

Otra medida es el coeficiente de correlación, que cae en el rango [-1, 1], se utiliza para evitar los impactos negativos de las clases desequilibradas. el coeficiente de correlación es -1 si las predicciones son completamente contrarias al valor real, 1 si las predicciones son 100 % correcto, y 0 si las predicciones se producen al azar. El coeficiente de correlación se calcula como sigue:

$$\text{Coeficiente de correlación} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

En las tablas 6.33,6.34 y 6.35 se resumen los cinco parámetros para conjuntos de entrenamiento (training) y conjuntos de prueba (testing). Los resultados de entrenamiento indican lo bien que el modelo de clasificación se ajusta al conjunto de entrenamiento, mientras que los resultados de prueba reflejan el poder de predicción real del modelo. Por lo tanto, los resultados de prueba determinan la calidad de los clasificadores.

Entre los seis métodos, SVM (radial) logra el mejor rendimiento en las medidas de la calidad de ajuste. Por otra parte LDA y MSD proporcionan buenas precisiones al clasificar los conjuntos, tasas de error, valor KS, y el coeficiente de correlación son medidas de clasificación aceptables en los mencionados modelos y además se puede observar que estos resultados se conservan tanto en la muestra de entrenamiento como en la muestra de prueba lo cual nos permite concluir que los modelos no son subestimados ni sobrestimados, es decir el ajuste de estos modelos es bueno.

De las tablas 6.33,6.34 y 6.35 se deduce que los modelos de programación lineal MSD, MCLP y MC2LP tiene un rendimiento satisfactorio. El estudio experimental indica:

1. MSD y MC2LP pueden clasificar los datos de riesgo de crédito y lograr resultados comparables con las técnicas de clasificación conocidas

2. El funcionamiento de los métodos de clasificación puede variar cuando los conjuntos de datos tienen diferentes características.

Tabla 6.33: Resumen de los diferentes métodos

		LDF		SVM		MSD		MMD		MCLP		MC2LP		
	GRUPO	BUENO	MALO	BUENO	MALO	BUENO	MALO	BUENO	MALO	BUENO	MALO	BUENO	MALO	TOTAL
ORIGINAL	BUENO	11479	602	11467	614	11466	615	8970	3111	11498	583	11469	612	12081
	MALO	1288	3491	1274	3505	1281	3498	3011	1768	1460	3319	1418	3361	4779
PROPORCIÓN	BUENO	0.95	0.05	0.9491	0.0508	0.9491	0.0519	0.7425	0.2575	0.9517	0.0483	0.9493	0.0507	1.00
	MALO	0.27	0.73	0.2665	0.7334	0.2680	0.7320	0.63	0.37	0.3055	0.6945	0.2967	0.7033	1.00
Porcentaje de buena clasificación		88.79 %		88.80 %		88.75 %		63.68 %		87.88 %		87.96 %		
Tiempo ejecución (seg.)		10		480		93.08		5.15		117.5		313.39		

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla 6.34: Calidad de ajuste de los modelos de clasificación, entrenamiento

MODELO	PREDICCION GENERAL (%)	PREDICCION DE MALOS(%)	PREDICCION DE BUENOS(%)	ERROR TIPO I(%)	ERROR TIPO II(%)	K-S	COEFICIENTE DE CORRELACION
LDA	88.79	73.0488	95.017	4.983	26.9512	0.8837	0.7155
SVM	88.8019	73.3417	94.9176	5.0824	26.6583	0.8811	0.716
MSD	88.7544	73.1952	94.9094	5.0906	26.8048	0.6998	0.7147
MMD	63.6892	36.9952	74.2488	25.7512	63.0048	0.1131	0.1117
MCLP	87.8826	69.4497	95.1742	4.8258	30.5503	0.6595	0.6905
MC2LP	87.9597	70.3285	94.9342	5.0658	29.6715	0.6595	0.6930

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla 6.35: Calidad de ajuste de los modelos de clasificación, back testing

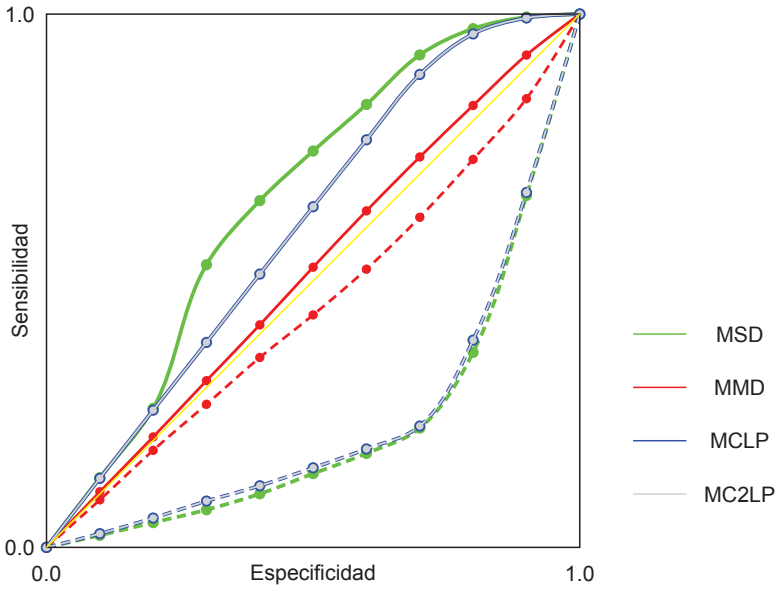
MODELO	PREDICCION GENERAL (%)	PREDICCION DE BUENOS (%)	PREDICCION DE MALOS(%)	ERROR TIPO(%) I	ERROR TIPO II(%)	K-S	COEFICIENTE DE CORRELACION
LDA	88.6833	72.093	94.9902	5.0098	27.907	0.8837	0.7069
SVM	88.9205	72.4376	95.1866	4.8134	27.5624	0.8811	0.7131
MSD	88.9205	72.4376	95.1866	4.8134	27.5624	0.6998	0.7131
MMD	63.8909	35.6589	74.6234	25.3766	64.3411	0.1131	0.1021
MCLP	87.9478	68.8200	95.2194	4.7806	31.1800	0.6509	0.6859
MC2LP	87.9004	69.4229	94.9247	5.0753	30.5771	0.6509	0.6851

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

La figura 6.9 representa las curvas ROC de los modelos de programación lineal: MMD, MSD, MCLP y MC2LP respectivamente, se puede observar que la curva del modelo MSD está más cerca de la esquina superior izquierda del diagrama, obteniendo la mayor área bajo la curva y por tanto mayor la discriminación con las curvas de los modelos MMD, MCLP y MC2LP, mientras que las curvas de los modelos MCLP y MC2LP al no estar cercanas a la diagonal principal permiten afirmar que estos métodos empleados tienen una calidad de predicción buena, por el contrario la curva del modelo MMD se encuentra cerca de la diagonal principal lo cual permite afirmar que su área bajo la curva es menor y por tanto la calidad de predicción del mismo no es buena.

De esta manera, una vez analizadas las características expuestas en esta sección para cada uno de los modelos implementados en este estudio, se determina que los modelos

Figura 6.9: Curva ROC, Modelos: MSD, MMD, MCLP y MC2LP



Elaborado por: Autor

MSD, MCLP y MC2LP proporcionan resultados óptimos en la clasificación de clientes de una entidad financiera, y que éstos son semejantes a los conseguidos con los métodos tradicionales, por lo que su aplicación contribuye a un análisis financiero más eficiente y preciso dentro de la institución.

Capítulo 7

CONCLUSIONES Y RECOMENDACIONES

7.1 Conclusiones

Los modelos de optimización basados en la teoría de la programación lineal utilizados en el presente estudio se justifican por la violación de varios supuestos en los métodos estadísticos discriminantes tradicionales. Estos supuestos son: normalidad de la distribución, la homogeneidad de la matriz de varianza-covarianza, tamaño de la muestra y la ausencia de valores atípicos en la data.

En esta línea en lo que se refiere a modelos no paramétricos en este proyecto se han explicado y se han implementado los modelos matemáticos bajo las herramientas de la programación lineal: MSD, MMD, MCLP y MC2LP, además se ha utilizado el modelo SVM en la línea de la Inteligencia Artificial cuya base es la programación cuadrática. Además se ha usado una de las técnicas y herramientas más tradicionales de la estadística multivariante como lo es el Análisis Discriminante (LDA). Estos modelos fueron aplicados al problema de clasificación, específicamente en la clasificación de sujetos de crédito en la línea de los microcréditos del sector agropecuario. Los resultados de cada uno de estos modelos fueron comparados unos con otros obteniendo lo siguiente: la tasa de error de SVM (radial) fue 11.20 % mientras que para LDA fue 11.21 % para la muestra de entrenamiento y 11.07 % y 11.31 % para el conjunto de datos de prueba, respectivamente, concluyendo que por una diferencia mínima SVM (radial) se desempeña mejor que LDA. De manera similar para MSD, MMD, MCLP y MC2LP la tasa de error fue 11.25 %, 36.31 %, 12.12 % y 12.04 % para la muestra de entrenamiento y 11.07 %, 36.10 %, 12.05 % y 12.10 % en el back testing, lo cual indica que MSD (PL) muestra un mejor comportamiento que otros modelos de programación lineal. En general, los resultados obtenidos muestran que la técnica de SVM es tan buena como la técnica LDA o la técnica MSD. Por lo tanto se concluye que las técnicas de programación lineal MSD, MCLP, MC2LP desempeñan un buen papel al tratar los problemas de clasificación.

Las diferencias en tiempo de resolución de los algoritmos desarrollados, es uno de los

factores más relevantes que hay que considerar cuando se trabaja con este tipo de algoritmos, el costo computacional varía entre algoritmos debido a factores tales como tamaño de las observaciones, restricciones, parámetros o variables e incluso al tipo de software empleado así como la versión del mismo. Para este proyecto se ha empleado un ordenador con un sistema operativo Windows 8.1 de 64 bits, procesador Intel(R) Core (TM) i7-4700MQ CPU @ 2.40GHz, 8.00 GB de memoria RAM; en el cual se han obtenido los siguiente costos computacionales: para el modelo SVM que proporciona la mejor segmentación, un costo de 480 segundos a diferencia de la técnica LDA que tiene por tiempo de ejecución 10 segundos, en este sentido el mejor modelo en cuanto a tasa de error de clasificación y tiempo de ejecución es el MSD, el cual se ejecuta en 93.8 segundos.

Los resultados de afrontar un problema de clasificación a un conjunto de datos en relación a una calificación crediticia mediante la aplicación de los diferentes métodos expuestos en este estudio, muestran que en la mayoría de los casos, el resultado de la clasificación de los modelos de programación lineal es más eficiente que el resultado del método estadístico, se comprueba que los modelos de programación lineal presentan una mayor flexibilidad al permitir al analista incorporar alguna información a priori en los modelos. Sin embargo, esto no excluye el uso de técnicas estadísticas una vez que las hipótesis requeridas sean satisfechas.

De manera general se concluye que tanto la técnica de análisis discriminante como la técnica de las máquinas de vectores de soporte y las de programación lineal son adecuadas para el estudio y predicción de la morosidad, ya que cualquiera de ellas presenta una elevada eficacia predictiva, debido a que las condiciones exigidas para la aplicación del análisis discriminante, se verifican en lo posible en la muestra de clientes objeto del estudio, por lo cual los resultados obtenidos mediante esa técnica son similares a los conseguidos mediante la aplicación de los otros métodos, que son métodos alternativos aplicados para el análisis de variables cuantitativas y cualitativas cuando se incumplen tales restricciones.

7.2 Recomendaciones

Los modelos de programación lineal multicriterio y de restricción múltiple son de gran utilidad, por tanto se recomienda su aplicación a diferentes campos de estudio, analizando otros parámetros para controlar y minimizar los tipos de errores: error tipo I y error tipo II.

En este proyecto, además de los modelos de programación lineal, las máquinas de vectores de soporte obtuvieron resultados satisfactorios en términos de predicción de

clasificación, razón por la cual se recomienda el estudio sobre la selección del mejor núcleo para mejorar los resultados obtenidos, pues el núcleo tiene un efecto significativo sobre el clasificador propuesto.

Investigaciones futuras deben enfocarse en estudiar el desempeño de las metodologías analizadas considerando otros tipos de escenarios en los cuales se pueden estudiar aspectos como: mayor número de grupos a clasificar, distintos ámbitos del crédito (comercial, microcrédito, consumo) o líneas de las tarjetas de crédito, con sus respectivas variables explicativas y medidas de desempeño sobre la tasa de clasificación errónea.

Se recomienda seguir avanzando en el estudio de métodos de clasificación no paramétricos (modelos de programación lineal multicriterio con conjuntos difusos) para evaluar el poder de predicción frente a los métodos tradicionales y poder utilizarlos como herramientas de pronóstico de calificación de riesgo.

La aplicación de tecnología crediticia como por ejemplo, desarrollo y aplicación de modelos de clasificación de clientes, consigue: mejorar la rentabilidad incrementando la tasa de aceptación y/o la reducción de la morosidad, ahorrar recursos, mejorar el servicio al cliente y apoyar a la toma de decisiones sin eliminar el criterio del analista de riesgo y la experiencia de los asesores de crédito, por lo cual se debe considerar factible aplicar modelos de clasificación en cada una de las etapas del ciclo de crédito de la entidad financiera para aprovechar la información disponible y alcanzar objetivos propios de la fase de aplicación.

Se recomienda escoger una ventana de muestreo bajo los parámetros siguientes: estabilidad, madurez y representatividad. Para medir la estabilidad se recomienda la construcción de un indicador de comportamiento, se puede considerar una variable relevante en función de los objetivos planteados, de tal forma de obtener un periodo estable del indicador para poder hacer inferencias sobre los clientes. La madurez debe garantizar que la cartera a analizar provea la suficiente información en relación a su tiempo de vida, como para concluir a cerca del comportamiento crediticio del cliente, es decir, si es bueno o malo. La representatividad hace referencia al tamaño de muestra adecuado.

Para definir la variable dependiente se recomienda el uso de una matriz de confusión entre las variables atraso promedio y atraso máximo del historial crediticio de la entidad. En base a la matriz y los porcentajes de pérdida asociados que la entidad está dispuesta a asumir se clasifica a los clientes en buenos, malos e indeterminados. Se recomienda considerar solo las definiciones de buenos y malos para construir la variable dependiente del modelo.

La selección de variables independientes se realizó mediante árboles de decisión. To-

mando en cuenta siempre que las variables explicativas estén correlacionadas con la variable dependiente, así como la representatividad de la característica en la población. Este tipo de herramientas además de identificar las variables independientes que más se correlacionan con la variable dependiente nos ayudan a encontrar combinaciones de variables donde la concentración de clientes buenos o malos aumenta significativamente.

Se recomienda, antes de iniciar con la construcción de un modelo de clasificación, realizar una auditoría informática sobre los datos a ser utilizados; con el fin de garantizar la fidelidad y la consistencia de la información.

Dentro de la estructura organizacional de la institución microfinanciera, se debe incluir un departamento de generación de información (data warehouse), para que trabaje de manera conjunta con el departamento de riesgo, pues sin la información pertinente y confiable no se pueden realizar análisis ni modelos para la administración de riesgos.

Un modelo de clasificación puede padecer cambios en su eficiencia y estructura a través del tiempo, por lo que se debe aplicar un monitoreo al mismo, con la finalidad de determinar si las aplicaciones se están calificando correctamente, si la población actual es diferente a la de construcción, si la razón de rechazo difiere o no de lo inicialmente establecido, es decir, el monitoreo se utilizará para comparar el comportamiento inicial del modelo con el comportamiento actual del mismo.

Se recomienda el análisis y discusión de la metodología propuesta como un modelo evaluación de la calificación de riesgo de crédito, que será de mucha ayuda en el control ejercido por organismos importantes como la Superintendencia de Bancos y la Superintendencia de la Economía Popular y Solidaria.

Finalmente, hay que recalcar que a pesar de que los modelos de clasificación basados en programación lineal analizados en este estudio se han aplicado directamente en el descubrimiento del conocimiento para la gestión de la cartera de crédito, se pueden utilizar en la investigación biomédica, farmacéutica y el análisis de ADN; en las telecomunicaciones, la salud, industrias para la gestión de fraude; y en las industrias para el análisis de marketing.

Bibliografía

[Aizerman et al., 1964] Aizerman, A., Braverman, A. and Rozonoer, L. (1964). *Theoretical foundations of the potential function method in pattern recognition learning*. Automation and Remote Control, 25, page 821-837.

[Apte, 2003] Apte, Chid (2003). *The big (data) dig*. OR/MS Today, February 2003.

[Ayala, 2008] Ayala G. Guillermo. (2008). *Análisis de datos con R para Ingeniería Informática*, Capítulo 6, pag. 1-20.

[Bajgier and Hill, 1982] Bajgier, S. and Hill, A. (1982). *An experimental comparison of statistical and linear programming approaches to the discriminant problem*, Decision Sciences, 13 (1982), 604-618.

[Brooke et al., 1998] Brooke, A., Kendrick, D. and Meeraus, A. (1998). *GAMS: A user's guide*. Scientific Press, California.

[Carmona, 2014] Carmona Suárez, Enrique J. (2014). *Tutorial sobre Máquinas de Vectores Soporte (SVM)*. Universidad Nacional de Educación a Distancia (UNED), Madrid.

[Carrizosa y Martín, 2005] Carrizosa, E. y Martín-Barragán, B. (2005). *Problemas de clasificación: una mirada desde la localización*. Avances en localización de servicios y sus aplicaciones. B. Pelegrín (Ed.), pp.249276. Servicio de Publicaciones de la Universidad de Murcia.

[Colin and Yiming, 2011] Colin, C. and Yiming, Y. (2011). *Learning with Support Vector Machines, Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool publishers. ISBN: 9781608456178.

[Cuadras, 2014] Cuadras Carles M. (2014). *Nuevos Métodos de Análisis Multivariante*, CMC Editions, Barcelona-España, pag.211-222.

[Ehrgott, 2005] Ehrgott, Matthias. (2005). *Multicriteria Optimization*, Springer, Segunda Edición, 2005, ISBN 3-540-21398-8, 323p.

[Freed and Glover, 1981] Freed, N. and Glover, F. (1981). *Simple but Powerful Goal Programming Models for Discriminant Problems*, European Journal Operational Research, vol. 7: 4460.

[Freed and Glover, 1986] Freed, N. and Glover, F. (1986). *Evaluating alternative linear, programming models to solve the two-group discriminant problem*, Decision Science, Vol. 17, pp. 151-162.

[Geletu, 2008] Geletu, Abebe. (2008) *GAMS - Modeling and Solving Optimization Problems*, Institute of Mathematics, Department of Operations Research & Stochastic,Ilmenau University of Technology.

[Gochet et al., 1997] Gochet, W., Stam, A., Srinivasan, V. and Chen, S. (1997). *Multi-group discriminant analysis using linear programming*. Operations Research 45, 213-225.

[González, 2002] González Luis. (2002). *Análisis discriminante utilizando máquinas núcleos de vectores soporte. Función núcleo similitud*. Tesis doctoral, Dpto. Economía Aplicada I. Universidad de Sevilla.

[González, 2002] González Luis. (2002). *Modelos de Clasificación basados en Máquinas de Vectores Soporte*. Departamento de Economía Aplicada I. Universidad de Sevilla.

[González, 2000] González, L. (2000). *Teoría del aprendizaje estadístico de la regresión. Máquinas de regresión de vector base*. Tesina, Facultad de Ciencias Económicas y Empresariales, Universidad de Sevilla.

[He et al., 2004] He, J., Liu, X., Shi, Y., Xu, W. and Yan, N. (2004). *Classifications of credit cardholder behavior by using fuzzy linear programming*. Int. J. Inf. Technol. Decis. Mak. 3(4), 633-650.

[He et al., 2010] He, J., Zhang, Y., Shi, Y. and Huang, G. (2010). *Domain-Driven Classification Based on Multiple Criteria and Multiple Constraint-Level Programming for Intelligent Credit Scoring*, IEEE Transactions on Knowledge and Data Engineering, vol. 22, No. 6:826-838.

[Hand, 1981] Hand, D. (1981). *Discrimination and classification*, John Wiley, New York.

[Hand et al., 2001] Hand, H., Mannila, H. and Smyth, P. (2001). *Principles of Data Mining*. MIT Press.

[Hardle, 2013] Hardle Simar. (2013). *Applied Multivariate Statistical Analysis*, pag. 323.

[Hastie, 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.

[Jiménez y Rengifo, 2010] Jiménez M. Leonardo y Rengifo R. Pervys. (2010). A*l Interior de una Máquina de Soporte Vectorial*. Universidad del Valle.

[Kalvelagen, 2002] Kalvelagen, Erwin. (2002). *Solving multi-objective models with GAMS*.

[Joachimsthaler and Stam, 1990] Joachimsthaler, E. and Stam, A. (1990). *Mathematical programming approaches for the classification problem in two-group discriminant analysis*. Multivariate Behavioral Research 25/4, 427-454.

[Koch, 2014] Koch, Thorsten. (2014). *Zimpl User Guide*, Zuse Institute Mathematical Programming Language.

[Koehler and Erenguc, 1990] Koehler, G. J. and Erenguc, S. S. (1990), M*inimizing Misclassifications in Linear Discriminant Analysis*. Decision Sciences, 21: 63-85.

[Kou et al., 2003] Kou, G., Liu, X., Peng, Y., Shi, Y., Wise, M. and Xu, W. (2003). *Multiple criteria linear programming to data mining: Models, algorithm designs and software developments*, Optim. Methods Softw., 18, pp. 453473.

[Kou et al., 2004] Kou, G., Peng, Y., Shi, Y., Chen, Z. and Chen, X. (2004). *A Multiple-Criteria Quadratic Programming Approach to Network Intrusion Detection*, in CASDMKM 2004, LNAI 3327, (Y. Shi, et al eds.), Springer-Verlag Berlin, pp. 145-153.

[Mangasarian, 1965] Mangasarian, O.L. (1965). *Linear and nonlinear separation of patterns by liner programming*. Operations Research 13, 444-452.

[Mangasarian, 1968] Mangasarian, O.L. (1968). *Multi-surface method of pattern separation*. IEEE Transactions on Information Theory 14(6), 801-807.

[Qi, Tian and Shi, 2012] Qi, Z., Tian, Y. and Shi, Y. (2012). *Laplacian twin support vector machine for semi-supervised classification Neural Networks*, 35 (2012), pp. 4653.

[Qi, Tian and Shi, 2012] Qi, Z., Tian, Y. and Shi, Y. (2012). *Twin support vector machine with universum data Neural Networks*, 36, pp. 112119.

[Qi, Tian and Shi, 2013] Qi, Z., Tian, Y. and Shi, Y. (2013). *Robust twin support vector machine for pattern classification Pattern Recognition*, 46 (1), pp. 305316.

[Scholkopf et al., 1997] Scholkopf, B., Sung, K., Burges, C., Girosi, F., Niyogi, P., Poggio, T. and Vapnik, V. (1997). *Comparing support vector machines with gaussian kernels to radial basis function classifiers*. IEEE Trans. Sign. Processing, 45:2758-2765.

[Shi, 2001] Shi Y. (2001). *Multiple Criteria and Multiple Constraint Levels Linear Programming: Concepts, Techniques and Applications*, World Scientific Pub Co Inc, New Jersey, USA.

[Shi, 2010] Shi Y. (2010). *Multiple Criteria Optimizationbased Data Mining Methods and Applications: A Systematic Survey, Knowledge Information System*, 24: 369-391.

[Shi et al., 2001] Shi, Y., Wise, W., Lou, M., et al. (2001). *Multiple Criteria Decision Making in Credit Card Portfolio Management*. Multiple Criteria Decision Making in New Millennium. (M. Koksalan and S. Zionts eds.), Springer-Verlag, Berlin, pp. 427-436.

[Shi et al., 2002] Shi, Y., Peng, Y., Xu, W., and Tang, X. (2002). *Data mining via multiple criteria linear programming: applications in credit card portfolio management* Int. J. Inf. Technol. Decis. Mak., pp. 131151.

[Shi et al., 2005] Shi, Y., Peng, Y., Kou, G., et al. (2005). *Classifying credit card accounts for business intelligence and decision making: a multiple-criteria quadratic programming approach*. International Journal of Information Technology and Decision Making, 2005, 4: 581-600.

[Shi et al., 2005] Shi, Y., He, J., Wang, L. and Fan, W. (2005) *Computer-based Algorithms for Multiple Criteria and Multiple Constraint Level Integer Linear Programming*, Comput. Math. Appl. 49(5): 903-921.

[Shi et al., 2008] Shi, Y., Tian, Y., Chen, X. and Zhang, P. (2008). *Regularized Multiple Criteria Linear Programs for Classification*.

[Shi et al., 2011] Shi, Y., Tian, Y., Kou, G., Peng, Y. and Li, J. (2011). *Optimization Based Data Mining: Theory and Applications, Advanced Information and Knowledge Processing*, Springer.

[Thomas, Edelman & Crook, 2002] Thomas L.C., Edelman D.B. and Crook J.N. (2002). *Credit Scoring and Its Applications*, SIAM, Philadelphia.

[Tian et al., 2012] Tian, Y., Shi, Y. and Liu, X. (2012). *Recent advances on support vector machines research Technological and Economic Development of Economy*, 18 (1), pp. 533.

[Vapnik, 1998] Vapnik V.N. (1998). *Statistical Learning Theory*. Jhon Wiley & Sons, Inc.

[Wang and Shi, 2012] Wang, B. and Shi, Y. (2012). *Error correction method in classification by using multiplecriteria and multipleconstraint levels linear programming*, International Journal of Computers, Communications and Control.

[Wang and Shi, 2013] Wang, B. and Shi, Y. (2013). *A multipleCriteria and Multiple-Constraint Levels Linear Programming Based Error Correction Classification Model*, Internacional Conference on Information Technology and Quantitive Management, ITQM.

[Weston and Watkins, 1998] Weston, J. and Watkins, C. (1998). *Multi-class support vector machines*. Techinical Report CSD-TR-98-04, Royal Holloway, University of London, Egham, UK.

[Zhang et al., 2004] Zhang, J., Zhuang, W., Yan, N., et al. (2004). *Classification of HIV-1 Mediated Neuronal Dendritic and Synaptic Damage Using Multiple Criteria Linear Programming*. Neuroinformatics, 2004, 2: 303-326.

[Zhang et al., 2010] Zhang, Z., Zhang, D., Tian, Y., and Shi, Y. (2010).*Kernel-based Multiple Criteria Linear Programming Classifier*, Procedia CS 1(1): 2407-2415.

[Zhu and Golderg, 2009] Zhu Xiaojin and Golderg Andrew B. (2009). *Introduction to SemiSupervised Learning, Synthesis Lectures on Artificial Intelligence and Machine Learning*.

Apéndice A

INTRODUCCIÓN A LA TEORÍA DE OPTIMIZACIÓN

Las Máquinas de Soporte Vectorial y otros métodos hacen un amplio uso de la teoría de optimización. En este apéndice, damos una descripción introductoria de algunos elementos de la teoría de optimización más relevantes a este tema.

Una de las principales aplicaciones de la teoría de la optimización es la maximización o minimización de una función de una o más variables (función objetivo). Esta tarea de optimización puede estar sujeta a posibles limitaciones. Un problema de optimización se puede clasificar de acuerdo con el tipo de función objetivo y las limitaciones consideradas.

En la programación lineal (LP), la función objetivo y las restricciones son funciones lineales de las variables, y estas variables son de valor real. En la programación no lineal, la función y/o limitaciones objetivo son funciones no lineales de las variables. Para la programación entera (IP), las variables se fijan a valores enteros. La programación dinámica es un tipo diferente de problema de optimización en el que tenemos un número de subtareas (por ejemplo, los buenos o los malos se mueven en un juego) y un objetivo final (ganar o perder el juego). La programación dinámica es relevante para la tarea de construir la secuencia de núcleos mencionados en la Sección 3.4.

A.1 Conceptos Básicos

La teoría de optimización o programación matemática está constituida por un conjunto de resultados y métodos analíticos y numéricos enfocados a encontrar e identificar al mejor candidato de entre una colección de alternativas, sin tener que enumerar y evaluar explícitamente todas esas alternativas. Un problema de optimización es, en general, un problema de decisión.

Los problemas de optimización se componen generalmente de estas tres partes:

- **Función objetivo**.- Es la medida cuantitativa del funcionamiento del sistema que se desea optimizar (maximizar o minimizar). Como ejemplo de funciones objetivo se pueden mencionar:
  - la minimización de los costos variables de operación de un sistema eléctrico,
  - la maximización de los beneficios netos de venta de ciertos productos,
  - la minimización del cuadrado de las desviaciones con respecto a unos valores observados,
  - la minimización del material utilizado en la fabricación de un producto.
- **Variables de decisión.-** Representan las decisiones que se pueden tomar para afectar el valor de la función objetivo. Desde un punto de vista funcional se pueden clasificar en variables independientes o principales o de control y variables dependientes o auxiliares o de estado.

  El primer elemento clave en la formulación de problemas de optimización es la selección de las variables independientes que sean adecuadas para caracterizar los posibles diseños candidatos y las condiciones de funcionamiento del sistema. Como variables independientes se suelen elegir aquellas que tienen un impacto significativo sobre la función objetivo.

  Las variables independientes se representarán mediante vectores columna de $\mathbb{R}^n$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

  o vectores fila

$$x^T = (x_1, \cdots, x_n)$$

  Aunque para los casos $n = 1, 2\ y\ 3$ se emplearán las notaciones usuales de x, (x, y) y (x, y, z) respectivamente.

- **Restricciones**.- Representan el conjunto de relaciones (expresadas mediante sistema de igualdades y desigualdades) que ciertas variables están obligadas a satisfacer. Por ejemplo, las potencias máxima y mínima de operación de un grupo de generación, la capacidad de producción de la fábrica para los diferentes productos, las dimensiones del material bruto del producto, etc.

Resolver un problema de optimización consiste en encontrar el valor que deben tomar las variables para hacer óptima la función objetivo satisfaciendo el conjunto de restricciones.

Los métodos de optimización los podemos clasificar en: métodos clásicos y métodos metaheurísticos

- **Métodos clásicos.-** De forma muy general y aproximada se puede decir que los métodos clásicos buscan y garantizan un óptimo local, dentro estos se encuentra la optimización lineal, lineal entera mixta, no lineal, estocástica, dinámica, etc.
- **Métodos heurísticos.-** incluyen los algoritmos evolutivos (genéticos entre otros), el método del recocido simulado (simulated annealing), las búsquedas heurísticas. Los métodos metaheurísticos tienen mecanismos específicos para alcanzar un óptimo global aunque no garantizan su alcance.

A.1.1 Programación Lineal

Una de las áreas más importantes y activas de la Optimización es la Programación Lineal. Los problemas que trata se basan en la optimización (minimización o maximización) de una función lineal conocida como función objetivo, sujeta a una serie de restricciones lineales de igualdad o desigualdad.

Matricialmente, un problema de PL en notación estándar (sistema de igualdades) se puede expresar como:

$$\begin{cases} \text{máx}\, z = c^T x \\ s.a. \\ Ax = b \\ x \geq 0 \end{cases} \tag{A.1}$$

donde $c^T x$ es la función objetivo a maximizar (ó minimizar), $x \in \mathbb{R}^n$ representa el vector de variables a determinar, $c \in \mathbb{R}^n$ es el vector de costos asociado a las variables, $A \in \mathbb{M}_{m \times n}$ es la matriz de coeficientes y $b \in \mathbb{R}^m$ el vector de términos independientes (ó rhs) relativos a las restricciones. Se puede introducir una variable de holgura para transformar una desigualdad como $x \leq b$ en una igualdad $x + s = b$ e igualmente unas variables superávit para transformar $x \geq b$ en $x - s = b$.

Definición 6. ***Región factible****.- se llama región factible del problema al conjunto de posibles valores que satisfacen todas las restricciones,* $\mathcal{R} = \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$. *Una solución del problema (A.1) se dice* ***solución factible*** *si satisface todas las*

restricciones del mismo, es decir, $x \in \mathcal{R}$. Una solución factible se dice ***solución óptima*** *si proporciona el valor más favorable a la función objetivo, es decir, $x^* \in \mathcal{R}$ es óptima si $\forall x \in \mathcal{R}$, $c^T x^* \geq c^T x$.*

La región factible se dice que constituyen un *conjunto convexo* de puntos, ya que, si trazamos una línea que une dos puntos de la región, la línea se encuentren por completo dentro de la región factible.

Conjuntos Convexos

El concepto de convexidad es de gran importancia en el estudio de los problemas de optimización desde el punto de vista de la aplicación práctica, puesto que en algunos casos, bajo condiciones de convexidad, se puede garantizar que un extremo local de un problema es realmente un extremo global y por tanto la solución óptima del problema. Se describen en esta sección algunos conceptos básicos de convexidad útiles para el desarrollo de la programación matemática y aunque es posible definirlos en el ámbito de cualquier espacio topológico, en lo sucesivo consideraremos el espacio vectorial $\mathbb{R}^n$.

Definición 7. ***(Segmento lineal)*** *Dados dos puntos $x, y \in \mathbb{R}^n$, segmento lineal cerrado que une x con y es el conjunto definido por*

$$[x, y] = \{\lambda x + (1 - \lambda)y \in \mathbb{R}^n : 0 \leq \lambda \geq 1\}$$

del mismo modo, el segmento lineal abierto que une x con y es el conjunto

$$(x, y) = \{\lambda x + (1 - \lambda)y \in \mathbb{R}^n : 0 < \lambda < 1\}$$

Definición 8. ***(Conjunto convexo)****. Sea $\Omega \subset \mathbb{R}^n$ entonces*

$$\Omega \textit{ es convexo ssi } \forall x, y \in \Omega, [x, y] \subseteq \Omega$$

Esta definición se interpreta de forma que un conjunto será convexo si el segmento lineal cerrado que une cualquier par de puntos del conjunto está contenido en dicho conjunto.

Funciones convexas

En esta sección se proporciona la definición de función convexa y se presentan alguna de sus propiedades más importantes, sobre todo aquellas que pueden utilizarse para resolver problemas de optimización.

Definición 9. ***(Función convexa)*** *Diremos que la función $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$, con Ω un conjunto convexo no vacío,*

f *es convexa sobre* Ω ssi $\forall x, y \in \Omega, \lambda \in [0,1] \Rightarrow f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda)f(y)$

se dice que f es estrictamente convexa sobre Ω

$$ssi\ \forall x, y \in \Omega, \lambda \in [0,1] \Rightarrow f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$$

Definición 10. ***(Función cóncava)*** *Diremos que la función* $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$*, con* Ω *un conjunto convexo no vacío es cóncava sobre* Ω *ssi* $g = -f$ *es convexa. Esta definición es equivalente a decir que*

f es cóncava ssi $\forall x, y \in \Omega, \lambda \in [0,1] \Rightarrow f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda)f(y)$

Teorema 8. *Sea* $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ *con* Ω *un conjunto convexo. Si* $f(x)$ *es convexa (cóncava), entonces el conjunto* Γ *donde* $f(x)$ *alcanza su mínimo (máximo) es convexo y cualquier mínimo (máximo) local de* $f(x)$ *es mínimo (máximo) global.*

A.2 Dualidad

Asociado a un problema de programación lineal aparece otro problema de programación lineal denominado problema dual (asociado al primero), en general se refiere al primero como *problema primal* y se denota mediante [**P**], y al segundo como *problema dual* y se denota con [**D**].

Definición 11. *Un problema de programación lineal se dice que está en forma simétrica si todas las variables están restringidas a ser no negativas y todas las restricciones son de tipo "*$\leq$*"en el caso del máximo y de tipo "*$\geq$*" en el caso del mínimo, es decir, toman la forma:*

$$\begin{cases} \text{máx}\, Z = c^T x \\ s.a.: \\ Ax \leq b \\ x \geq 0 \end{cases} \qquad \begin{cases} \text{mín}\, Z = c^T x \\ s.a.: \\ Ax \geq b \\ x \geq 0 \end{cases}$$

Definición 12. *Dado un problema de programación lineal en forma simétrica de maxi-*

mización:

$$[P]\begin{cases}\text{máx}\, Z = c^T x \\ s.a.: \\ Ax \leq b \\ x \geq 0\end{cases}$$

entonces su problema dual asociado es

$$[D]\begin{cases}\text{mín}\, \tilde{Z} = b^T y \\ s.a.: \\ A^T y \geq c \\ y \geq 0\end{cases}$$

*donde **y** es un vector de m componentes, (tantas como restricciones en el problema primal).*

Teorema 9. ***(De la dualidad débil)**. Si **x** e **y** son soluciones factibles de un par de problemas primal $[P]$ y dual $[D]$, respectivamente, entonces se verifica*

$$Z(x) = c^T x < b^T y = \tilde{Z}(y)$$

Es decir, para cualquier par de soluciones factibles de un problema primal y su dual el valor de la función objetivo del problema primal es menor o igual que el valor de la función objetivo del problema dual.

Corolario 2. *El valor de la función objetivo del problema primal de maximización para cualquier solución factible es una cota inferior del mínimo valor de la función objetivo de su problema dual asociado, y recíprocamente, el valor de la función objetivo del problema dual de minimización para cualquier solución factible dual es una cota superior del máximo valor de la función objetivo del primal.*

Corolario 3. *Si el problema primal es factible y su solución es no acotada, $(Z \to \infty)$, entonces el problema dual no tiene solución. Análogamente, si el problema dual es factible pero con solución no acotada, $(\tilde{Z} \to \infty)$, entonces el problema primal no tiene solución factible.*

Teorema 10. ***(Criterio de Optimalidad).** Si existen soluciones factibles para los problemas primal $[P]$ y dual $[D]$ tales que los valores correspondientes de las funciones objetivo coinciden, entonces dichas soluciones factibles son óptimas para sus respectivos problemas*

Corolario 4. *Si el problema primal tiene solución óptima finita entonces el problema dual también, y recíprocamente, si el problema dual tiene solución óptima finita entonces el primal también.*

Corolario 5. *Si el problema primal es no factible entonces su dual o no tiene solución o tiene solución no acotada, y recíprocamente, si el dual es no factible entonces el primal o es no factible o tiene solución no acotada.*

Teorema 11. ***(Teorema fundamental de la dualidad)**. Dado un problema primal y su dual sólo una de las siguientes afirmaciones es cierta:*

P tiene solución óptima (finita) $\Leftrightarrow$ D tiene solución óptima (finita).

P es no factible $\Rightarrow$ D es no factible o D tiene solución no acotada.

D es no factible $\Rightarrow$ P es no factible o P tiene solución no acotada.

P tiene solución no acotada $\Rightarrow$ D es no factible.

D tiene solución no acotada $\Rightarrow$ P es no factible.

Así, para cada problema de programación lineal primal, hay una formulación dual. Si se resuelve, este problema dual puede dar solución al problema primal original. Los problemas primal y dual están estrechamente relacionados y tienen la importante propiedad de que los valores óptimos de ambas funciones objetivo son iguales a la solución. Para una máquina de vectores de soporte, la formulación primal implica optimización sobre el vector de pesos ω mientras que la formulación dual implica la optimización sobre los α_i. Curiosamente, la dimensionalidad de ω es el número de características mientras que α_i es indexado por el número de observaciones en la muestra. Por lo tanto, para el problema dual, el proceso de optimización es menos intensivo computacionalmente, ya que se realiza sobre un menor número de parámetros.

A.3 Optimización con restricciones

A manera de respuesta a dos de las observaciones más frecuentes de la PL, a saber, la restrictividad de la hipótesis de linealidad y la dificultad de definir una única función objetivo, surge la Programación no lineal (PNL) y la Teoría de la Decisión Multicriterio. Realmente, un supuesto importante en PL es que todas sus funciones (objetivo y restricciones) son lineales. Aunque, en esencia, esta hipótesis se cumple en el caso de muchos problemas prácticos, con frecuencia no es así. Por lo que es necesario abordarlo desde la Programación No Lineal (PNL).

De manera general, un problema de Programación No Lineal consisten en encontrar $x = (x_1, \cdots, x_n)$ tal que

$$\begin{cases} \text{máx } f(x) \\ s.a. \\ g_i(x) \leq b_i, \ \forall i = 1, \cdots, m \\ x \geq 0 \end{cases}$$

donde $f(x)$ y $g_i(x)$ son funciones dadas de n variables de decisión.
En notación extendida:

$$\begin{cases} \text{máx } f(x_1, \cdots, x_n) \\ s.a. \\ g_1(x_1, \cdots, x_n) \leq b_1 \\ g_2(x_1, \cdots, x_n) \leq b_2 \\ \vdots \\ g_m(x_1, \cdots, x_n) \leq b_m \\ x_j \geq 0 \ \forall j = 1, \cdots, n \end{cases}$$

Existen muchos tipos de problemas de PNL, en función de las características de estas funciones, por lo que se emplean varios algoritmos para resolver los distintos tipos. Para ciertos casos donde las funciones tienen formas sencillas, los problemas pueden resolverse de manera relativamente eficiente. En algunos otros casos, incluso la solución de pequeños problemas representa un verdadero reto.

En Programación No Lineal un máximo local no es necesariamente un máximo global y en general, los algoritmos de PNL no pueden distinguir cuando se encuentra en un óptimo local o en uno global. Por tanto, es crucial conocer las condiciones en las que se garantiza que un máximo local es un máximo global en la región factible. Recuerde que en Análisis Matemático, cuando se maximiza una función ordinaria (doblemente diferenciable) de una sola variable $f(x)$ sin restricciones, esta garantía está dada cuando

$$\forall, \quad \frac{d^2 f}{dx^2} \leq 0$$

es decir, cuando la función es cóncava hacia abajo, o simplemente cóncava.

Si un problema de PNL no tiene restricciones, el hecho de que la función objetivo sea cóncava garantiza que un máximo local es un máximo global. (De igual manera, una función objetivo convexa asegura que un mínimo local es un mínimo global.) Si existen restricciones, se necesita una condición más para dar esta garantía, a saber, que la

región factible sea un conjunto convexo. Por esta razón, los conjuntos convexos tienen un papel fundamental en la programación no lineal.

A.3.1 Optimización Restringida

La formulación de este tipo de problemas responde a la formulación general presentada al comienzo de esta sección. El estudio del problema con restricciones comenzó abordando solamente el problema con restricciones de igualdad, teniendo sus orígenes en el siglo XVIII. El problema con restricciones de desigualdad tiene una historia más reciente, de hecho hasta los años cincuenta del pasado siglo no se tratan éstos.

Nótese que los problemas con restricciones de desigualdad permiten reflejar la realidad en términos matemáticos mejor que los problemas con restricciones de igualdad, ya que no "limitan" tanto la elección de los valores de las variables de decisión. Los problemas con restricciones de igualdad suelen considerarse como poco realistas debido a lo restrictivo de su planteamiento. Sin embargo, el estudio de ellos se considera interesante ya que es de utilidad en distintas áreas de conocimiento como Economía, Estadística...

Para el caso restringido general, se estudia las condiciones de Karush-Kuhn-Tucker (condiciones KKT), que fueron desarrolladas de manera independiente por Karush (tesis de master, 1939) y por Kuhn y Tucker (1951).

Cuando se estudió las Máquinas de Vectores de Soporte en el capítulo 3, obsérvese que la tarea de aprendizaje involucra la optimización de una función objetivo 3.21. En el contexto de la teoría de optimización, esto requiere la introducción de los multiplicadores de Lagrange α_i para manipular las restricciones. En esta sección, se revisa brevemente la optimización restringida para proporcionar alguna formación más matemática.

Para propósitos ilustrativos, se considera un problema simple de optimización de dos variables con función objetivo $z = f(x, y)$ y con las variables componentes x e y para cumplir la restricción de igualdad $g(x, y) = 0$.

Esta ecuación define implícitamente una relación entre x e y, que podríamos escribir $y = h(x)$, y por lo tanto:

$$z = f(x, y) = f(x, h(x)) \tag{A.2}$$

que ahora es una función de una variable independiente x. En un óptimo de z, se tiene:

$$\frac{dz}{dx} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx} = 0 \tag{A.3}$$

Además, de

$$dg(x,y) = \left(\frac{\partial g}{\partial x}\right) dx + \left(\frac{\partial g}{\partial y}\right) dy = 0$$

se deduce:

$$\frac{dy}{dx} = -\left[\frac{\partial g}{\partial y}\right]^{-1} \frac{\partial g}{\partial x} \tag{A.4}$$

Por lo tanto:

$$\frac{\partial f}{\partial x} - \frac{\partial f}{\partial y}\left[\frac{\partial g}{\partial y}\right]^{-1} \frac{\partial g}{\partial x} = 0 \tag{A.5}$$

Si se define:

$$\lambda = -\left[\frac{\partial g}{\partial y}(x,y)\right]^{-1} \left[\frac{\partial f}{\partial y}(x,y)\right] \tag{A.6}$$

entonces, en el óptimo, se tiene:

$$\frac{\partial f}{\partial x}(x,y) + \lambda \frac{\partial g}{\partial x}(x,y) = 0 \tag{A.7}$$

Un argumento similar con x e intercambiada por y da:

$$\frac{\partial f}{\partial y}(x,y) + \lambda \frac{\partial g}{\partial y}(x,y) = 0 \tag{A.8}$$

y, por supuesto, tenemos $g(x,y) = 0$. Estas tres condiciones pueden ser generadas a partir de una función de Lagrange:

$$F(x,y,\lambda) = f(x,y) + \lambda g(x,y) \tag{A.9}$$

como los tres derivados con respecto a x, y y λ, respectivamente. $\lambda?$ es un multiplicador de Lagrange y la función objetivo ha sido corregida, como la suma del objetivo original y λ multiplicando la restricción de igualdad. Esto generaliza a n variables y m restricciones de igualdad con las correspondientes funciones de Lagrange:

$$F(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) \tag{A.10}$$

También se puede incorporar fácilmente restricciones de desigualdad. Desde una restricción $c_i(x) \geq b_i$ puede ser visto como $-c_i(x) \leq -b_i$, se puede considerar genéricamente todas las desigualdades de la forma:

$$g_i(x) \leq b_i \tag{A.11}$$

Estas desigualdades se pueden transformar en restricciones de igualdad mediante la adición de variables de holgura no negativos u_i^2 (cuadrado para asegurar la positividad):

$$g_i(x) + u_i^2 - b_i = 0 \tag{A.12}$$

por lo que el problema es optimizar $f(x)$ sujeto a las restricciones de igualdad $g_i(x) + u_i^2 - b_i = 0$ con la función de Lagrange:

$$F(x, \lambda, u) = f(x) + \sum_{i=1}^{m} \lambda_i \left[g_i(x) + u_i^2 - b_i \right] \tag{A.13}$$

Las condiciones necesarias que deben cumplir en un punto fijo son entonces:

$$\frac{\partial F}{\partial x_j} = 0 = \frac{\partial f}{\partial x_j} + \sum_{i=1}^{m} \lambda_i \frac{\partial g_i}{\partial x_j}; \quad j = 1, 2, \cdots, n \tag{A.14}$$

$$\frac{\partial F}{\partial \lambda_i} = 0 = g_i(x) + u_i^2 - b_i; \quad i = 1, 2, \cdots, m \tag{A.15}$$

$$\frac{\partial F}{\partial u_i} = 0 = 2\lambda_i u_i; \quad i = 1, 2, \cdots, m \tag{A.16}$$

La última condición, cuando se multiplica por $u_i/2$, da:

$$\lambda_i u_i^2 = 0; \quad i = 1, 2, \cdots, m \tag{A.17}$$

es decir:

$$\lambda_i [b_i - g_i(x)] = 0 \tag{A.18}$$

Esta última ecuación indica que, o bien λ_i o $(b_i - g_i(x^*))$ es cero, donde x^* denota el valor de x en el óptimo. Si el multiplicador de Lagrange λ_i no es cero, entonces $g_i(x^*) = b_i$, y la restricción se dice que es *activa*. Por otro lado, si $g_i(x^*) < b$, entonces el correspondiente multiplicador de Lagrange λ_i debe ser cero. Para una máquina de vectores de soporte, cuando el multiplicador de Lagrange α_i no es cero, entonces $y_i[\omega \cdot z + b] = 1$ correspondiente a i es un vector de soporte. Cuando $\alpha_i = 0$ entonces $y_i[\omega \cdot z + b] > 1$ e i no es un vector de soporte y este punto de datos no hace una contribución a la discusión dentro de la función de decisión.

Los multiplicadores de Lagrange se ven limitados en signo como consecuencia del siguiente argumento. A partir $g_k(x) + u_k^2 = b_k$:

$$\frac{\partial g_k}{\partial b_i} = \begin{cases} 0 \; si \; i \neq k \\ 1 \; si \; i = k \end{cases} \tag{A.19}$$

Así, si tenemos en cuenta el óptimo de la función de Lagrange con respecto a b_i

$$\frac{\partial F}{\partial b_i} = \frac{\partial f}{\partial b_i} + \sum_{k=1}^{m} \lambda_k^* \frac{\partial g_k}{\partial b_i} = \frac{\partial f}{\partial b_i} + \lambda_i^* = 0 \tag{A.20}$$

donde λ_i^* es el valor de λ_i en el óptimo. Por lo tanto:

$$\frac{\partial f}{\partial b_i} = -\lambda_i^* \tag{A.21}$$

Si b_i se incrementa, la región de restricción se amplía de modo que el mínimo de $f(x)$ podría seguir siendo el mismo o podría disminuir en valor. Por tanto, se deduce que:

$$\frac{\partial f}{\partial b_i} \leq 0 \tag{A.22}$$

y por lo que $\lambda_i^* \geq 0$. Así, por restricciones de desigualdad, en un mínimo de $f(x)$ las siguientes condiciones se satisfacen:

$$\frac{\partial f}{\partial x_j} + \sum_{i=1}^{m} \lambda_i \frac{\partial g_i}{\partial x_j} = 0, \quad j = 1, 2, \cdots, n. \tag{A.23}$$

$$g_i(x) \leq b_i, \quad i = 1, 2, \cdots, m \tag{A.24}$$

$$\lambda_i [g_i(x) - b_i] = 0, \quad i = 1, 2, \cdots, m \tag{A.25}$$

$$\lambda_i \geq 0, \quad i = 1, 2, \cdots, m \tag{A.26}$$

Este es un ejemplo de las *condiciones Karush-Kuhn-Tucker (KKT)*: la serie completa de las condiciones necesarias para ser satisfecha en un óptimo. Para la maximización, este argumento llevaría a $\lambda_i \leq 0$.

Sin embargo, para nuestra proposición en la tarea de aprendizaje de SVM en (3.14), tenemos que restar el segundo término que incorpora los multiplicadores de Lagrange y de ahí el requisito es $\alpha_i \geq 0$. Dada la formulación primal del clasificador SVM en (3.14):

$$L = \frac{1}{2} \langle \omega, \omega \rangle - \sum_{i=1}^{m} \alpha_i \left[y_i (\langle \omega, x_i \rangle + b) - 1 \right] \tag{A.27}$$

Las condiciones KKT son por lo tanto

$$\frac{\partial L}{\partial \omega} = \omega - \sum_{i=1}^{m} \alpha_i y_i x_i = 0$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{m} \alpha_i y_i = 0$$

$$y_i(\omega \cdot x_i + b) \geq 1$$

$$\alpha_i \left[y_i(\langle \omega, x_i \rangle + b) - 1 \right] = 0$$

$$\alpha_i \geq 0$$

En resume el siguiente teorema, denominado teorema de Karush-Kuhn-Tucker establece las condiciones suficientes (también conocidas como condiciones KKT) para que un punto x^* sea solución del problema primal.

Teorema 12. ***(Teorema de Karush-Kuhn-Tucker)*** *Si en el problema primal*

$$\begin{cases} \text{mín } f(x), \quad x \in \Omega \\ s.a.: \\ g_i(x) \leq 0, \quad i = 1, 2, \cdots, m \end{cases} \tag{A.28}$$

las funciones $f : \mathbb{R}^d \to \mathbb{R}$ *y* $g_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1, 2, \cdots, m$ *son todas ellas funciones convexas, y existen constantes* $\alpha_i \geq 0,\ i = 1, 2, \cdots, m$ *tales que:*

$$\frac{\partial f(x^*)}{\partial x_j} + \sum_{i=1}^{m} \alpha_i \frac{\partial g_i(x^*)}{\partial x_j} = 0, \quad j = 1, 2, \cdots, d \tag{A.29}$$

$$\alpha_i g_i(x^*) = 0, \quad i = 1, 2, \cdots, m \tag{A.30}$$

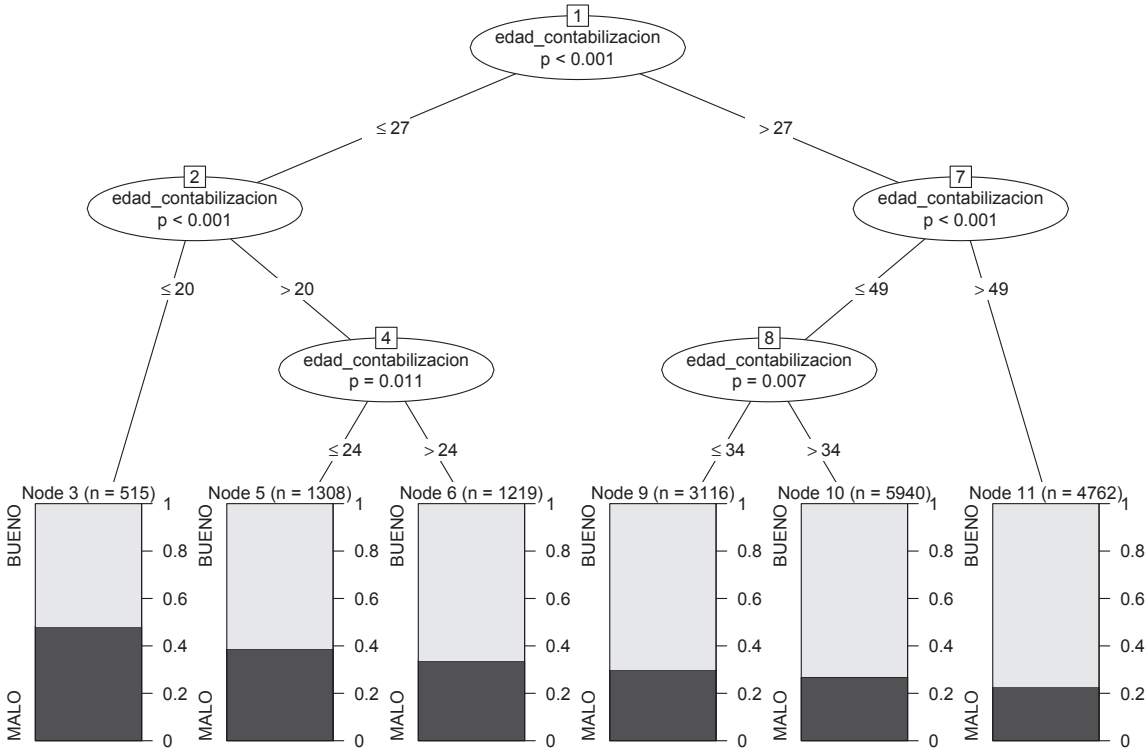
entonces el punto x^* *es un mínimo global del problema primal.*

Apéndice B

TABLAS Y GRÁFICOS

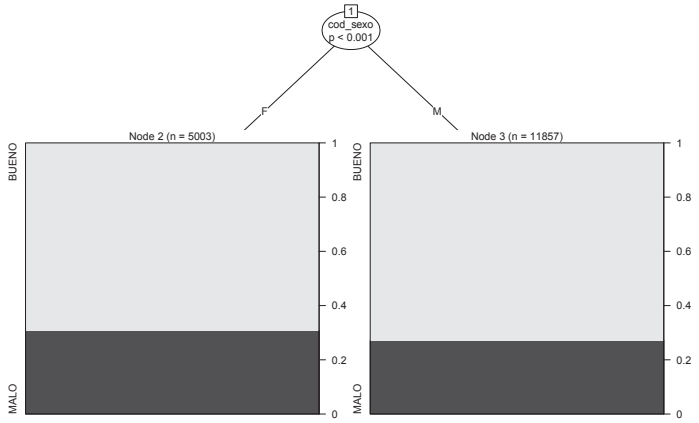
B.1 ÁRBOLES DE DECISIÓN

Figura B.1: Árbol de decisión: Variable edad



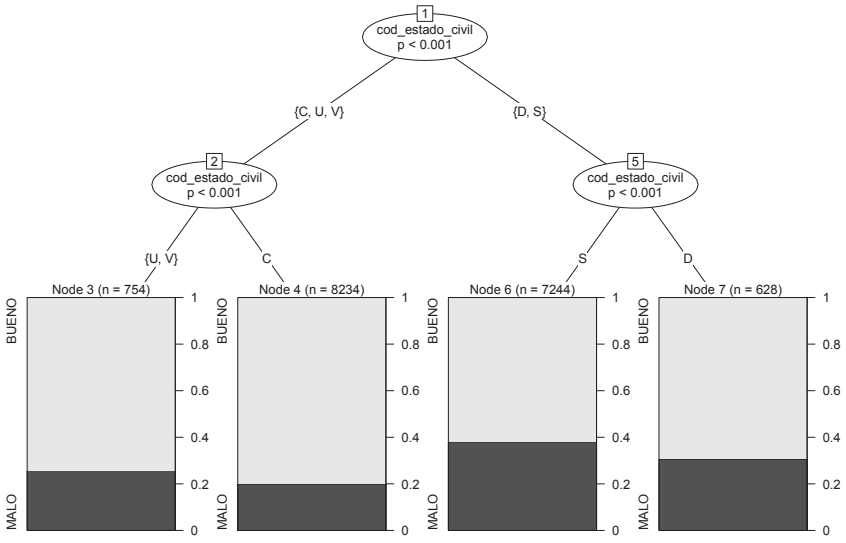
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.2: Árbol de decisión: Variable sexo



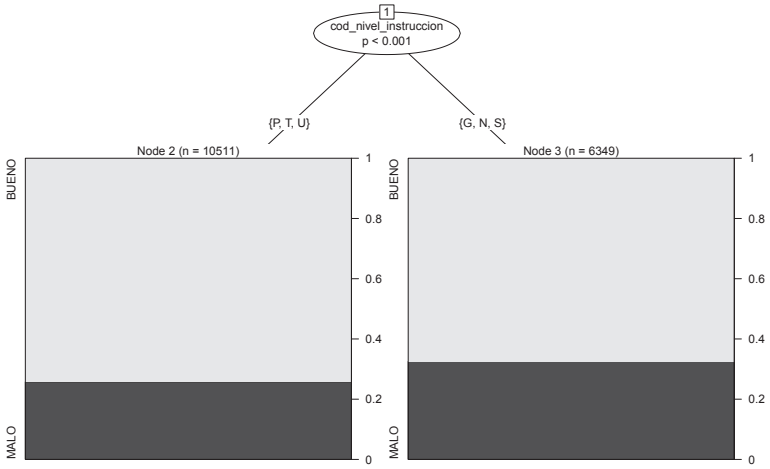
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.3: Árbol de decisión: Variable Estado Civil



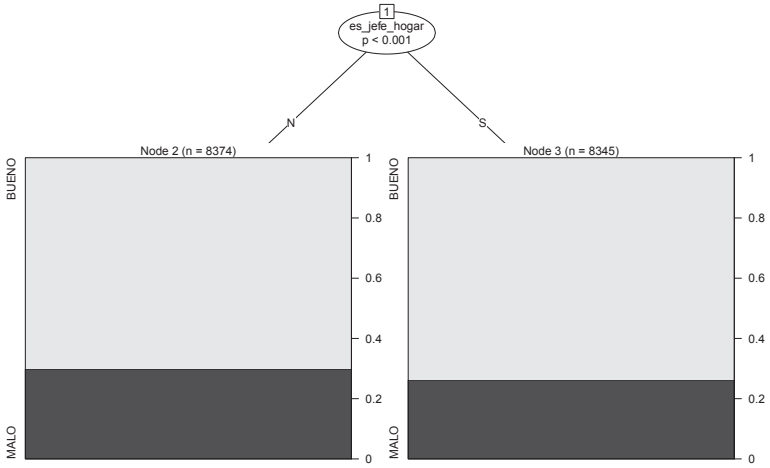
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.4: Árbol de decisión: Variable Nivel de Instrucción



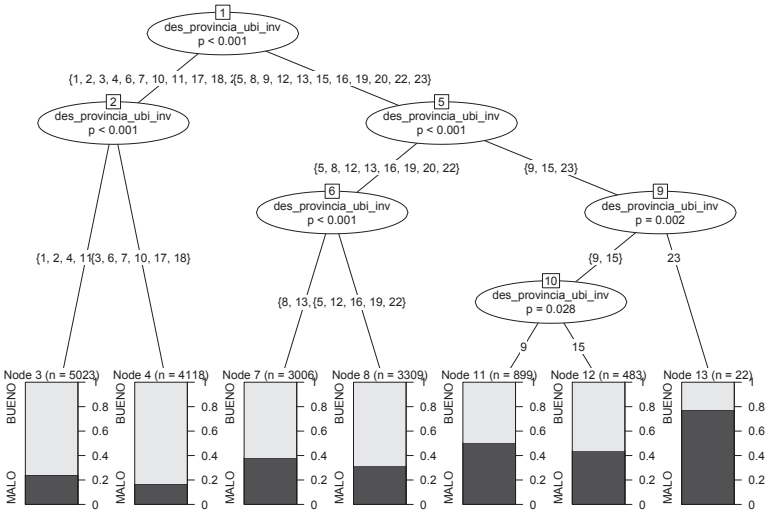
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.5: Árbol de decisión: Variable Jefe de Hogar



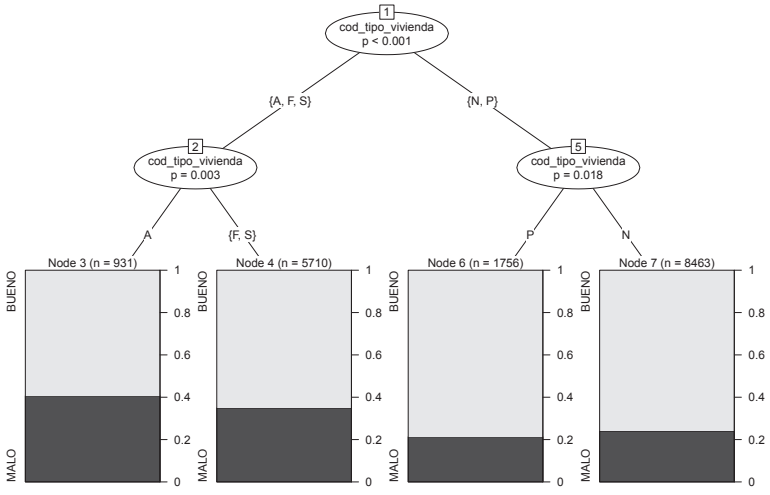
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.6: Árbol de decisión: Provincia de ubicación de la inversión



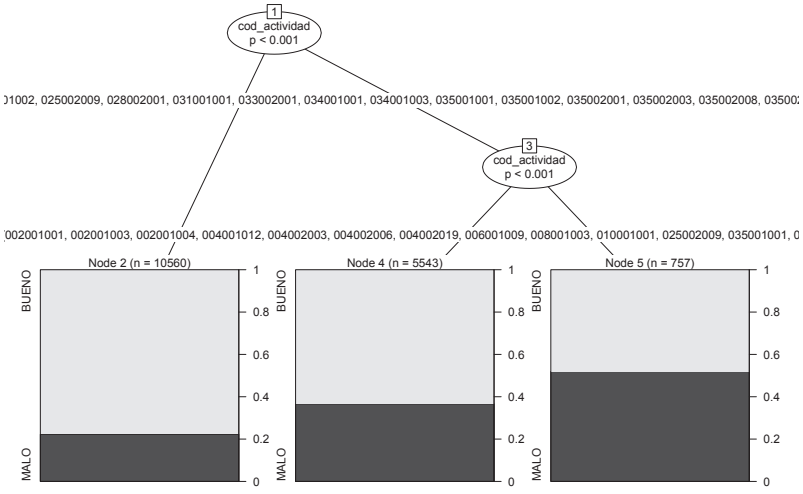
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.7: Árbol de decisión: Variable tipo de vivienda



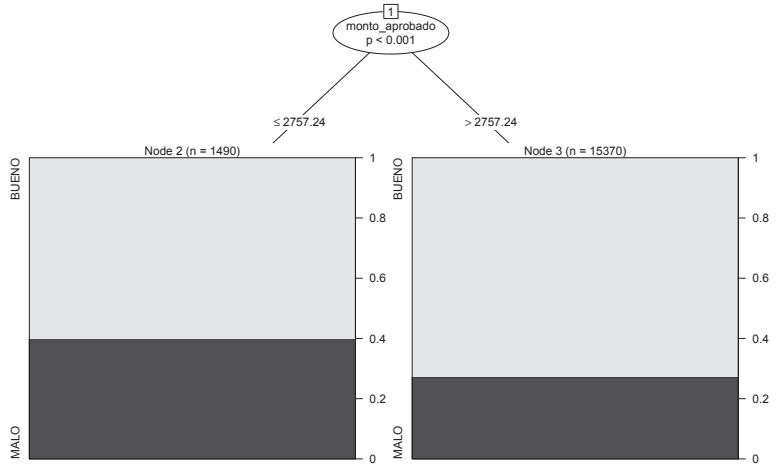
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.8: Árbol de decisión: Variable actividad económica



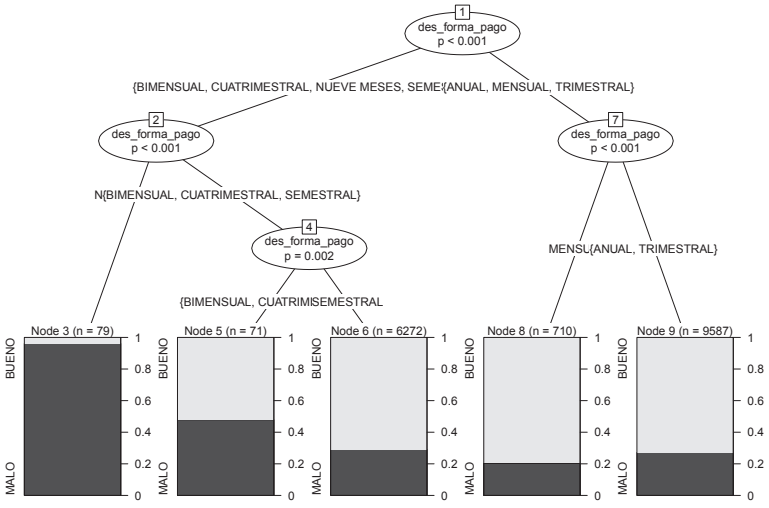
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.9: Árbol de decisión: Variable monto aprobado (exposición)



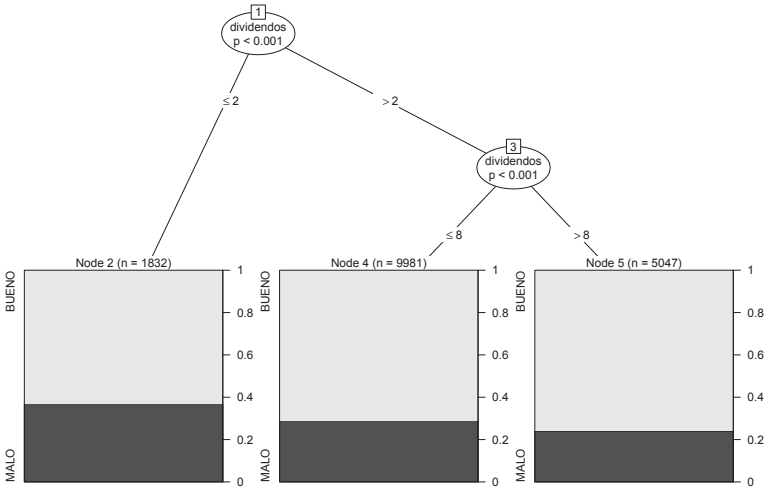
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.10: Árbol de decisión: Variable forma de pago



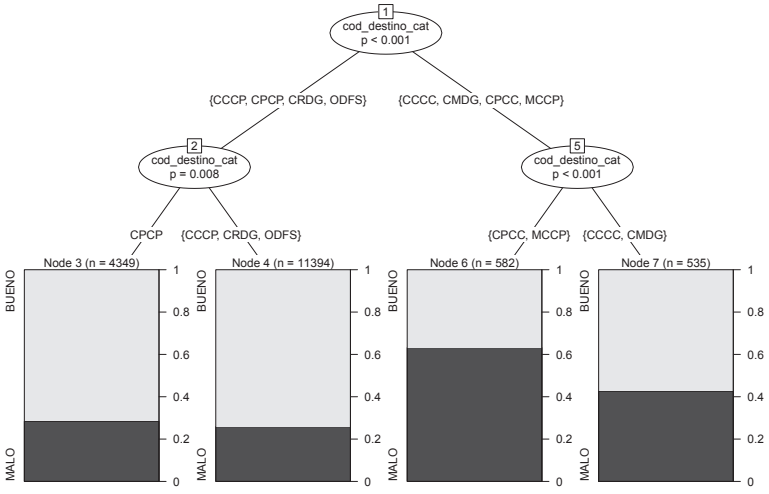
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.11: Árbol de decisión: Variable dividendos



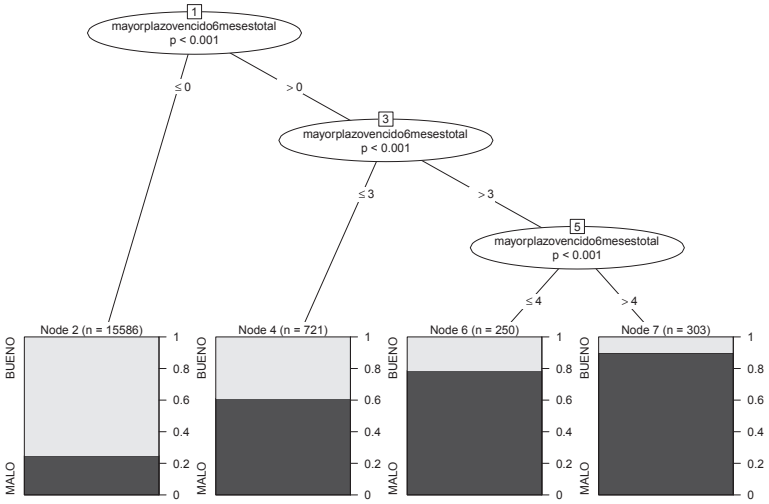
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.12: Árbol de decisión: Variable destino final categorizado



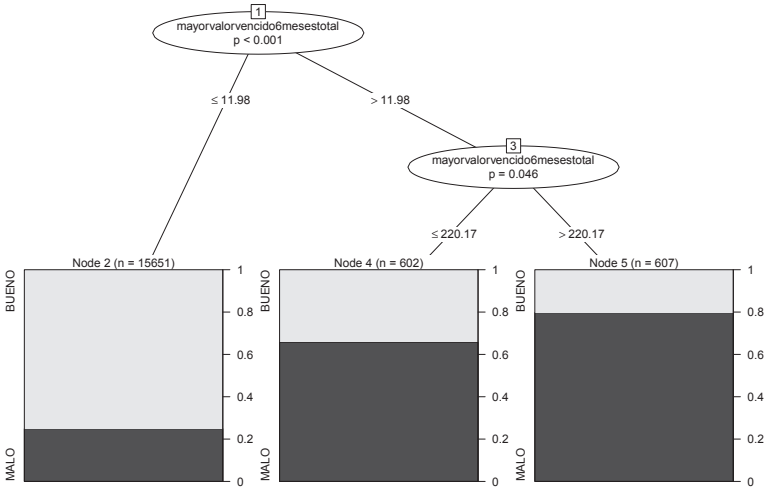
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.13: Árbol de decisión: Variable mayor plazo vencido 6 meses total



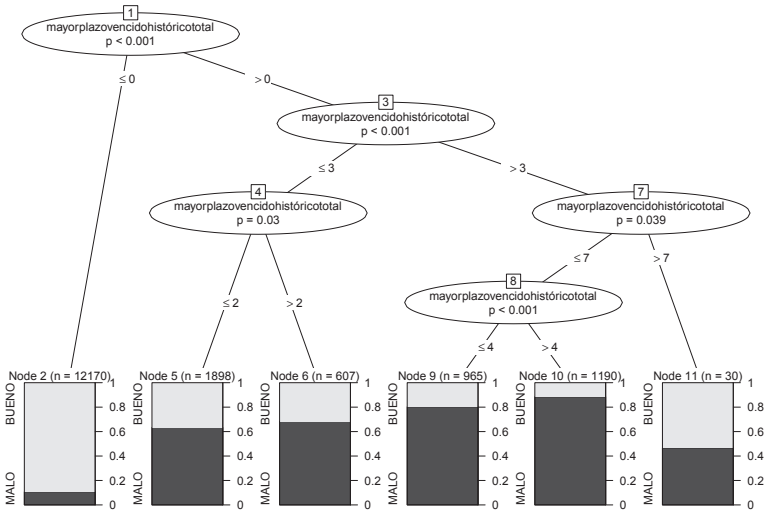
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.14: Árbol de decisión: Variable mayor valor vencido 6 meses total



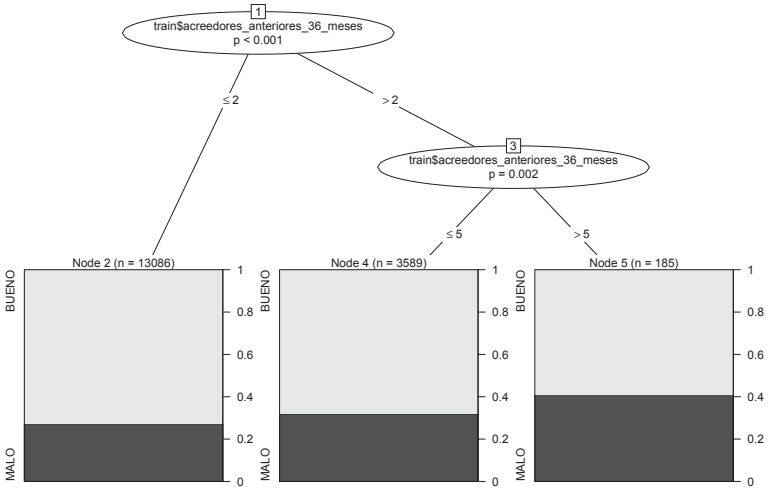
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.15: Árbol de decisión: Variable mayor plazo vencido histórico total



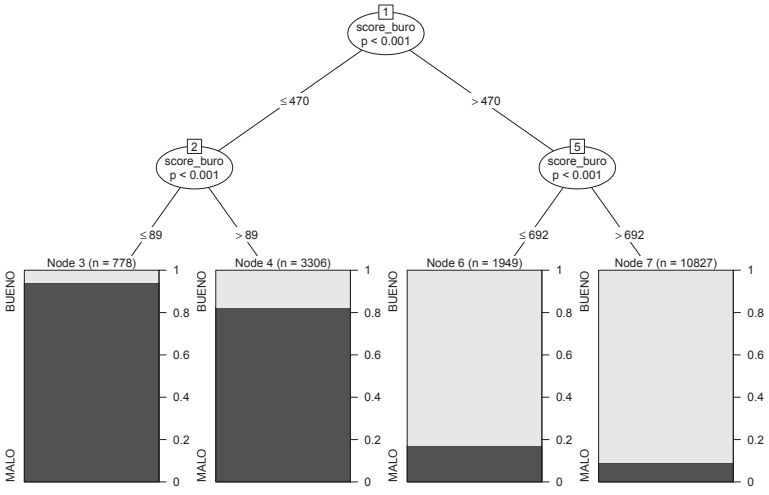
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.16: Árbol de decisión: Variable número de acreedores anteriores 36 meses



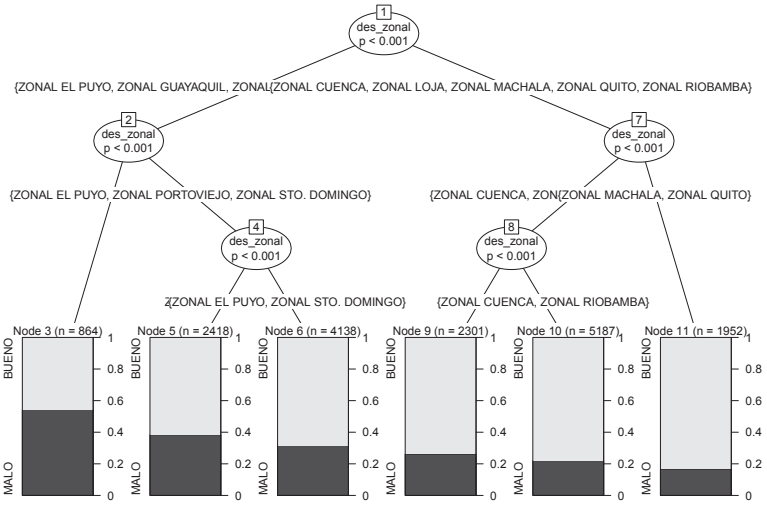
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.17: Árbol de decisión: Variable score de buró



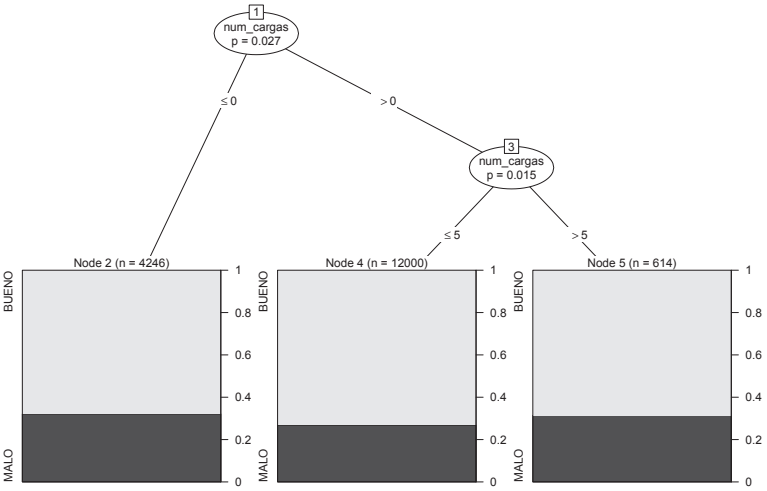
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.18: Árbol de decisión: Variable zonal



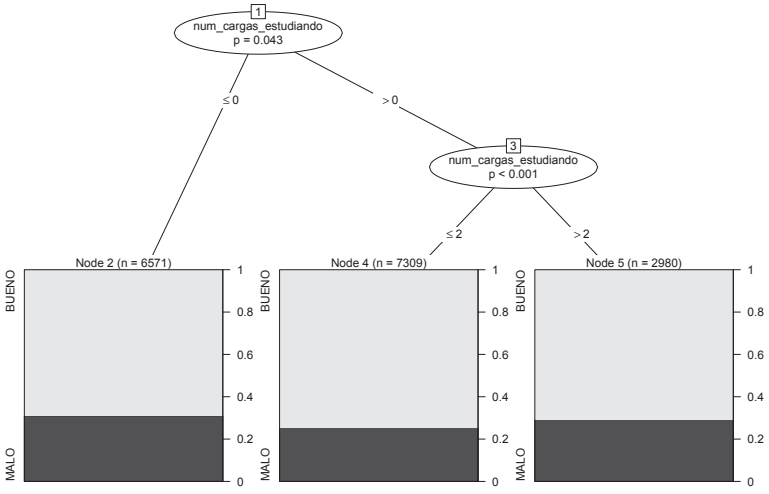
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.19: Árbol de decisión: Variable número de cargas



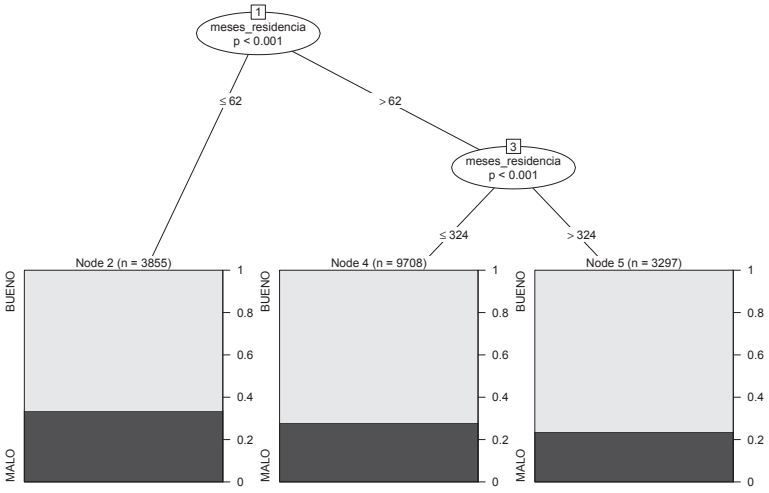
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.20: Árbol de decisión: Variable número de cargas estudiando



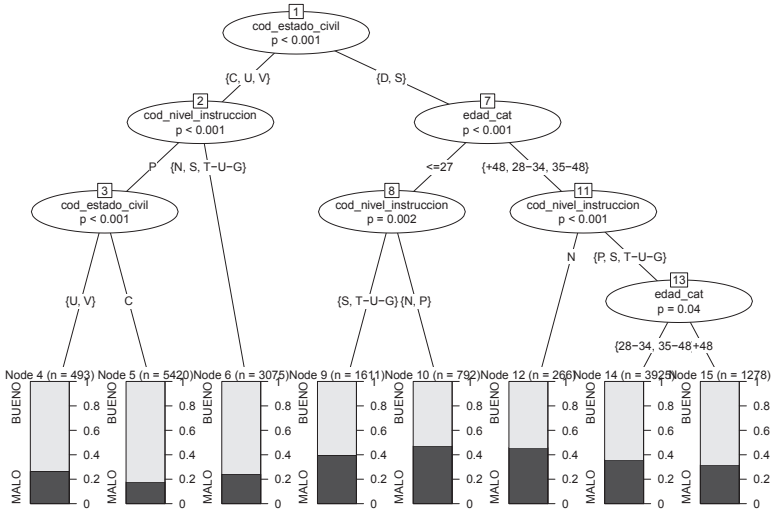
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.21: Árbol de decisión: Variable meses residencia



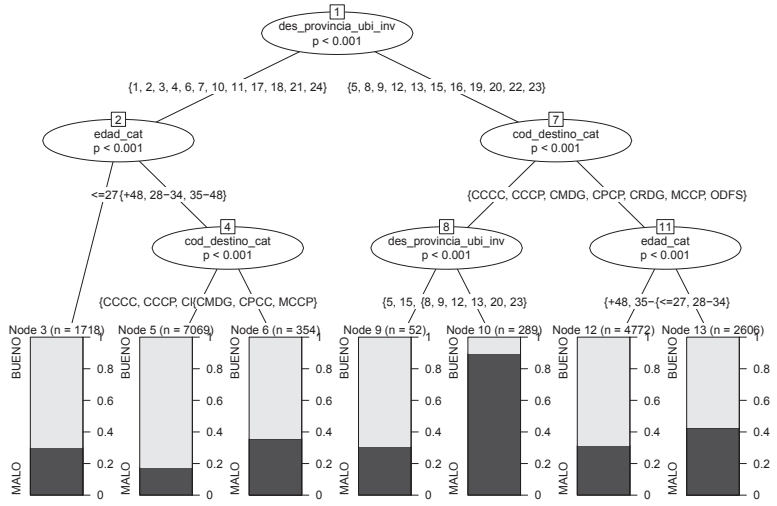
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.22: Árbol de decisión cruce variables:Edad, Estado Civil, Nivel de Instrucción



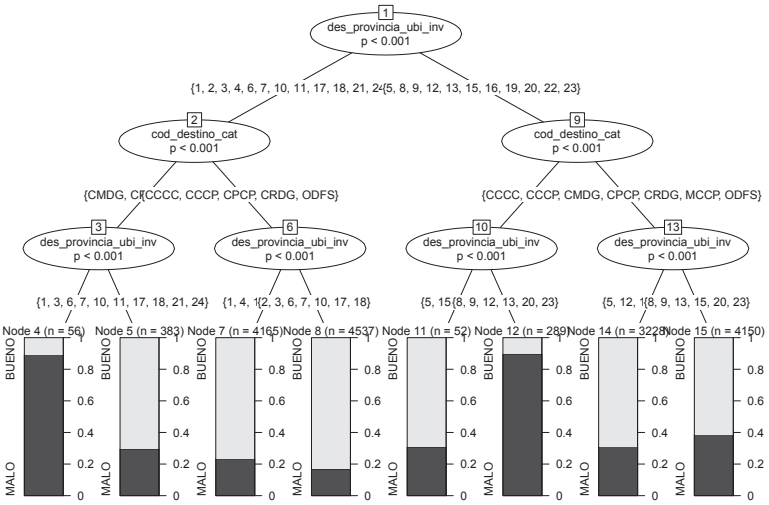
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.23: Árbol de decisión variables: Destino y provincia inversión, Edad



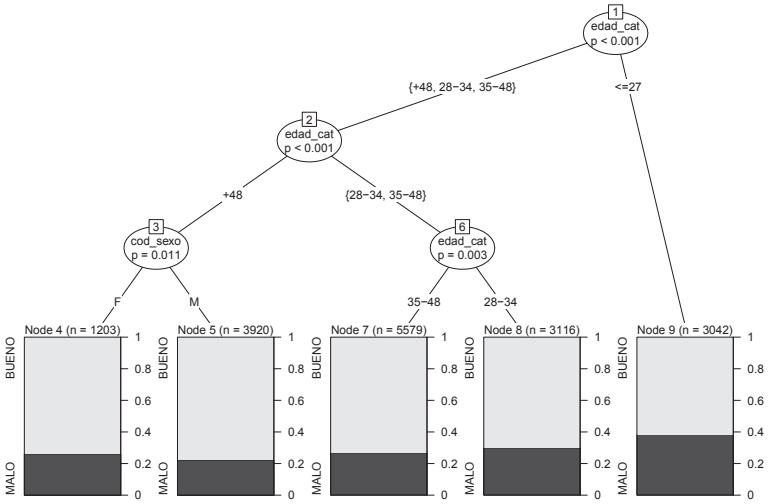
Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.24: Árbol de decisión variables: Destino y provincia inversión, Endeudamiento



Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Figura B.25: Árbol de decisión cruce variables:Edad, Género



Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

B.2 ANÁLISIS DESCRIPTIVO DE VARIABLES EXPLICATIVAS

Tabla B.1: Análisis de Frecuencias Variable Destino final de la inversión categorizada

	Frecuencia	Frecuencia_relativa	Porcentaje_Valido	Frecuencia_relativa_acumulada
Comercio de cultivos de ciclo corto	106.00	0.01	0.01	0.01
Comercio de cultivos de ciclo permanente	68.00	0.00	0.00	0.01
Comercio de ganado	429.00	0.03	0.03	0.04
Cría de ganado	10502.00	0.62	0.62	0.66
Cultivo de productos de ciclo corto	577.00	0.03	0.03	0.69
Cultivo de productos de ciclo permanente	4349.00	0.26	0.26	0.95
Movilización de cultivos de ciclo permanente	5.00	0.00	0.00	0.95
Otro destino final	824.00	0.05	0.05	1.00
Total	16860.00	1.00	1.00	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla B.2: Análisis de Frecuencias Variable Provincia de la inversión

Código	Provincia	Frecuencia	Frecuencia_relativa	Porcentaje_Valido	Frecuencia_relativa_acumulada
1	AZUAY	599.00	0.04	0.04	0.04
2	BOLIVAR	633.00	0.04	0.04	0.07
3	CAÑAR	268.00	0.02	0.02	0.09
4	CARCHI	442.00	0.03	0.03	0.12
5	COTOPAXI	1213.00	0.07	0.07	0.19
6	CHIMBORAZO	1854.00	0.11	0.11	0.30
7	EL ORO	633.00	0.04	0.04	0.33
8	ESMERALDAS	574.00	0.03	0.03	0.37
9	GUAYAS	899.00	0.05	0.05	0.42
10	IMBABURA	358.00	0.02	0.02	0.44
11	LOJA	1608.00	0.10	0.10	0.54
12	LOS RIOS	622.00	0.04	0.04	0.58
13	MORONA SANTIAGO	2420.00	0.14	0.14	0.72
15	NAPO	483.00	0.03	0.03	0.75
16	PASTAZA	280.00	0.02	0.02	0.76
17	PICHINCHA	478.00	0.03	0.03	0.79
18	TUNGURAHUA	527.00	0.03	0.03	0.82
19	ZAMORA CHINCHIPE	704.00	0.04	0.04	0.87
20	GALAPAGOS	12.00	0.00	0.00	0.87
21	SUCUMBIOS	552.00	0.03	0.03	0.90
22	ORELLANA	490.00	0.03	0.03	0.93
23	SANTA ELENA	22.00	0.00	0.00	0.93
24	SANTO DOMINGO	1189.00	0.07	0.07	1.00
Total		16860.00	1.00	1.00	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla B.3: Análisis de Frecuencias Variable Estado Civil

	Código	Frecuencia	Frecuencia_relativa	Porcentaje_Valido	Frecuencia_relativa_acumulada
CASADO	C	8234.00	0.49	0.49	0.49
DIVORCIADO	D	628.00	0.04	0.04	0.53
SOLTERO	S	7244.00	0.43	0.43	0.96
UNION LIBRE	U	318.00	0.02	0.02	0.97
VIUDO	V	436.00	0.03	0.03	1.00
TOTAL		16860.00	1.00	1.00	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla B.4: Análisis de Frecuencias Variable Nivel de Instrucción

	Código	Frecuencia	Frecuencia_relativa	Porcentaje_Valido	Frecuencia_relativa_acumulada
FORMACION INTERMEDIA (TECNICA)	T	5.00	0.00	0.00	0.00
POSTGRADO	G	3.00	0.00	0.00	0.00
PRIMARIA	P	9711.00	0.58	0.58	0.58
SECUNDARIA	S	5594.00	0.33	0.33	0.91
SIN ESTUDIOS	N	752.00	0.04	0.04	0.95
UNIVERSITARIA	U	795.00	0.05	0.05	1.00
TOTAL		16860.00	1.00	1.00	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla B.5: Análisis de Frecuencias Variable Tipo Vivienda

	Código	Frecuencia	Frecuencia_relativa	Porcentaje_Valido	Frecuencia_relativa_acumulada
ARRENDADA	A	931.00	0.06	0.06	0.06
PRESTADA	S	223.00	0.01	0.01	0.07
PROPIA HIPOTECADA	P	1756.00	0.10	0.10	0.17
PROPIA NO HIPOTECADA	N	8463.00	0.50	0.50	0.67
VIVE CON FAMILIARES	F	5487.00	0.33	0.33	1.00
TOTAL		16860.00	1.00	1.00	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla B.6: Análisis de Frecuencias Variable Numero cargas familiares

	Frecuencia	Frecuencia_relativa	Porcentaje_Valido	Frecuencia_relativa_acumulada
0	4246.00	0.25	0.25	0.25
1	3535.00	0.21	0.21	0.46
2	3751.00	0.22	0.22	0.68
3	2637.00	0.16	0.16	0.84
4	1366.00	0.08	0.08	0.92
5	711.00	0.04	0.04	0.96
6	340.00	0.02	0.02	0.98
7	142.00	0.01	0.01	0.99
8	73.00	0.00	0.00	1.00
9	28.00	0.00	0.00	1.00
10	18.00	0.00	0.00	1.00
11	6.00	0.00	0.00	1.00
12	5.00	0.00	0.00	1.00
13	1.00	0.00	0.00	1.00
16	1.00	0.00	0.00	1.00
TOTAL	16860.00	1.00	1.00	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla B.7: Análisis de Frecuencias Variable Numero cargas estudiando

	Frecuencia	Frecuencia_relativa	Porcentaje_Valido	Frecuencia_relativa_acumulada
0	6571.00	0.39	0.39	0.39
1	3873.00	0.23	0.23	0.62
2	3436.00	0.20	0.20	0.82
3	1774.00	0.11	0.11	0.93
4	769.00	0.05	0.05	0.97
5	249.00	0.01	0.01	0.99
6	127.00	0.01	0.01	1.00
7	40.00	0.00	0.00	1.00
8	17.00	0.00	0.00	1.00
9	2.00	0.00	0.00	1.00
10	1.00	0.00	0.00	1.00
12	1.00	0.00	0.00	1.00
TOTAL	16860.00	1.00	1.00	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla B.8: Análisis de Frecuencias Variable Género

	Código	Frecuencia	Frecuencia_relativa	Porcentaje_Valido	Frecuencia_relativa_acumulada
FEMENINO	F	5003.00	0.30	0.30	0.30
MASCULINO	M	11857.00	0.70	0.70	1.00
TOTAL		16860.00	1.00	1.00	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla B.9: Análisis de Frecuencias Variable Forma de Pago

	Código	Frecuencia	Frecuencia_relativa	Porcentaje_Valido	Frecuencia_relativa_acumulada
ANUAL	AN	8055.00	0.48	0.48	0.48
BIMENSUAL	BM	69.00	0.00	0.00	0.48
CUATRIMESTRAL	CM	2.00	0.00	0.00	0.48
MENSUAL	M	719.00	0.04	0.04	0.52
NUEVE MESES	NM	83.00	0.00	0.00	0.53
SEMESTRAL	SE	6313.00	0.37	0.37	0.90
TRIMESTRAL	TR	1619.00	0.10	0.10	1.00
TOTAL		16860.00	1.00	1.00	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla B.10: Análisis de Frecuencias Variable mayor plazo vencido histórico total

	Frecuencia	Frecuencia_relativa	Porcentaje_Valido	Frecuencia_relativa_acumulada
0	12170.00	0.72	0.72	0.72
1	860.00	0.05	0.05	0.77
2	1038.00	0.06	0.06	0.83
3	607.00	0.04	0.04	0.87
4	965.00	0.06	0.06	0.93
5	526.00	0.03	0.03	0.96
6	369.00	0.02	0.02	0.98
7	295.00	0.02	0.02	1.00
8	16.00	0.00	0.00	1.00
9	14.00	0.00	0.00	1.00
TOTAL	16860.00	1.00	1.00	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla B.11: Análisis de Frecuencias Variable mayor plazo vencido 6 meses total

	Frecuencia	Frecuencia_relativa	Porcentaje_Valido	Frecuencia_relativa_acumulada
0	15586.00	0.92	0.92	0.92
1	257.00	0.02	0.02	0.94
2	312.00	0.02	0.02	0.96
3	152.00	0.01	0.01	0.97
4	250.00	0.01	0.01	0.98
5	104.00	0.01	0.01	0.99
6	94.00	0.01	0.01	0.99
7	101.00	0.01	0.01	1.00
8	4.00	0.00	0.00	1.00
TOTAL	16860.00	1.00	1.00	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla B.12: Análisis de Frecuencias Variable acreedores anteriores 36 meses

	Frecuencia	Frecuencia_relativa	Porcentaje_Valido	Frecuencia_relativa_acumulada
1	8143.00	0.48	0.48	0.48
2	4943.00	0.29	0.29	0.78
3	2313.00	0.14	0.14	0.91
4	919.00	0.05	0.05	0.97
5	357.00	0.02	0.02	0.99
6	124.00	0.01	0.01	1.00
7	45.00	0.00	0.00	1.00
8	11.00	0.00	0.00	1.00
9	4.00	0.00	0.00	1.00
10	1.00	0.00	0.00	1.00
TOTAL	16860.00	1.00	1.00	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla B.13: Análisis de Frecuencias Variable Zonal

	Frecuencia	Frecuencia_relativa	Porcentaje_Valido	Frecuencia_relativa_acumulada
ZONAL CUENCA	859.00	0.05	0.05	0.05
ZONAL EL PUYO	2786.00	0.17	0.17	0.22
ZONAL GUAYAQUIL	864.00	0.05	0.05	0.27
ZONAL LOJA	2301.00	0.14	0.14	0.40
ZONAL MACHALA	637.00	0.04	0.04	0.44
ZONAL PORTOVIEJO	2418.00	0.14	0.14	0.59
ZONAL QUITO	1315.00	0.08	0.08	0.66
ZONAL RIOBAMBA	4328.00	0.26	0.26	0.92
ZONAL STO. DOMINGO	1352.00	0.08	0.08	1.00
TOTAL	16860.00	1.00	1.00	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Tabla B.14: Análisis de Frecuencias Variable Dividendos

	Frecuencia	Frecuencia_relativa	Porcentaje_Valido	Frecuencia_relativa_acumulada
1	623.00	0.04	0.04	0.04
2	1209.00	0.07	0.07	0.11
3	1182.00	0.07	0.07	0.18
4	1978.00	0.12	0.12	0.30
5	2985.00	0.18	0.18	0.47
6	1522.00	0.09	0.09	0.56
7	928.00	0.06	0.06	0.62
8	1386.00	0.08	0.08	0.70
9	24.00	0.00	0.00	0.70
10	2207.00	0.13	0.13	0.83
12	650.00	0.04	0.04	0.87
13	8.00	0.00	0.00	0.87
14	578.00	0.03	0.03	0.91
15	1.00	0.00	0.00	0.91
16	220.00	0.01	0.01	0.92
18	44.00	0.00	0.00	0.92
19	1.00	0.00	0.00	0.92
20	563.00	0.03	0.03	0.96
21	1.00	0.00	0.00	0.96
23	1.00	0.00	0.00	0.96
24	126.00	0.01	0.01	0.96
27	1.00	0.00	0.00	0.96
28	15.00	0.00	0.00	0.96
30	20.00	0.00	0.00	0.97
36	210.00	0.01	0.01	0.98
38	1.00	0.00	0.00	0.98
42	2.00	0.00	0.00	0.98
48	139.00	0.01	0.01	0.99
60	212.00	0.01	0.01	1.00
72	10.00	0.00	0.00	1.00
84	13.00	0.00	0.00	1.00
TOTAL	16860.00	1.00	1.00	1.00

Fuente: Registro Crediticio de la IMF.
Elaborado por: Autor.

Apéndice C

RESULTADOS DEL EXPERIMENTO

C.1 Simulaciones del modelo MCLP

Tabla C.1: Matriz de confusión: Prueba 1

(a) Muestra de Entrenamiento

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	11635	446	12081	96.31 %
MALO	2105	2674	4779	55.95 %
PORCENTAJE GLOBAL				84.87 %
Tiempo de ejecución: 345.64 seg.				

(b) Muestra de prueba

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	2934	120	3054	96.07 %
MALO	502	659	1161	56.76 %
PORCENTAJE GLOBAL				85.24 %

Tabla C.2: Matriz de confusión: Prueba 2

(a) Muestra de Entrenamiento

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	11618	463	12081	96.17 %
MALO	2048	2731	4779	57.15 %
PORCENTAJE GLOBAL				85.11 %
Tiempo de ejecución: 364.45 seg.				

(b) Muestra de prueba

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	2933	121	3054	96.04 %
MALO	492	669	1161	57.62 %
PORCENTAJE GLOBAL				85.46 %

Tabla C.3: Matriz de confusión: Prueba 3

(a) Muestra de Entrenamiento

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	11498	583	12081	95.17 %
MALO	1460	3319	4779	69.45 %
PORCENTAJE GLOBAL				87.88 %
Tiempo de ejecución: 117.50 seg.				

(b) Muestra de prueba

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	2908	146	3054	95.22 %
MALO	362	799	1161	68.82 %
PORCENTAJE GLOBAL				87.95 %

Tabla C.4: Matriz de confusión: Prueba 4

(a) Muestra de Entrenamiento

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	11600	481	12081	96.02 %
MALO	1988	2791	4779	58.40 %
PORCENTAJE GLOBAL				85.36 %
Tiempo de ejecución: 125.69 seg.				

(b) Muestra de prueba

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	2928	126	3054	95.87 %
MALO	477	684	1161	58.91 %
PORCENTAJE GLOBAL				85.69 %

Tabla C.5: Matriz de confusión: Prueba 5

(a) Muestra de Entrenamiento

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	11521	560	12081	95.36 %
MALO	1555	3224	4779	67.46 %
PORCENTAJE GLOBAL				87.46 %
Tiempo de ejecución: 170.5 seg.				

(b) Muestra de prueba

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	2913	141	3054	95.38 %
MALO	384	777	1161	66.93 %
PORCENTAJE GLOBAL				87.54 %

C.2 Simulaciones del modelo MC2LP

Tabla C.6: Matriz de confusión: $\lambda_1 = 0{,}10$, $\lambda_2 = 0{,}90$, $b \in [-3200, 3200]$

(a) Muestra de Entrenamiento

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	11472	609	12081	94.96 %
MALO	1320	3459	4779	72.38 %
PORCENTAJE GLOBAL				88.56 %
Tiempo de ejecución: 503.20 seg.				

(b) Muestra de prueba

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	2908	146	3054	95.22 %
MALO	327	834	1161	71.83 %
PORCENTAJE GLOBAL				88.78 %

Tabla C.7: Matriz de confusión: $\lambda_1 = 0{,}20$, $\lambda_2 = 0{,}80$, $b \in [-3200, 3200]$

(a) Muestra de Entrenamiento

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	11470	611	12081	94.94 %
MALO	1342	3437	4779	71.92 %
PORCENTAJE GLOBAL				88.42 %
Tiempo de ejecución: 528.74 seg.				

(b) Muestra de prueba

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	2905	149	3054	95.12 %
MALO	334	827	1161	71.23 %
PORCENTAJE GLOBAL				88.54 %

Tabla C.8: Matriz de confusión: $\lambda_1 = 0{,}30$, $\lambda_2 = 0{,}70$, $b \in [-3200, 3200]$

(a) Muestra de Entrenamiento

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	11497	584	12081	95.17 %
MALO	1448	3331	4779	69.70 %
PORCENTAJE GLOBAL				87.95 %
Tiempo de ejecución: 416.4 seg.				

(b) Muestra de prueba

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	2909	145	3054	95.25 %
MALO	352	809	1161	69.68 %
PORCENTAJE GLOBAL				88.21 %

Tabla C.9: Matriz de confusión: $\lambda_1 = 0{,}40$, $\lambda_2 = 0{,}60$, $b \in [-3200, 3200]$

(a) Muestra de Entrenamiento

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	11555	526	12081	95.65 %
MALO	1760	3019	4779	63.17 %
PORCENTAJE GLOBAL				86.44 %
Tiempo de ejecución: 417.47 seg.				

(b) Muestra de prueba

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	2922	132	3054	95.68 %
MALO	431	730	1161	62.88 %
PORCENTAJE GLOBAL				86.64 %

Tabla C.10: Matriz de confusión: $\lambda_1 = 0{,}50$, $\lambda_2 = 0{,}50$, $b \in [-3200, 3200]$

(a) Muestra de Entrenamiento

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	11809	272	12081	97.75 %
MALO	2784	1995	4779	41.75 %
PORCENTAJE GLOBAL				81.87 %
Tiempo de ejecución: 269.69 seg.				

(b) Muestra de prueba

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	2977	77	3054	97.48 %
MALO	681	480	1161	41.34 %
PORCENTAJE GLOBAL				82.02 %

Tabla C.11: Matriz de confusión: $\lambda_1 = 0{,}60$, $\lambda_2 = 0{,}40$, $b \in [-3200, 3200]$

(a) Muestra de Entrenamiento

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	11716	365	12081	96.98 %
MALO	2389	2390	4779	50.01 %
PORCENTAJE GLOBAL				83.67 %
Tiempo de ejecución: 339.54 seg.				

(b) Muestra de prueba

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	2951	103	3054	96.63 %
MALO	580	581	1161	50.04 %
PORCENTAJE GLOBAL				83.80 %

Tabla C.12: Matriz de confusión: $\lambda_1 = 0{,}65$, $\lambda_2 = 0{,}35$, $b \in [-3200, 3200]$

(a) Muestra de Entrenamiento

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	11612	469	12081	96.12 %
MALO	1991	2788	4779	58.34 %
PORCENTAJE GLOBAL				85.41 %
Tiempo de ejecución: 330.05 seg.				

(b) Muestra de prueba

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	2930	124	3054	95.94 %
MALO	480	681	1161	58.66 %
PORCENTAJE GLOBAL				85.67 %

Tabla C.13: Matriz de confusión: $\lambda_1 = 0{,}70$, $\lambda_2 = 0{,}30$, $b \in [-3200, 3200]$

(a) Muestra de Entrenamiento

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	11563	518	12081	95.71 %
MALO	1800	2979	4779	62.34 %
PORCENTAJE GLOBAL				86.25 %
Tiempo de ejecución: 305.41 seg.				

(b) Muestra de prueba

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	2920	134	3054	95.61 %
MALO	435	726	1161	62.53 %
PORCENTAJE GLOBAL				86.50 %

Tabla C.14: Matriz de confusión: $\lambda_1 = 0{,}75$, $\lambda_2 = 0{,}25$, $b \in [-3200, 3200]$

(a) Muestra de Entrenamiento

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	11522	559	12081	95.37 %
MALO	1641	3138	4779	65.66 %
PORCENTAJE GLOBAL				86.95 %
Tiempo de ejecución: 281.80 seg.				

(b) Muestra de prueba

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	2916	138	3054	95.48 %
MALO	406	755	1161	65.03 %
PORCENTAJE GLOBAL				87.09 %

Tabla C.15: Matriz de confusión: $\lambda_1 = 0{,}80$, $\lambda_2 = 0{,}20$, $b \in [-3200, 3200]$

(a) Muestra de Entrenamiento

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	11469	612	12081	94.93 %
MALO	1418	3361	4779	70.33 %
PORCENTAJE GLOBAL				87.96 %
Tiempo de ejecución: 313.39 seg.				

(b) Muestra de prueba

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	2899	155	3054	94.92 %
MALO	355	806	1161	69.42 %
PORCENTAJE GLOBAL				87.90 %

Tabla C.16: Matriz de confusión: $\lambda_1 = 0{,}90$, $\lambda_2 = 0{,}10$, $b \in [-3200, 3200]$

(a) Muestra de Entrenamiento

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	10900	1181	12081	90.22 %
MALO	1111	3668	4779	76.75 %
PORCENTAJE GLOBAL				86.41 %
Tiempo de ejecución: 183.71 seg.				

(b) Muestra de prueba

	PRONOSTICADO			
OBSERVADO	BUENO	MALO	TOTAL	PORCENTAJE
BUENO	2763	291	3054	90.47 %
MALO	272	889	1161	76.57 %
PORCENTAJE GLOBAL				86.64 %