
UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA

# Análisis y formulación matemática del modelo de máquina de soporte vectorial

**TESIS**

Para obtener el título de:

**LICENCIADO EN MATEMÁTICAS APLICADAS**

PRESENTA:

**Diego Josue Bautista Reyes**

Director de Tesis

*Dr. Tomás Pérez Becerra*

Huajuapan de León, Oaxaca, Febrero de 2026

*A las personas que más admiro en todo el mundo,*
*Mi padre Santiago Bautista y mi madre, Irma Reyes . . .*

*A mis compañeros de viaje en esta aventura llamada vida,*
*Ariadna Bautita y Santiago Bautista*

# Agradecimientos

A mi padre Santiago Bautista Rodríguez, quien me ha impulsado a ser mejor persona, por todas sus palabras de aliento, sus consejos y demostrarme que siempre se puede alcanzar el éxito con el suficiente esfuerzo. Por toda la confianza que ha depositado en mí y por siempre ayudarme a levantarme cuando no podía por mi propia cuenta.

A mi madre Irma Margarita Reyes Mendoza le agradezco por ser mi maestra no solo de escuela sino también de vida, me mostró a ver el mundo de manera seria y firme, a poder encarar problemas. Por forjar en mí hábitos de disciplina y responsabilidad que me han ayudado a alcanzar mis metas en la vida. Por siempre darme un abrazo y un beso los cuales se sentían hasta dentro del alma. Agradezco a ambos por su amor infinito y sus caricias, por sus enseñanzas, sus consejos, regaños y también por sus palabras, esos momentos son el tesoro más grande que tengo.

A mis hermanos Ariadna Bautista Reyes y Santiago Bautista Reyes, quienes son mi mayor inspiración y uno de mis más grandes motivos para seguir adelante. Les agradezco por todo su apoyo, las risas que compartimos, los corajes que pasamos entre muchas otras cosas más, son el brillo de mis ojos y la luz de mi alma.

A mi flor de Noviembre quien fue una parte fundamental en esta etapa de mi vida, me llenó de tantos hermosos momentos que sería imposible describirlos a detalle, fue un apoyo fundamental que tuve por lo que le estaré eternamente agradecido.

A Andrea Jazive Martínez José y Jacqueline Barajas Gamas quienes me apoyaron en todo momento, de aliento pero sobre todo, me brindaron su amistad y su confianza, al igual que su paciencia.

Al Dr. Tomás Pérez Becerra le agradezco la confianza y la paciencia depositada en mí para poder realizar este trabajo, también por enseñarme lo necesario para realizar investigación y pensar como matemático. Sus palabras fueron fundamentales para mi crecimiento tanto profesional como personal.

A mis sinodales el Dr. Jesús Ferando Tenorio Arvide, el Dr. Sergio Palafox Delgado y al Dr. Pedro Alberto Antonio Soto por la dedicación prestada para leer este trabajo y por las observaciones realizadas.

A mis profesores por las enseñanzas las cuales me nutrieron cada día para llegar a ser el profesionista que soy.

A todos los que han contribuido de alguna forma en mi formación académica, amigos colegas y profesores que me acompañaron a lo largo de este lustro de aprendizaje. Gracias a todos.

# Índice general

# Introducción

En la actualidad, los modelos de aprendizaje automático se enfrentan al manejo de grandes volúmenes de datos y en la resolución de problemas de clasificación o regresión complejos; estos modelos han demostrado ser eficaces para este tipo de tareas. Sin embargo, la naturaleza exponencialmente creciente de los datos, especialmente en aplicaciones de alta dimensionalidad, impone limitaciones a los modelos de aprendizaje en términos de tiempo de procesamiento y eficiencia.

El avance de la computación ha abierto nuevas oportunidades en el ámbito del aprendizaje automático, permitiendo el desarrollo de modelos que prometen mejorar significativamente el procesamiento de grandes volúmenes de datos y la resolución de problemas complejos. Uno de estos modelos es el denominado máquina de soporte vectorial (SVM, por sus siglas en inglés), que utiliza las propiedades del álgebra lineal, de la optimización y del análisis funcional para manejar datos en espacios de características de dimensiones muy elevadas. Estas propiedades sugieren que la SVM podría realizar operaciones de clasificación o regresión de manera eficiente. Sin embargo, a pesar de su potencial, en la revisión bibliográfica realizada se observó que la SVM enfrenta una limitación importante: la falta de un documento que muestre la formalización matemática que sustente su teoría con una notación clara. Estos fundamentos se basan en la teoría matemática de optimización convexa, pero carecen de una formulación rigurosa que permita comprender y predecir su comportamiento en escenarios prácticos. Esta carencia genera incertidumbre en cuanto a sus propiedades esenciales, como la estabilidad, la precisión y la capacidad de generalización.

En este contexto, surge la necesidad de recopilar una base teórica para la SVM que permita aprovechar su potencial y establecer un marco de referencia que guíe su aplicación en el aprendizaje automático. Este trabajo de tesis formaliza matemáticamente la SVM, proponiendo un

marco teórico matemático que permita comprender sus características. La formalización de la SVM no solo contribuirá al avance del aprendizaje automático, sino que también facilitará el diseño de algoritmos y hardware específicos para maximizar sus ventajas en la clasificación y regresión de datos complejos.

Esta investigación proporciona los fundamentos matemáticos para la SVM y, posteriormente, evalúa aspectos críticos como la estabilidad, la precisión y la eficiencia de la SVM. Además, con la presencia de una base matemática clara, se derrumba una barrera para el desarrollo de hardware y algoritmos específicos que maximicen el potencial de este modelo.

Por lo anterior, el objetivo general del trabajo es realizar una memoria autocontenida, en la medida de lo posible, que contenga una formalización matemática rigurosa para el modelo de máquina de soporte vectorial (SVM) que permita comprender sus propiedades, limitaciones y comportamiento en aplicaciones de clasificación y regresión complejas, con el fin de mejorar su fiabilidad, eficiencia y aplicabilidad en el ámbito del aprendizaje automático. Para alcanzarlo, se plantean los objetivos específicos:

1. Identificar y analizar la teoría que da origen y fundamento a la formulación matemática del modelo de máquina de soporte vectorial (SVM).

2. Definir un marco teórico matemático que permita describir formalmente las propiedades fundamentales de la SVM, tales como su estabilidad, convergencia y capacidad de generalización en problemas de clasificación y regresión.

3. Desarrollar y validar teoremas y lemas que expliquen el comportamiento de la SVM en términos de optimización.

La tesis se divide de la siguiente forma:

**Capítulo 1:** La matemática que sustenta a la SVM incluye conceptos de espacios vectoriales pre-hilbertianos, espacios normados, transformaciones lineales, funcionales y operaciones matriciales. Así como tópicos particulares en optimización convexa, tales como conjuntos convexos, combinaciones convexas y envolturas convexas, esto dará paso al planteamiento del problema de optimización con restricciones convexas y afines. Una parte

relevante es el planteamiento de problemas cuadráticos convexos y la existencia de hiperplanos de separación bajo funcionales continuos. Todo esto será exhibido en este capítulo de fundamentos. Adicionalmente, y tomando un papel principal, se mostrará el "Teorema de dualidad fuerte y recuperación primal desde el dual", que establece las condiciones bajo las cuales la solución del problema primal puede ser recuperada a partir del dual y que, en la revisión bibliográfica, comúnmente son omitidas tales condiciones. Con esto se logra el primer objetivo específico.

**Capítulo 2:** Denominado simplemente "Modelo de Máquina de Soporte vectorial". Se inicia con una breve introducción al aprendizaje automático, resaltando las características del aprendizaje supervisado, donde se enmarca el tema de esta investigación. Es en este capítulo donde se formula el modelo SVM de una manera detallada; en específico, se inicia con los conjuntos linealmente separables y se establece la existencia de un hiperplano de separación como consecuencia de que las envolventes convexas de los conjuntos linealmente separables sean disjuntas. En la literatura, esta existencia en ocasiones se obvia.

Un tema interesante, y que no se puede pasar por alto, es el "Teorema de separabilidad mediante aumento de dimensionalidad", que da fundamento a la técnica de introducir un kernel en el modelo de SVM; esto ocasiona que, en dimensiones altas, ¡los conjuntos no linealmente separables se vuelven separables! Esto permite darle el poder a las SVM's para procesar datos en dimensiones altas, o incluso elevarlas más. Claro está que en la mayoría de las referencias de este tema no es incluido, dado que las condiciones son generales y el kernel gaussiano las cumple por de facto.

Finaliza este capítulo con la descripción detallada de la SVM para regresión, junto con el caso particular de las máquinas de soporte vectorial basadas en mínimos cuadrados (LS-SVMR), cuyo entrenamiento se encuentra enfocado en la resolución de un sistema de ecuaciones mediante la matriz inversa.

Con lo anterior, se logran los objetivos específicos dos y tres.

A manera de justificación para la elaboración de este trabajo de tesis, el hecho de formali-

zar matemáticamente la SVM no solo contribuye a consolidar su potencial, sino que también permite un avance significativo en el campo de la computación aplicada. Una estructura teórica clara facilita la identificación de aplicaciones óptimas para la SVM, así como la creación de algoritmos y arquitecturas diseñadas específicamente para su implementación. Por tanto, este estudio brindará un impulso al potencial uso y aplicaciones del modelo en cuestión, los cuales serán herramientas confiables que puedan responder a la creciente demanda de procesamiento de datos en problemas complejos de clasificación y regresión.

CAPÍTULO 1

# Fundamentos matemáticos

Un objetivo específico de esta investigación es mostrar, en un primer paso, las bases matemáticas que brindan soporte al desarrollo de modelos de SVM, en especial, la teoría de espacios vectoriales y normados. Se hablará de espacios dotados de un producto interior, que se denominan prehilbertianos, así como de funcionales, transformaciones lineales y algunos preliminares en optimización convexa, haciendo énfasis en las envolventes convexas debido a su relación con los conjuntos linealmente separables. Posteriormente, se plantearán los problemas de optimización convexa y los teoremas de recuperación de la solución del problema primal a partir del dual. Por lo que este capítulo está dedicado a exhibir las bases teóricas sin demostración; sin embargo, se encuentra debidamente referida, salvo algunos resultados especiales que, al no hallarse referencia alguna, se presentan con sus pruebas.

## 1.1. Notaciones y convenciones

Esta sección inicia con un conjunto de definiciones a utilizar en esta tesis; algunas están acompañadas de observaciones generales con el objetivo de contextualizar. Posteriormente, se introducen algunos de los conceptos que forman parte de los fundamentos matemáticos que brindan los cimientos de los modelos SVM. Se inicia con la definición de campo y de espacio vectorial con producto interior.

**Definición 1.1.1. (Campo).** Un campo (o cuerpo) $F$ consiste en un conjunto en el que están definidas dos operaciones (llamadas adición y multiplicación), tal que para cualquier par de elementos $x$ y $y$ en $F$ existen dos únicos elementos $x + y$ en $F$ y $xy$ en $F$, de tal manera que se cumplan las siguientes condiciones para cada $x,\ y$ y $z \in F$.

1. $x + y$ es elemento de $F$ (Ley de la cerradura para la suma).
2. $x + y = y + x$ (Ley conmutativa para la suma).
3. $x + (y + z) = (x + y) + z$ (Ley asociativa para la suma).
4. Existe un único elemento $0 \in F$ tal que $x + 0 = x$ (Existencia del neutro aditivo).
5. Para cada $x$ existe un $y$ tal que $x + y = 0$ (Existencia del inverso aditivo).
6. $xy$ en $F$ (Ley de la cerradura para el producto).
7. $xy = yx$ (Ley conmutativa para el producto).
8. $x(yz) = (xy)z$ (Ley asociativa para el producto).
9. Existe $1 \in F \setminus \{0\}$ tal que para cada $x$ en $F$ se tiene que $1x = x$ (Existencia del neutro multiplicativo).
10. Para cada $x \in F \setminus \{0\}$, existe $y$ tal que $xy = 1$ (Existencia del inverso multiplicativo).
11. $x(y + z) = xy + xz$ (Ley distributiva).

Un ejemplo de campo es el conjunto de los números reales $\mathbb{R}$ y los números complejos $\mathbb{C}$, bajo las operaciones de suma y producto usuales. En el capítulo 2 se considerará el campo como uno de estos espacios; sin embargo, la definición introducida a continuación se plantea en términos generales.

**Definición 1.1.2. (Espacio vectorial).** Un espacio vectorial (o espacio lineal) $V$ sobre un campo $F$ consiste en un conjunto en el que están definidas dos operaciones (llamadas adición y multiplicación por escalares, respectivamente), tal que para cualquier par de elementos $x$ y $y$ en $V$ existe un único elemento $x + y$ en $V$, y para cada elemento $a$ en $F$ y cada elemento $x$ en $V$, exista un elemento único $ax$ en $V$, de manera que se cumplan las siguientes condiciones:

1. Para todo par $x, y$ en $V$, $x + y = y + x$ (Conmutatividad de la adición).
2. Para toda tripleta $x, y, z$ en $V$, $(x + y) + z = x + (y + z)$ (Asociatividad para la suma).
3. Existe un único elemento $0 \in V$ tal que $x + 0 = x$, para toda $x$ en $V$.
4. Para cada elemento $x$ en $V$, existe un elemento $y$ en $V$ tal que $x + y = 0$.
5. Para cada elemento $x$ en $V$, $1x = x$.
6. Para cada par $a, b$ de elementos en $F$ y cada elemento $x$ en $V$, $(ab)x = a(bx)$.
7. Para cada par de elementos $a, b$ en $F$ y cada elemento $x$ en $V$, $(a + b)x = ax + bx$.
8. Para cada elemento $a$ en $F$ y para cada par de elementos $x, y$ en $V$, $a(x+y) = ax+ay$.

Los elementos $x + y$ y $ax$ se denominan, respectivamente, la suma de $x$ y $y$ y el producto de $a$ y $x$.

Una operación dentro de algunos espacios vectoriales es el producto interior; además, se utiliza para formular la SVM. Formalmente, esta operación está definida como una función cuyo dominio es el espacio producto $V \times V$, tal como se muestra en la siguiente definición.

**Definición 1.1.3. (Producto interior).** Sea $V$ un espacio vectorial definido sobre un campo $F = \mathbb{R}$ o $\mathbb{C}$. Un producto interior en $V$ es una función $\langle \cdot, \cdot \rangle : V \times V \to F$ que asigna a cada par ordenado de vectores $x$ y $y$ en $V$ un escalar en $F$, representado como $\langle x, y \rangle$, tal que para toda $x, y$ y $z$ en $V$ y toda $c$ en $F$ se tiene que:

1. $\langle x, x \rangle > 0$ si $x \neq 0$.
2. $\langle x, y \rangle = \overline{\langle y, x \rangle}$, donde la barra indica conjugación compleja.
3. $\langle cx, y \rangle = c\langle x, y \rangle$.
4. $\langle x + z, y \rangle = \langle x, y \rangle + \langle z, y \rangle$.
5. $\langle x, x \rangle = 0 \Rightarrow x = 0$.

En caso de que el espacio vectorial esté definido sobre el campo real, se denomina espacio vectorial real; de igual manera, se define el espacio vectorial complejo. Para espacios vectoriales reales, la propiedad 2 se escribe como $\langle x, y \rangle = \langle y, x \rangle$.

**Ejemplo 1.1.4.** Sea el espacio vectorial $V = \mathbb{R}^n$. Para $x = (a_1, a_2, \ldots, a_n)$ y $y = (b_1, b_2, \ldots, b_n) \in \mathbb{R}^n$ defínase

$$\langle x, y \rangle = \sum_{i=1}^{n} a_i b_i. \tag{1.1.1}$$

Esta función satisface las condiciones de la Definición 1.1.3 y se denomina *producto interior ordinario en* $\mathbb{R}^n$ o también producto punto.

Aquellos espacios vectoriales que cuentan con un producto interior reciben un nombre característico; formalmente, se introduce en la definición a continuación.

**Definición 1.1.5.** Un espacio prehilbertiano es un espacio vectorial $V$ real o complejo dotado de un producto interior.

A continuación, se define la norma sobre ciertos espacios vectoriales, la cual es una función que tiene relevancia en las SVM debido a que determina, de cierta manera, la distancia entre los puntos del espacio.

**Definición 1.1.6. (Norma).** Sea $V$ un espacio vectorial real o complejo. Una función $\|\cdot\| : V \to \mathbb{R}$ es llamada norma si cumple lo siguiente para cada $x, y \in V$ y $\alpha$ en el campo:

1. $\|x\| \geq 0$.
2. $\|x\| = 0$ si y solo si $x = 0$.
3. $\|\alpha x\| = |\alpha|\|x\|$.
4. $\|x + y\| \leq \|x\| + \|y\|$.

Donde $|\cdot|$ denota el valor absoluto si el espacio vectorial es real o el módulo para el caso de espacios vectoriales complejos.

**Ejemplo 1.1.7.** Sea el espacio vectorial $V = \mathbb{R}^n$. La función $\|\cdot\|_2 : \mathbb{R}^n \to \mathbb{R}$ definida como

$$\|x\|_2 = \left( \sum_{i=1}^{n} x_i^2 \right)^{1/2}, \tag{1.1.2}$$

para cada $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, es una norma en $\mathbb{R}^n$ y es llamada norma euclidiana.

**Definición 1.1.8.** Un espacio vectorial real o complejo dotado de una norma se denomina espacio normado.

Dentro de los espacios prehilbertianos, el producto interior induce una norma; para demostrarlo, se introduce una función candidata a tal norma.

**Definición 1.1.9.** Sea $V$ un espacio real prehilbertiano. Se define la función $\phi : V \to \mathbb{R}$ como $\phi(x) = \sqrt{\langle x, x \rangle}$.

La siguiente desigualdad se denomina la *desigualdad de Cauchy-Schwarz*. Muestra la relación entre el producto interno y el producto de la función $\phi$ consigo misma.

**Teorema 1.1.10. (Desigualdad de Cauchy-Schwarz)** [20, Teorema 7.1, p. 151].
Si $V$ es un espacio real prehilbertiano, entonces

$$|\langle x, y\rangle| \leq \phi(x)\phi(y). \tag{1.1.3}$$

**Proposición 1.1.11.** Si $V$ es un espacio vectorial real prehilbertiano, entonces $\phi$ es una norma sobre $V$. En consecuencia, $V$ es un espacio normado.

*Demostración.* Sean $x, y \in V$ y $\alpha$ un escalar.

1. Por definición, $\phi(x) = \sqrt{\langle x, x\rangle}$. Dado que $\langle x, x\rangle \geq 0$, se tiene que $\phi(x) \geq 0$.

2. Se tiene que $\phi(x) = 0$, por lo que $\langle x, x\rangle = 0$, es decir $x = 0$.

3. $\phi(\alpha x) = \sqrt{\langle \alpha x, \alpha x\rangle} = \sqrt{\alpha^2\langle x, x\rangle} = |\alpha|\sqrt{\langle x, x\rangle} = |\alpha|\phi(x)$.

4. Tomando la norma al cuadrado se tiene

$$\begin{aligned}\phi(x+y)^2 &= \langle x+y, x+y\rangle \\ &= \langle x, x\rangle + 2\langle x, y\rangle + \langle y, y\rangle \\ &\leq \phi(x)^2 + 2\phi(x)\phi(y) + \phi(y)^2 \\ &= (\phi(x) + \phi(y))^2,\end{aligned}$$

   por lo que:

$$\phi(x+y)^2 \leq (\phi(x) + \phi(y))^2$$

   , es decir

$$\phi(x+y) \leq \phi(x) + \phi(y).$$

Tras haber demostrado las propiedades requeridas, se concluye que $\phi$ es una norma sobre el espacio $V$. □

Por lo tanto, de ahora en adelante, se utilizará la notación $\|\cdot\|$ para hacer referencia a la función $\phi$ sobre un espacio prehilbert. A continuación, se introducen las definiciones de conjuntos abiertos y cerrados, las cuales son relevantes para el desarrollo de conceptos matemáticos más complejos en las siguientes páginas. Las definiciones están dadas dentro del contexto de los espacios normados debido a la idoneidad con la temática de la investigación. A manera de identificación, se utilizará el símbolo $V$ para los espacios vectoriales y $X$ para normados con su respectiva norma $\|\cdot\|$.

**Definición 1.1.12. (Punto y conjunto interior).** Sea $P$ un subconjunto de un espacio normado $X$. El punto $p \in P$ es llamado *punto interior* de $P$ si existe un $\varepsilon > 0$ tal que todos los vectores $x$ que satisfagan $\|x - p\| < \varepsilon$ sean elementos de $P$. La colección de todos los puntos interiores de $P$ se denomina *interior* de $P$ y es denotado por $\mathring{P}$.

Los conjuntos en los que todos sus elementos son puntos interiores reciben un nombre en particular, el cual se muestra en la siguiente definición.

**Definición 1.1.13. (Conjunto abierto).** Sea $P$ un subconjunto de un espacio normado $X$. El conjunto $P$ es llamado *abierto* si $P = \mathring{P}$.

Otro tipo de puntos específicos son los puntos de clausura, que permiten definir la clausura de un conjunto y caracterizar cuándo un conjunto es cerrado. Sus definiciones se presentan a continuación

**Definición 1.1.14. (Punto clausura y conjunto clausura).** Sea $P$ un subconjunto de un espacio normado $X$. Un punto $x \in X$ es *punto clausura* del conjunto $P$ si dado $\varepsilon > 0$ existe un punto $p \in P$ tal que $\|x - p\| < \varepsilon$. La colección de todos los puntos clausura de $P$ se llama la *clausura* de $P$ y es denotada por $\overline{P}$.

Los conjuntos cuyos elementos son todos puntos clausura reciben la denominación que se establece en la siguiente definición.

**Definición 1.1.15. (Conjunto cerrado.)** Sea $P$ un subconjunto de un espacio normado $X$. El conjunto $P$ es llamado *cerrado* si $P = \overline{P}$.

Ahora bien, ciertas funciones desempeñan un papel fundamental en el entrenamiento de las SVM dentro del marco de los espacios vectoriales: las transformaciones lineales y los funcionales lineales. Las definiciones correspondientes se presentan a continuación.

**Definición 1.1.16. (Transformación).** Sean $V$ y $W$ espacios vectoriales sobre un campo $F$ y sea $D$ un subconjunto de $V$. Una regla $T : D \subseteq V \to W$, que asocia a cada elemento $v$ en $D$ un único elemento $w$ en $W$, se dice que es una transformación de $V$ a $W$ con dominio $D$. Si $w$ corresponde a $x$ bajo $T$, se escribe $w = T(x)$.

**Definición 1.1.17. (Transformación lineal).** Sean $V$ y $W$ espacios vectoriales sobre un campo $F$. Una transformación $T : V \to W$ se llama transformación lineal de $V$ en $W$ si para toda $x, y \in V$ y $c \in F$ se tiene que:

1. $T(x + y) = T(x) + T(y)$.

2. $T(cx) = cT(x)$.

**Definición 1.1.18. (Funcional).** Sea $V$ un espacio vectorial definido sobre un campo $F$. Una transformación $f : V \to F$ es llamada un funcional sobre $V$.

En los problemas de optimización convexa (introducidos más adelante en el Capítulo 2), una de las condiciones suficientes para la existencia de la solución es que las restricciones sean transformaciones afines. Este concepto se define de la siguiente manera:

**Definición 1.1.19. (Transformación afín).** Sean $X$ y $Y$ dos espacios vectoriales. Una transformación $f : X \to Y$ se dice que es afín si existen una aplicación lineal $L : X \to Y$ y un vector fijo $b$ en el espacio $Y$ tales que

$$f(x) = L(x) + b.$$

Con el fin de mantener esta tesis lo más autocontenida posible, se revisan a continuación tres tipos especiales de matrices.

**Definición 1.1.20. (Matriz transpuesta).** Sea $A$ una matriz de $m \times n$ con elementos en un campo $F$. Se define la transpuesta de $A$ como la matriz $A^T$ de $n \times m$ tal que $A^T_{ij} = A_{ji}$ para $i \in \{1, 2, , ..., n\}$ y $j \in \{1, 2, ..., m\}$.

**Definición 1.1.21. (Matriz conjugada).** Sea $A$ una matriz de dimensión $m \times n$ con entradas reales o complejas. Se define la transpuesta conjugada (o adjunta) de $A$ como la matriz $A^*$ de $n \times m$ tal que $A^*_{ij} = \bar{A}_{ji}$.

**Definición 1.1.22. (Matriz de Gram).** Sea $S = \{v_1, v_2, ..., v_n\}$ un conjunto de vectores de un espacio prehilbertiano $V$. Sea $G$ una matriz de $n \times n$ cuyas entradas están dadas por $G_{ij} = \langle v_i, v_j \rangle$, donde $i, j \in \{1, 2, ..., n\}$. La matriz $G$ se denomina matriz de Gram.

**Definición 1.1.23. (Matriz semidefinida positiva)** Una matriz cuadrada $M$ de tamaño $n \times n$ con coeficientes reales es semidefinida positiva si cumple las dos condiciones:

- $M = M^T$,
- $x^T M x \geq 0$, para todo $x \in \mathbb{R}^n$ distinto del vector nulo.

## 1.2. Principios de optimización convexa

Las SVM son modelos de optimización con restricciones convexas y afines. Para deducir este proceso, dentro de esta sección se contextualiza esta teoría; se incluyen definiciones y teoremas que permiten comprender de mejor manera el modelo cuadrático con restricciones convexas, objetivo de este trabajo de tesis.

### 1.2.1. Conceptos básicos

**Definición 1.2.1. (Conjunto convexo).** Sea $V$ un espacio vectorial real. Un conjunto $C \subseteq V$ es convexo si para cualesquiera dos puntos, $x_1,\ x_2$ de $C$ y cualquier $\theta$, tal que $0 \leq \theta \leq 1$, se tiene que

$$\theta x_1 + (1-\theta)x_2 \tag{1.2.1}$$

pertenece a $C$.

La definición anterior se puede interpretar como: dados cualesquiera dos puntos en el conjunto, el subconjunto de puntos que forman el segmento que los une debe estar totalmente contenido en el conjunto $C$.

Un caso que se considera en la formulación del modelo de máquina de soporte vectorial es cuando se entrena con conjuntos discretos finitos, esto es, con conjuntos que no son convexos; por ello, en esta sección se introducen algunos conceptos que permiten realizar el ajuste de la SVM con conjuntos de cardinalidad finita.

**Definición 1.2.2. (Combinación convexa).** Sean $V$ un espacio vectorial real, un conjunto finito $P = \{v_1, v_2, \ldots, v_n\} \subseteq V$ y $x \in V$. Se dice que $x$ es una combinación convexa de los elementos de $P$ si

$$x = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n,$$

donde $\alpha_i \in \mathbb{R}$ y $\alpha_i \geq 0$ para cada $i \in \{1, \ldots, n\}$, además $\sum_{i=1}^{n} \alpha_i = 1$.

**Definición 1.2.3. (Envolvente convexa).** Sean $V$ un espacio vectorial real y $S = \{x_1, x_2, \ldots, x_n\} \subseteq V$ un conjunto finito. La envolvente convexa de $S$ se denota por $\text{conv}(S)$ y es el conjunto

$$\text{conv}(S) = \left\{ \sum_{i=1}^{k} \alpha_i x_i \;\middle|\; x_i \in S,\ \alpha_i \geq 0,\ \sum_{i=1}^{n} \alpha_i = 1 \right\}.$$

El conjunto $\text{conv}(S)$ está formado por todas las combinaciones convexas de los elementos

de $S$. De este modo, se obtiene un conjunto convexo que contiene al conjunto finito $S$, hecho que se establece en la siguiente proposición.

**Proposición 1.2.4.** Sea $V$ un espacio vectorial real y $S = \{x_1, x_2, \ldots, x_n\} \subseteq V$ un conjunto finito. Se cumple lo siguiente:

1. $S \subseteq \text{conv}(S)$.

2. $\text{conv}(S)$ es un conjunto convexo.

*Demostración.* 1. Sean $x, y \in S$ dos elementos distintos arbitrarios. Sin pérdida de generalidad, supóngase que $x = x_1$, por lo que se puede escribir el elemento $x$ de la siguiente forma:

$$x = 1x + 0x_2 + 0x_3 + ... + 0x_n,$$

el cual está en $\text{conv}(S)$.

2. Sean $x, y \in \text{conv}(S)$ y $\theta$ una constante tal que $0 \leq \theta \leq 1$. Se tiene que

$$x = \sum_{i=1}^{k_x} \alpha_i^x x_i \qquad \text{y} \qquad y = \sum_{j=1}^{k_y} \alpha_j^y x_j;$$

así,

$$\theta x + (1-\theta)y = \theta \sum_{i=1}^{k_x} \alpha_i^x x_i + (1-\theta) \sum_{j=1}^{k_y} \alpha_j^y x_j.$$

Nótese que los coeficientes $\alpha_i^x$, $\alpha_j^y$, $\theta$ y $(1-\theta)$ son positivos, además, por la definición de una envolvente convexa, se tiene que:

$$\sum_{i=1}^{k_x} \alpha_i^x = 1 = \sum_{j=1}^{k_y} \alpha_j^y,$$

por lo que se concluye que

$$\theta \sum_{i=1}^{k_x} \alpha_i^x + (1-\theta) \sum_{j=1}^{k_y} \alpha_j^y = \theta(1) + (1-\theta)(1) = 1.$$

Por lo tanto, el conjunto conv($S$) es convexo.

□

Un resultado que se deduce de manera directa es la igualdad de un conjunto convexo con su envoltura convexa; esto es:

**Proposición 1.2.5.** [23, Teorema 2.2, p. 11] Sea $V$ un espacio vectorial real y $P \subseteq V$. Si $P$ es convexo, entonces $P = \text{conv}(P)$.

Las Máquinas de Soporte Vectorial se fundamentan directamente en los principios de la teoría de optimización, la cual proporciona el marco matemático necesario para formular y resolver el problema central. Esta tarea se traduce en un problema de optimización convexa, en el que se busca minimizar una función objetivo sujeta a un conjunto de restricciones que aseguran la correcta clasificación de los datos. De esta manera, la formulación general del problema de optimización sirve como punto de partida para derivar las SVM lineales. En la definición siguiente se plantea este problema.

**Definición 1.2.6. (Problema de Optimización).** Un problema matemático de optimización (también llamado problema de optimización) tiene la forma

$$\begin{aligned} &\text{Minimizar} \quad f(x) && && (1.2.2)\\ &\text{sujeto a } f_i(x) \leq b_i, && i = 1, \ldots, m,\\ &\qquad h_j(x) = c_j, && j = 1, \ldots, p \end{aligned}$$

donde:

- El vector $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ es la variable de optimización del problema.

- La función $f : \mathbb{R}^n \to \mathbb{R}$ se denomina función objetivo.
- Las funciones $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$ y $h_j : \mathbb{R}^n \to \mathbb{R}$, $j = 1, \ldots, p$ son llamadas funciones de restricción (restricciones).
- Las constantes $b_1, \ldots b_m \in \mathbb{R}$ y $c_1, \ldots c_p \in \mathbb{R}$ son llamadas límites o fronteras de las restricciones.

El conjunto $Z = \{x \in \mathbb{R}^n : f_i(x) \leq b_i \text{ y } h_j(x) = c_j\}$ se denomina conjunto factible o zona factible.

Con solución:

**Definición 1.2.7.** Un vector $x^* \in \mathbb{R}^n$ se denomina óptimo (o una solución) del problema dado en la Definición (1.2.6) si $f(x^*)$ tiene el valor objetivo más pequeño entre todos los vectores que satisfacen las restricciones; es decir, para cualquier $z \in \mathbb{R}^n$, con $f_1(z) \leq b_1, \ldots, f_m(z) \leq b_m$ y $h_1(z) = c_1, ..., h_p(z) = c_p$, se tiene que $f(z) \geq f(x^*)$.

Los problemas de optimización se agrupan en familias de acuerdo con la función objetivo y las restricciones. El problema de la Definición (1.2.6) se llama *de programación lineal* si la función objetivo y las restricciones son lineales. En dado caso de que el problema no sea lineal, se denomina *de programación no lineal*. Un caso particular es cuando dentro de la función objetivo se encuentra un término cuadrático; en ese caso, se le conoce como *programación cuadrática* (Véase la Definición 1.2.11).

**Definición 1.2.8.** Un problema de la forma

$$\begin{aligned} &\text{Minimizar} \quad f(x) && \\ &\text{sujeto a } f_i(x) \leq b_i, && i = 1, \ldots, m, \end{aligned} \tag{1.2.3}$$

con $f : \mathbb{R}^n \to \mathbb{R}$ y $f_i : \mathbb{R}^n \to \mathbb{R}$ con $i = 1, ..., m$ es llamado un problema de optimización convexa si para cualesquiera $x, y \in \mathbb{R}^n$ y

$\alpha, \beta \in \mathbb{R}$ tales que $\alpha + \beta = 1$ con $\alpha, \beta \geq 0$ se satisface

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y). \tag{1.2.4}$$

y además la función objetivo $f(x)$ es una función convexa.

De aquí en adelante, esta investigación estará centrada en los problemas de optimización convexa.

Desde finales de la década de 1940, se han realizado distintos esfuerzos para desarrollar algoritmos que permitan encontrar soluciones a los problemas de optimización, analizar sus propiedades y desarrollar implementaciones de software. La efectividad de los algoritmos de resolución del problema de optimización (1.2.3) con restricciones (1.2.4) es variada y depende de factores como las formas particulares de las funciones objetivo y las restricciones; también depende de la cantidad de variables, así como de la cantidad de restricciones; incluso la dimensionalidad del espacio de entrada el cual es el espacio donde es encuentran los datos de entrenamiento. En este trabajo de tesis se toma como espacio de entrada $\mathbb{R}^n$.

Por lo tanto, los enfoques al problema general implican un tiempo de cálculo de la solución muy largo o la posibilidad de no encontrarla. Sin embargo, para algunas clases de problemas existen algoritmos efectivos que pueden resolver de manera confiable distintos problemas con cientos de variables y restricciones; un ejemplo de ello son los problemas de mínimos cuadrados y los programas lineales. Por esta razón, se introduce esta metodología en la sección siguiente.

### 1.2.2. Mínimos cuadrados

Supongamos que se han seleccionado $m$ números reales y han sido organizados como las $m$ componentes de un vector $y$. A menudo, la naturaleza de la fuente de los datos lleva a suponer que el vector $y$, en lugar de consistir en $m$ componentes independientes, es una función lineal dada de unos pocos parámetros desconocidos. Si estos parámetros se disponen como los componentes de un vector $n$-dimensional $\beta$ (donde $n \leq m$), es equivalente a asumir que el vector $y$

tiene la forma

$$y = W\beta. \tag{1.2.5}$$

La matriz $W$ de dimensión $m \times n$ se supone que es conocida y está determinada por el experimento o por la situación física en cuestión. El vector $y$ también es conocido, así que el problema se basa en encontrar el vector $\beta$. Sin embargo, dado que $n \leq m$, generalmente no es posible determinar el vector $\beta$ que satisfaga exactamente $y = W\beta$. Una forma útil de determinar el valor de $\beta$ que mejor aproxima la solución del problema consiste en minimizar la norma de la diferencia:

$$\|y - W\beta\|. \tag{1.2.6}$$

Si la norma se toma como la norma euclidiana, entonces la norma mostrada en (1.2.6) se convierte en:

**Definición 1.2.9. (Problema de mínimos cuadrados).** Un problema de optimización se denomina problema de mínimos cuadrados si es de la forma:

$$\text{Minimizar} \qquad f(\beta) := \|y - W\beta\|_2^2 = \sum_{i=1}^{k} (y_i - w_i^T \beta)^2, \tag{1.2.7}$$

donde $W \in \mathbb{R}^{m \times n}$ con $n \leq m$, $w_i^T$ son las filas de $W$, el vector $\beta \in \mathbb{R}^n$ es la variable de optimización y $f : \mathbb{R}^n \to \mathbb{R}$.

La solución del problema de mínimos cuadrados de la Definición (1.2.9) se encuentra como consecuencia del siguiente teorema.

**Teorema 1.2.10. (Estimador de mínimos cuadrados)** [19, Teorema 1, p. 83] Supongamos que $y$ es un vector en $\mathbb{R}^m$ y $W$ una matriz de $m \times n$ con columnas linealmente independientes. Entonces existe un único vector $\beta^*$ de dimensión $n$ el cual minimiza $\|y - W\beta\|_2$ sobre todos los $\beta$ en $\mathbb{R}^n$. Además

$$\beta^* = (W^t W)^{-1} W^t y. \tag{1.2.8}$$

La existencia y unicidad del vector solución en el teorema de estimador de mínimos cua-

drados se deben al teorema de proyección (véase [19]) y a la independencia de las columnas de $W$. Además, la matriz de Gram correspondiente a los vectores columna de $W$ es $W^tW$. El vector $W^ty$ tiene como componentes los productos internos de las columnas de $W$ con el vector $y$. Por lo tanto, las ecuaciones normales son $W^tW\beta = W^ty$. Dado que se supone que las columnas de $W$ son linealmente independientes, la matriz de Gram $W^tW$ es no singular y la conclusión del Teorema 1.2.10 se sigue.

Para los problemas de mínimos cuadrados, existen algoritmos de resolución que encuentran el óptimo en un tiempo aproximadamente proporcional a $n^2k$, con $k$ constante conocida. Un programa de computadora puede resolver problemas con cientos de variables y miles de términos en unos pocos segundos; esto respecto a la capacidad de la computadora con la que se resuelva el problema. Para problemas con millones de variables, o para problemas con requisitos relativamente altos de computación en tiempo real, resolver un problema de mínimos cuadrados puede ser un reto. En la mayoría de los casos, existen métodos que son efectivos y extremadamente fiables.

Los problemas de mínimos cuadrados son la base para el análisis de regresión, el control óptimo y los métodos de estimación de parámetros y ajuste de datos. Tiene una serie de interpretaciones estadísticas; por ejemplo, la estimación de máxima verosimilitud de un vector $x$, dadas algunas mediciones corruptas por errores de medición gaussianos, entre otras.

Un problema de optimización cuadrática (o problema cuadrático QP) es uno de la forma de la Definición 1.2.6, donde la función objetivo es cuadrática y las respectivas restricciones son funciones afines. Por lo tanto, el conjunto que satisface las restricciones del problema de optimización (conjunto factible) de un QP es un poliedro (como en los problemas lineales), pero el objetivo es cuadrático en lugar de lineal. La forma estándar de un QP es la siguiente:

**Definición 1.2.11. (Problema de optimización cuadrática.)** Se define un problema de optimización cuadrática de la forma:

$$\text{Minimizar} \quad \frac{1}{2}x^T Hx + c^T x, \tag{1.2.9}$$

$$\text{sujeto a} \quad A_{eq}x = b_{eq}, \tag{1.2.10}$$

$$Ax \leq b, \tag{1.2.11}$$

donde:

- $x \in \mathbb{R}^n$ es la variable de optimización.
- El vector $c$ de (1.2.9) es un vector en $\mathbb{R}^n$ que contiene los coeficientes de los términos de grado uno.
- $H$ de (1.2.9) es una matriz simétrica en $\mathbb{R}^{n\times n}$,
- La matriz $A_{eq}$ de (1.2.10) está en $\mathbb{R}^{m\times n}$, la cual contiene los coeficientes de las restricciones de igualdad.
- La matriz $A$ de (1.2.11) está en $\mathbb{R}^{k\times n}$ y contiene los coeficientes de desigualdad.
- El vector $b_{eq}$ de (1.2.10) está en $\mathbb{R}^m$ y tiene los coeficientes de igualdad.
- El vector $b$ de (1.2.11) se encuentra en $\mathbb{R}^k$ y tiene como entradas los coeficientes de desigualdad.

**Teorema 1.2.12.** [9, p. 71] Sea $f : \mathbb{R}^n \to \mathbb{R}$ de la forma $f(x) = \frac{1}{2}x^T Hx + c^T x$, con $H$ una matriz cuadrada de tamaño $n$ y $c \in \mathbb{R}^n$. La función $f$ es convexa si y solo si la matriz $H$ es semidefinida positiva.

Para identificar un problema de optimización como un problema de mínimos cuadrados se requiere verificar que la función objetivo sea una función cuadrática, seguido de probar si la

forma matricial de la función cuadrática es semidefinida positiva. Con lo que se observa su relación con los problemas de SVM.

### 1.2.3. Hiperplanos y funcionales lineales

Un componente central de los modelos de máquinas de soporte vectorial son los hiperplanos de separación. En consecuencia, esta sección tiene como objetivo introducir las definiciones y resultados asociados a ellos, apoyándose en el uso de transformaciones y funcionales lineales.

**Definición 1.2.13. (Hiperplano).** Sea $V$ un espacio vectorial sobre un campo $F$. Sea $f : V \to F$ un funcional lineal no nulo sobre $V$ y sea $c$ una constante en $F$. El conjunto $H = \{x : f(x) = c\}$ es un hiperplano en $V$.

En caso de que el conjunto

$$H = \{x : f(x) = c\}$$

es cerrado, entonces se dice que el hiperplano $H$ es cerrado.

Excluyendo los hiperplanos que contienen el origen, es posible establecer una correspondencia única entre los hiperplanos y los funcionales lineales, como se muestra en la siguiente proposición.

**Proposición 1.2.14.** [19, Proposición 2, p.130] Sean $V$ un espacio vectorial real y $H$ un hiperplano que no contiene al vector 0. Entonces existe un único funcional lineal no nulo $f : V \to \mathbb{R}$ tal que

$$H = \{x \in V : f(x) = 1\}.$$

Existe una relación estrecha entre los hiperplanos y los funcionales continuos, por lo que a continuación se introduce la definición correspondiente a esta noción de continuidad.

**Definición 1.2.15. (Transformación continua.)** Sean $X$ y $Y$ dos espacios normados con sus respectivas normas $\|\cdot\|_X$ y $\|\cdot\|_Y$. Una transformación $T : X \to Y$ se dice que es continua en el punto $x_0 \in X$ si, para cada $\varepsilon > 0$, existe un $\delta > 0$ tal que $\|x - x_0\|_X < \delta$ implica que $\|T(x) - T(x_0)\|_Y < \varepsilon$.

Si la transformación $T$ es continua en cada punto $x_0 \in X$, entonces se dice que la transformación $T$ es continua donde sea o, simplemente, que $T$ es continua. Particularmente, se puede establecer la continuidad de una transformación con el siguiente resultado.

**Proposición 1.2.16.** [19, Proposición 1, p. 144] Una transformación lineal $T : X \to Y$ es continua en cada punto de $X$ si es continua en un solo punto.

Las transformaciones lineales continuas pueden caracterizarse en términos de su acotación. Para establecer esta relación, se introduce la siguiente definición.

**Definición 1.2.17. (Transformación acotada).** Sean $X$ y $Y$ dos espacios normados y $T : X \to Y$ una transformación lineal. Se dice que $T$ es acotada si existe una constante $M \geq 0$ tal que $\|Tx\|_Y \leq M\|x\|_X$, para cada $x$ en $X$. La constante $M$ más pequeña que satisface la condición anterior se denota como $\|T\|$ y se denomina la norma de $T$.

**Proposición 1.2.18.** [19, Proposición 2, p. 144] Sean $X$ y $Y$ dos espacios normados y $T : X \to Y$ una transformación lineal. El operador $T$ es acotado si y sólo si es continuo.

La continuidad de los funcionales se define como:

**Definición 1.2.19. (Funcional semicontinuo superiormente).** Sea $X$ un espacio normado con campo $\mathbb{R}$ y un funcional $f$ sobre $X$. El funcional $f$ es semicontinuo superiormente en $x_0$ si dado un $\varepsilon > 0$ existe un $\delta > 0$, tal que si $\|x - x_0\| < \delta$, entonces $f(x) - f(x_0) < \varepsilon$. El funcional $f$ se dice que es semicontinuo inferiormente en $x_0$ si el funcional $-f$ es semicontinuo superiormente en $x_0$.

**Definición 1.2.20. (Funcional continuo).** Sea $X$ un espacio normado con campo $\mathbb{R}$ y $f$ un funcional sobre $X$. El funcional $f$ es continuo si es semicontinuo inferiormente y superiormente.

Ahora bien, la relación entre los funcionales continuos y los hiperplanos es establecida por la siguiente proposición.

**Proposición 1.2.21.** [19, Proposición 3, p.130] Sea $f : X \to \mathbb{R}$ un funcional no nulo sobre un espacio normado $X$. El hiperplano $H = \{x : f(x) = c\}$ es cerrado para cada $c$ si y solo si $f$ es continuo.

**Definición 1.2.22.** Sea $f : V \to \mathbb{R}$ un funcional no nulo sobre un espacio vectorial real $V$. Se asocia el hiperplano $H = \{x : f(x) = c\}$ a los cuatro conjuntos:

$$\{x : f(x) \leq c\}, \qquad \{x : f(x) < c\}, \qquad \{x : f(x) \geq c\}, \qquad \{x : f(x) > c\}. \tag{1.2.12}$$

Estos conjuntos son llamados semiespacios determinados por $H$.

Los semiespacios $\{x : f(x) \leq c\}$, $\{x : f(x) < c\}$ son denominados semiespacios negativos determinados por $f$, mientras que los semiespacios $\{x : f(x) \geq c\}$, $\{x : f(x) > c\}$ son los positivos. Si $f$ es continua, entonces los semiespacios $\{x : f(x) < c\}$ y $\{x : f(x) > c\}$ son conjuntos abiertos y los semiespacios $\{x : f(x) \geq c\}$ y $\{x : f(x) \leq c\}$ son cerrados.

**Definición 1.2.23. (Hiperplano de soporte).** Un hiperplano cerrado $H$ en un espacio normado $X$ es un soporte (hiperplano de soporte) para un conjunto convexo $K$, si $K$ está contenido en uno de los semiespacios cerrados determinados por $H$ y $H$ contiene un punto de $\overline{K}$.

**Definición 1.2.24. (Hiperplano orientado.)** Sea $w \in \mathbb{R}^n$ un vector no nulo y $b \in \mathbb{R}$. El **hiperplano** asociado a $(w, b)$ se define como

$$H = \{x \in \mathbb{R}^n : \langle w, x\rangle + b = 0\}.$$

Cuando el vector $w$ se fija, se le llama vector normal al plano $H$ y se dice que $H$ es un **hiperplano orientado**. Más aún, la orientación dada por $w$ permite distinguir entre los dos semiespacios:

$$H^+ = \{x \in \mathbb{R}^n : \langle w, x\rangle + b > 0\}, \qquad H^- = \{x \in \mathbb{R}^n : \langle w, x\rangle + b < 0\}.$$

**Ejemplo 1.2.25.** En $\mathbb{R}^2$, la recta $H = \{(x, y) : x + y = 0\}$ se escribe como

$$x + y = (1)x + (1)y = \langle (1, 1), (x, y)\rangle = 0.$$

Por lo que el hiperplano se orienta mediante el vector normal $w = (1, 1)$. Entonces $H^+$ corresponde al semiplano $\{(x, y) : x + y > 0\}$ y $H^-$ al semiplano $\{(x, y) : x + y < 0\}$.

**Ejemplo 1.2.26.** En $\mathbb{R}^3$, el plano $H = \{(x, y, z) : x + y + z = 0\}$ se representa como

$$x + y + z = (1)x + (1)y + (1)z = \langle (1, 1, 1), (x, y, z)\rangle = 0.$$

Por lo que el hiperplano se orienta mediante el vector normal $w = (1, 1, 1)$. Esto permite distinguir los semiespacios

$$H^+ = \{(x, y, z) : x + y + z > 0\}, \quad H^- = \{(x, y, z) : x + y + z < 0\}.$$

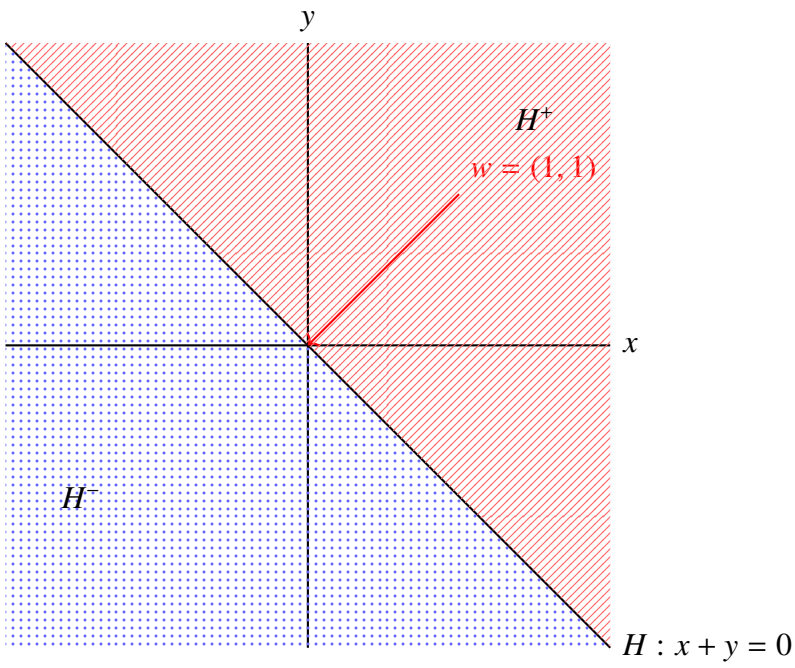


**Figura 1.1:** Hiperplano de separación con semiespacios diferenciados por texturas.

**Teorema 1.2.27. (Teorema del soporte)** [19, Teorema 2, p. 133] Sea $V$ un espacio vectorial y $K \subseteq V$ un conjunto convexo. Si $x$ no es un punto interior de $K$, entonces hay un hiperplano cerrado $H = \{x \in V : \langle w, x\rangle + b = 0\}$ que contiene a $x$ tal que $K$ se encuentra en un lado de $H$, es decir para cada $k \in K$, satisface que $\langle w, k\rangle + b > 0$ o bien $\langle w, k\rangle + b < 0$.

### 1.2.4. Separación e hiperplanos de soporte

En esta sección se estudia el uso de hiperplanos, o de manera equivalente funciones afines, como herramientas fundamentales para separar conjuntos convexos disjuntos. En particular, se analiza bajo qué condiciones es posible encontrar un hiperplano que distinga adecuadamente a dichos conjuntos, asignándolos a semiespacios opuestos. El resultado central que sustenta este análisis es el teorema del hiperplano de separación, el cual garantiza la existencia de un hiperplano separador para pares de conjuntos convexos que no se intersectan, y constituye una pieza clave en el desarrollo de la optimización convexa y de los métodos de clasificación.

**Teorema 1.2.28. (Teorema del hiperplano de separación)** [9, Teorema 2.5.1, p. 46]. Sea $V$ un espacio vectorial real prehilbertiano. Si $C$ y $D$ son dos conjuntos convexos de $V$ tales que $C \cap D = \emptyset$, entonces existe un vector $a \neq 0$ en $V$ y un escalar $b$ tales que $\langle a, x\rangle \leq b$, para todo $x$ en $C$, y $\langle a, x\rangle \geq b$, para cada $x$ en $D$. Es decir, la función afín $\langle a, x\rangle - b$ es no positiva sobre $C$ y no negativa sobre $D$. El hiperplano $H = \{x \mid \langle a, x\rangle = b\}$ se denomina hiperplano de separación para los conjuntos $C$ y $D$.

Una ilustración gráfica del hiperplano de separación se observa en la Figura 1.2. Cabe mencionar que si el hiperplano de separación satisface la condición $\langle a, x\rangle < b$, para todo $x$ en $C$, y $\langle a, x\rangle > b$, para todo $x$ en $D$, entonces se denomina separación estricta de los conjuntos $C$ y $D$.

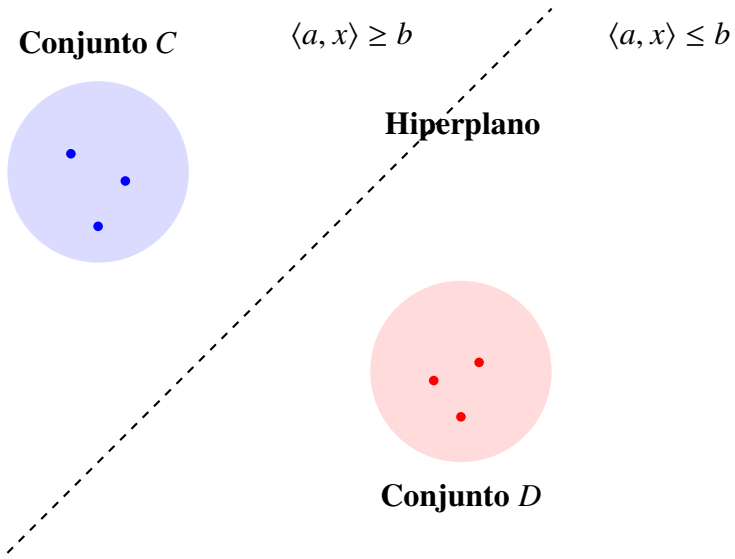


**Figura 1.2:** El hiperplano $H = \{x \mid \langle a, x\rangle = b\}$ separa los conjuntos convexos $C$ y $D$.

El recíproco del Teorema 1.2.28 no es cierto de manera general; es decir, es posible que exista un hiperplano de separación entre $C$ y $D$, y que $C \cap D \neq \emptyset$. Como se muestra en el siguiente ejemplo.

**Ejemplo 1.2.29.** Sea $C = \{(x, y) : (x + 1)^2 + y^2 \leq 1\}$ y el rectángulo $D = [0, 3] \times [-1, 1]$. Note que $(0, 0) \in D$ y $(0, 0) \in C$, por lo que $C \cap D \neq \emptyset$. Considere el hiperplano

$$H = \{(x, y) \in \mathbb{R}^2 : x = 0\}.$$

Ahora, se tiene que

$$c_1 \leq 0, \qquad \text{para todo } c = (c_1, c_2) \in C,$$

$$d_1 \geq 0, \qquad \text{para todo } d = (d_1, d_2) \in D,$$

por lo que el hiperplano $H$ separa a $C$ y $D$. Gráficamente se observa en la Figura 1.3.

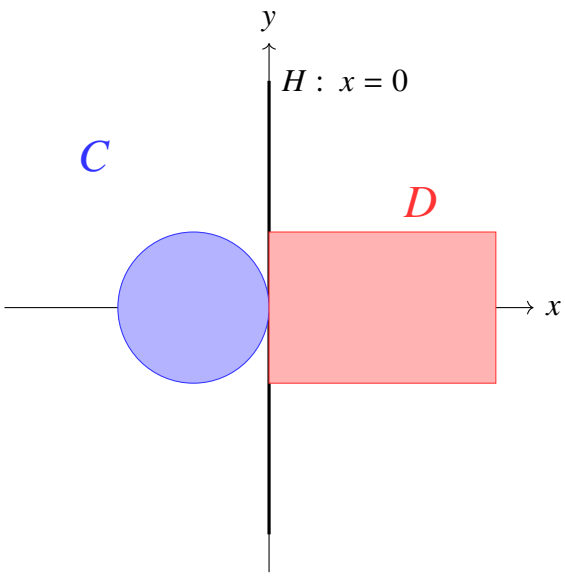


**Figura 1.3:** Separación débil a color: $C$ (disco) y $D$ (rectángulo) son convexos y satisfacen $C \cap D \neq \emptyset$, pero están separados por el hiperplano $H : x = 0$.

A aquellos conjuntos que sí cumplen con el recíproco se les denomina de una forma especial:

**Definición 1.2.30. (Conjuntos linealmente separables).** Sean $C$, $D$ dos conjuntos en un espacio vectorial prehilbertiano real $V$. Los conjuntos $C$ y $D$ se dicen linealmente separables si existe un hiperplano $H = \{x \in V : \langle w, x\rangle + b = 0\}$ tal que

$$\langle w, d\rangle + b > 0, \quad \text{para todo } d \in D,$$

y

$$\langle w, c\rangle + b < 0, \quad \text{para todo } c \in C.$$

En espacios finito dimensionales, y bajo ciertas condiciones adicionales sobre $C$ y $D$, más amplias que la convexidad, se cumple el recíproco; de hecho, caracteriza a los conjuntos linealmente separables:

**Teorema 1.2.31.** [18, Corolario 13, p. 5] Sean dos conjuntos $C$ y $D$ en $\mathbb{R}^n$. Los conjuntos $C$ y $D$ son linealmente separables si y solo si sus envolventes convexas $\text{conv}(C)$ y $\text{conv}(D)$ son disjuntas.

### 1.2.5. Funciones convexas

Como se ha mencionado previamente, las SVM son modelos de optimización cuyas restricciones son convexas en el sentido de la definición siguiente.

**Definición 1.2.32. (Función convexa).** Una función $f : D \subset \mathbb{R}^n \to \mathbb{R}$ es convexa si $D$ es un conjunto convexo, y si para todo $x$ y $y$ en el dominio de $f$ y $0 \leq \theta \leq 1$, se tiene que

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y). \tag{1.2.13}$$

Una interpretación geométrica de la desigualdad (1.2.13) es que el segmento entre los puntos $(x, f(x))$ y $(y, f(y))$, que es la cuerda de $x$ a $y$, se encuentra acotando superiormente a la función $f$ entre los puntos $x$ y $y$. Un ejemplo de esto se muestra a continuación.

**Ejemplo 1.2.33.** Sean $x, y \in \mathbb{R}$ y $\theta \in [0,1]$. Se probará que la función $f : \mathbb{R} \to \mathbb{R}$ dada por $f(x) = x^2$ es una función convexa:

$$\begin{aligned}\theta x^2 + (1-\theta)y^2 - (\theta x + (1-\theta)y)^2 &= \theta x^2 + (1-\theta)y^2 - (\theta^2 x^2 + 2\theta(1-\theta)xy \\ &\quad + (1-\theta)^2 y^2) \\ &= (\theta - \theta^2)x^2 - 2\theta(1-\theta)xy + ((1-\theta) \\ &\quad - (1-\theta)^2)y^2 \\ &= \theta(1-\theta)x^2 - 2\theta(1-\theta)xy + \theta(1-\theta)y^2 \\ &= \theta(1-\theta)(x^2 - 2xy + y^2) \\ &= \theta(1-\theta)(x-y)^2.\end{aligned}$$

Como $\theta \geq 0$, $1-\theta \geq 0$ y $(x-y)^2 \geq 0$, entonces el producto $\theta(1-\theta)(x-y)^2 \geq 0$, es decir

$$\begin{aligned}\theta x^2 + (1-\theta)y^2 - (\theta x + (1-\theta)y)^2 &\geq 0, \\ \theta x^2 + (1-\theta)y^2 &\geq (\theta x + (1-\theta)y)^2,\end{aligned}$$

esto es,

$$\theta f(x) + (1-\theta)f(y) \geq f(\theta x + (1-\theta)y),$$

con lo que se concluye que la función $f$ es convexa. Particularmente, el segmento que va del punto $(0,0)$ al punto $(2,4)$ va a acotar a la función $f(x) = x^2$ en el intervalo $[0,2]$. Esto se ilustra en la Figura 1.4.

**Definición 1.2.34.** Una función $f : C \subseteq \mathbb{R}^n \to \mathbb{R}$, es estrictamente convexa si la desigualdad (1.2.13) es estricta siempre que $x \neq y$ y $0 < \theta < 1$.

**Definición 1.2.35.** Una función $f : C \to \mathbb{R}^n$ con $C \subseteq \mathbb{R}^n$ es cóncava si $-f$ es convexa; más aún, $f$ es estrictamente cóncava si $-f$ es estrictamente convexa.

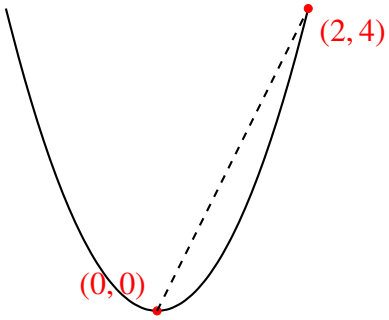


**Figura 1.4:** Función convexa. Obsérvese que la cuerda trazada del punto (0,0) al (2,4) se encuentra acotando superiormente a la función.

### 1.2.6. Dualidad

En esta sección se introduce el concepto de multiplicadores de Lagrange, que constituye una herramienta fundamental en la formulación y el análisis de problemas de optimización con restricciones. En particular, se explica cómo estos multiplicadores permiten transformar un problema primal con restricciones en un problema dual, cuyo estudio resulta esencial para el desarrollo teórico de las máquinas de soporte vectorial. Asimismo, se presenta el teorema de dualidad fuerte, el cual establece condiciones bajo las cuales los valores óptimos de los problemas primal y dual coinciden. Finalmente, se discute el proceso de recuperación de soluciones primales a partir de soluciones duales, aspecto clave para la interpretación geométrica y computacional de la SVM y para la obtención explícita del hiperplano de separación óptimo.

Considérese el problema de optimización con la forma de la Definición 1.2.6:

$$\begin{aligned} \text{Minimizar} \quad & f(x) \\ \text{sujeto a} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m, \\ & h_j(x) = 0, \quad j = 1, \ldots, p, \end{aligned} \tag{1.2.14}$$

con variable $x \in \mathbb{R}^n$. Defínase $\mathcal{D} = \bigcap_{i=0}^{m} \mathbf{dom}\ f_i \cap \bigcap_{j=1}^{p} \mathbf{dom}\ h_j$ como el conjunto dominio del problema, donde $\mathbf{dom}\ g_i$ y $\mathbf{dom}\ h_j$ son los dominios de cada función $g_i$ y $h_j$, respectivamente. Asúmase que $\mathcal{D}$ es no vacío y $x^*$ es el óptimo del problema; finalmente, $f(x^*) = p^* \in \mathbb{R}$ denota el valor óptimo del problema de optimización 1.2.14. Por el momento, no se asume que el

problema sea convexo.

La función que permite transformar un problema de optimización en su formulación dual es la función lagrangiana, la cual se construye a partir del problema primal incorporando sus restricciones. En la siguiente definición se presenta explícitamente la función lagrangiana asociada, así como la función dual correspondiente.

**Definición 1.2.36. (Lagrangiano)** El *lagrangiano* es la función $\mathcal{L} : \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ asociada al problema de optimización (1.2.14) definida como

$$\mathcal{L}(x, \alpha, \nu) = f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{j=1}^{p} \nu_j h_j(x), \tag{1.2.15}$$

para toda $x \in \mathcal{D}$, $\alpha \in \mathbb{R}^m$ y $\nu \in \mathbb{R}^p$. Cada $\alpha_i$ es el multiplicador de Lagrange asociado a la $i$-ésima restricción de desigualdad $g_i(x) \leq 0$; de manera similar, cada $\nu_j$ es el multiplicador de Lagrange respectivo a cada restricción de igualdad $h_j(x) = 0$. Los vectores $\alpha$ y $\nu$ son llamados las variables duales, o simplemente *multiplicadores de Lagrange* asociados al problema de optimización (1.2.14).

**Definición 1.2.37. (Función dual de la Lagrange)** La *función dual de Lagrange* (o solo *función dual*) $\theta : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$, asociada al problema de optimización (1.2.14), es el valor mínimo del Lagrangiano $\mathcal{L} : \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ tomado sobre $x \in \mathcal{D}$, es decir, para todo $\alpha \in \mathbb{R}^m,\ \nu \in \mathbb{R}^p$,

$$\theta(\alpha, \nu) = \inf_{x \in \mathcal{D}} \mathcal{L}(x, \alpha, \nu) = \inf_{x \in \mathcal{D}} \left( f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{j=1}^{p} \nu_j h_j(x) \right). \tag{1.2.16}$$

Cuando el ínfimo del Lagrangiano $\mathcal{L}(x, \alpha, \nu)$, para $x \in \mathcal{D}$, no existe, entonces la función dual de Lagrange $\theta$ toma el valor $-\infty$. Dado que la función dual es el ínfimo puntual de una familia de funciones afines de $(\alpha, \nu)$, entonces es cóncava, incluso cuando el problema de optimización (1.2.14) no es convexo.

El siguiente teorema es clave para establecer la relación entre el problema de optimización de la forma (1.2.3), que de ahora en adelante se llamará problema primal, y su representación

dual; además, establece la manera en la que se puede calcular la solución del problema primal a partir de la solución del problema dual. Es por ello que comúnmente se le denomina a esta técnica "recuperación del óptimo". Esta recuperación se puede dar si se cumple la condición de Slater, lo que implica que haya una brecha de dualidad nula y se pueda alcanzar el óptimo a partir de la solución del problema dual. La demostración ilustra la utilidad de las condiciones impuestas, por lo que también se exhibe.

**Teorema 1.2.38** (Dualidad fuerte y recuperación primal desde el dual)**.** Considérese el problema de optimización convexo

$$\begin{aligned} \text{Minimizar} \quad & f(x) \\ \text{sujeto a} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, m, \\ & h_j(x) = 0, \quad j = 1, \ldots, p, \end{aligned} \tag{1.2.17}$$

donde $f : \mathbb{R}^n \to \mathbb{R}$ y cada $g_i : \mathbb{R}^n \to \mathbb{R}$ son funciones convexas y cada $h_j : \mathbb{R}^n \to \mathbb{R}$ es afín. Sea $\mathcal{L} : \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ con $\mathcal{D}$ el dominio de $f$, el Lagrangiano

$$\mathcal{L}(x, \alpha, \nu) \;=\; f(x) \;+\; \sum_{i=1}^{m} \alpha_i \, g_i(x) \;+\; \sum_{j=1}^{p} \nu_j \, h_j(x),$$

con multiplicadores $\alpha \in \mathbb{R}^m$ y $\nu \in \mathbb{R}^p$, y la función dual $\theta : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$

$$\theta(\alpha, \nu) \;=\; \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \alpha, \nu).$$

El problema dual es

$$\begin{aligned} \text{Maximizar} \quad & \theta(\alpha, \nu) \\ \text{sujeto a} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, m, \end{aligned} \tag{1.2.18}$$

Supóngase que se cumple la condición de Slater: existe $\bar{x} \in \mathcal{D}$ tal que $g_i(\bar{x}) < 0$, para toda $i = 1, \ldots, m$, y $h_j(\bar{x}) = 0$, para toda $j = 1, \ldots, p$. Entonces:

1. (KKT $\Rightarrow$ optimalidad) Si existe $(x^\star, \alpha^\star, \nu^\star) \in \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$ que satisface las condiciones de Karush–Kuhn–Tucker:

$$\begin{aligned}
&\text{(factibilidad primal)} && g_i(x^\star) \leq 0, \quad h_j(x^\star) = 0, \\
&\text{(factibilidad dual)} && \alpha_i^\star \geq 0, \\
&\text{(complementariedad)} && \alpha_i^\star \, g_i(x^\star) = 0, \quad i = 1, \ldots, m, \\
&\text{(estacionariedad)} && 0 = \nabla f(x^\star) + \sum_{i=1}^{m} \alpha_i^\star \nabla g_i(x^\star) \; + \; \sum_{j=1}^{p} \nu_j^\star \nabla h_j(x^\star),
\end{aligned}$$

   entonces $x^\star$ es óptimo primal y $(\alpha^\star, \nu^\star)$ es óptimo dual.

2. (Dualidad fuerte) El óptimo dual es alcanzado, es decir, existen $(\alpha^\star, \nu^\star) \in \mathbb{R}^m \times \mathbb{R}^p$ tales que $g(\alpha^*, \nu^*) = d^*$. Los elementos $p^*$, el óptimo del problema (1.2.17), y $d^*$, el óptimo del problema (1.2.18), satisfacen

$$p^* = d^*.$$

3. (Recuperación primal desde el dual) Si $(\alpha^\star, \nu^\star)$ es solución de (1.2.18), todo punto $x^\star$ en el cual el mínimo de $\mathcal{L}(x, \alpha^\star, \nu^\star)$ es alcanzado, es decir,

$$x^\star \in \operatorname*{argmin}_x \mathcal{L}(x, \alpha^\star, \nu^\star),$$

   satisface las KKT y, por tanto, es solución óptima de (1.2.17).

*Demostración. (Debilidad de la dualidad).* Para todo $x$ factible en (1.2.17) y todo $(\alpha, \nu)$ con $\alpha \geq 0$,

$$\mathcal{L}(x, \alpha, \nu) = f(x) + \sum_i \alpha_i g_i(x) + \sum_j \nu_j h_j(x) \leq f(x),$$

pues $g_i(x) \leq 0$ y $h_j(x) = 0$. Tomando el ínfimo en $x$ de $\mathcal{L}$ se tiene que

$$\theta(\alpha, \nu) \;\leq\; f(x).$$

Particularmente, en el óptimo $x^\star$

$$\theta(\alpha, \nu) \;\leq\; f(x^\star),$$

para todo $(\alpha, \nu)$, es decir,

$$d^* \leq p^*.$$

*(KKT $\Rightarrow$ optimalidad).* Si $(x^\star, \alpha^\star, \nu^\star)$ satisfacen KKT, entonces $x^\star$ es factible y la estacionariedad dice que $0 = \nabla\mathcal{L}(x^\star, \alpha^\star, \nu^\star)$, por lo que $x^\star$ minimiza $\mathcal{L}(x, \alpha^\star, \nu^\star)$. El hecho de que $\alpha^\star \geq 0$ y la factibilidad primal implican

$$\mathcal{L}(x^\star, \alpha^\star, \nu^\star) = f(x^\star).$$

Usando la debilidad de la dualidad,

$$d^* = \theta(\alpha^\star, \nu^\star) \;\leq\; f(x^\star) \;=\; p^*.$$

Pero como $x^\star$ minimiza a $\mathcal{L}$, $\theta(\alpha^\star, \nu^\star) = \mathcal{L}(x^\star, \alpha^\star, \nu^\star) = f(x^\star)$, y esta igualdad implica que $x^\star$ y $(\alpha^\star, \nu^\star)$ son óptimos primal y dual, respectivamente.

*(Dualidad fuerte bajo Slater).* Bajo la hipótesis de la condición de Slater, el conjunto factible del problema primal posee interior no vacío. Dado que $f$ y las funciones $g_i$ son convexas y las funciones $h_j$ son afines, la condición de Slater garantiza que:

- existe ausencia de brecha de dualidad, es decir, se tiene la igualdad entre los valores óptimos primal y dual;
- existen los multiplicadores de Lagrange óptimos $\alpha^\star$ y $\nu^\star$;
- se validan las condiciones de Karush–Kuhn–Tucker (KKT).

Como consecuencia de estos hechos, existe un punto primal óptimo $x^\star$ y multiplicadores $\alpha^\star \geq 0$, $\nu^\star$ tales que la tripleta $(x^\star, \alpha^\star, \nu^\star)$ satisface las condiciones KKT.

Por la condición de estacionariedad de KKT se tiene

$$0 = \partial_x \mathcal{L}(x^\star, \alpha^\star, \nu^\star).$$

Esto implica que $x^\star$ es un minimizador de la función $\mathcal{L}(x, \alpha^\star, \nu^\star)$. Por definición de la función dual,

$$\theta(\alpha^\star, \nu^\star) = \inf_x \mathcal{L}(x, \alpha^\star, \nu^\star).$$

Como en $x^\star$ se alcanza el ínfimo, se cumple

$$\theta(\alpha^\star, \nu^\star) = \mathcal{L}(x^\star, \alpha^\star, \nu^\star),$$

y por tanto

$$\mathcal{L}(x^\star, \alpha^\star, \nu^\star) \leq \mathcal{L}(x, \alpha^\star, \nu^\star), \qquad \text{para toda } x \in \mathbb{R}^n. \tag{1.2.19}$$

Sea $(\alpha, \nu)$ arbitrario con $\alpha \geq 0$. Considere la diferencia

$$\mathcal{L}(x^\star, \alpha, \nu) - \mathcal{L}(x^\star, \alpha^\star, \nu^\star).$$

Por definición de la Lagrangiana,

$$\begin{aligned}\mathcal{L}(x^\star, \alpha, \nu) - \mathcal{L}(x^\star, \alpha^\star, \nu^\star) = f(x^\star) \; &+ \sum_{i=1}^{m} \alpha_i \, g_i(x^\star) + \sum_{j=1}^{p} \nu_j \, h_j(x^\star) - f(x^\star) - \sum_{i=1}^{m} \alpha_i^\star \, g_i(x^\star) \\ &- \sum_{j=1}^{p} \nu_j^\star \, h_j(x^\star) \\ &= \sum_{i=1}^{m} (\alpha_i - \alpha_i^\star) g_i(x^\star) + \sum_{j=1}^{p} (\nu_j - \nu_j^\star) h_j(x^\star).\end{aligned}$$

Por factibilidad primal, $h_j(x^\star) = 0$ para todo $j$, por lo que el segundo sumando desaparece.

Quedando

$$\mathcal{L}(x^\star, \alpha, \nu) - \mathcal{L}(x^\star, \alpha^\star, \nu^\star) = \sum_{i=1}^{m} (\alpha_i - \alpha_i^\star) g_i(x^\star).$$

Ahora, considere los siguientes casos para cada índice $i$:

- Si $g_i(x^\star) < 0$, entonces $\alpha_i^\star = 0$, y en consecuencia

$$(\alpha_i - \alpha_i^\star) g_i(x^\star) = \alpha_i g_i(x^\star) \leq 0,$$

  ya que $\alpha_i \geq 0$ y $g_i(x^\star) < 0$.

- Si $g_i(x^\star) = 0$, entonces

$$(\alpha_i - \alpha_i^\star) g_i(x^\star) = 0.$$

En todos los casos, cada término de la suma es no positivo, por lo que

$$\mathcal{L}(x^\star, \alpha, \nu) - \mathcal{L}(x^\star, \alpha^\star, \nu^\star) \leq 0,$$

o equivalentemente,

$$\mathcal{L}(x^\star, \alpha, \nu) \leq \mathcal{L}(x^\star, \alpha^\star, \nu^\star) \qquad \text{para toda } \alpha \geq 0 \text{ y } \nu. \tag{1.2.20}$$

De las desigualdades (1.2.19) y (1.2.20) se obtiene que $(x^\star, \alpha^\star, \nu^\star)$ satisface

$$\mathcal{L}(x^\star, \alpha, \nu) \leq \mathcal{L}(x^\star, \alpha^\star, \nu^\star) \leq \mathcal{L}(x, \alpha^\star, \nu^\star), \qquad \text{para toda } x,\ \alpha \geq 0,\ \nu.$$

Por lo tanto, $(x^\star, \alpha^\star, \nu^\star)$ es un *punto de silla* de la Lagrangiana.

Además, por factibilidad primal y complementariedad,

$$\mathcal{L}(x^\star, \alpha^\star, \nu^\star) = f(x^\star).$$

Usando la debilidad de la dualidad y la definición de la función dual, se obtiene:

$$d^* \leq p^* \leq f(x^\star) = \mathcal{L}(x^\star, \alpha^\star, \nu^\star) = \theta(\alpha^\star, \nu^\star) \leq d^*,$$

lo cual implica la igualdad y completa la demostración de la dualidad fuerte.

*(Recuperación primal).* Si $(\alpha^\star, \nu^\star)$ maximiza (1.2.18), cualquier $x^\star \in \operatorname{argmin}_x \mathcal{L}(x, \alpha^\star, \nu^\star)$ cumple $\theta(\alpha^\star, \nu^\star) = \mathcal{L}(x^\star, \alpha^\star, \nu^\star)$. Por dualidad fuerte (condición de Slater), este valor coincide con el óptimo primal, y $x^\star$ satisface KKT; en particular, es solución de (1.2.17). □

**Observaciones.**

**(i)** Si no se tiene un problema de optimización convexo o falla la condición de Slater en el problema de optimización, puede existir *brecha de dualidad* y el óptimo dual no permite recuperar una solución primal:

**Ejemplo 1.2.39.** Considérese el siguiente problema de optimización.

$$\begin{aligned} &\text{Minimizar} && x \\ &\text{sujeto a} && x^2 \leq 0. \end{aligned}$$

Las funciones $f(x) = x$ y $g(x) = x^2$ son funciones convexas. Nótese que no se cumple la condición de Slater pues se requiere que exista un $x \in \mathbb{R}$ tal que

$$x^2 < 0,$$

lo cual claramente no existe. La solución del primal es $x = 0$ dado que es el único que satisface la restricción, con lo que el valor óptimo primal es

$$p^\star = 0.$$

Se construye el Lagrangiano $\mathcal{L}(x, \alpha) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ como

$$\mathcal{L}(x, \alpha) = x + \alpha x^2,$$

con $\alpha \geq 0$. El dual es

$$\mathcal{W}(\alpha) = \text{ínf}\{x \in \mathbb{R} : x + \alpha x^2\}.$$

Se calcula el mínimo en $x$:

$$\frac{\partial \mathcal{L}}{\partial x} = 1 + 2\alpha x = 0,$$

es decir:

$$x^\star = -\frac{1}{2\alpha}.$$

Se sustituye en la función $\mathcal{W}$:

$$\mathcal{W}(\alpha) = -\frac{1}{2\alpha} + \alpha\frac{1}{4\alpha^2} = -\frac{1}{4\alpha}.$$

Como $\alpha > 0$:

$$q(\alpha) = -\frac{1}{4\alpha} \to 0 \text{ por la izquierda.}$$

Así tenemos que por más que se aumente el valor de $\alpha$ del valor óptimo $d^\star$ nunca será cero. Con esto se tiene que:

$$p^\star - d^\star > 0.$$

Es decir la brecha de dualidad no es nula.

**(ii)** En muchos problemas (por ejemplo, las máquinas de soporte vectorial, como es el presente caso), se resuelve el dual por eficiencia y luego se reconstruye $x^\star$ usando estacionariedad y complementariedad. Estas condiciones se utilizan para definir el problema de optimización dual de la SVM y, tras resolver ese problema de optimización dual, se obtienen los valores óptimos utilizados para la construcción del modelo SVM. Esto se muestra en el capítulo siguiente.

CAPÍTULO 2

# Máquina de Soporte Vectorial para clasificación

En este capítulo se presenta el análisis y formulación del modelo de máquina de soporte vectorial. Se inicia con una sección contextualizadora sobre los conceptos de aprendizaje automático, su funcionamiento y sus principales usos.

Una parte relevante de este capítulo es que se mostrará el teorema de Separabilidad mediante aumento de dimensionalidad, que afirma que los conjuntos que no son linealmente separables al mapearlos en un espacio de dimensión mayor (el cual se demuestra que existe) se vuelven linealmente separables. Este resultado no es común en la literatura sobre máquinas de soporte vectorial, salvo en trabajos muy especializados en kernels.

Se mostrará en detalle el modelo para clasificación en sus versiones de margen suave y rígido. Estas se construirán a partir de los fundamentos teóricos expuestos en el capítulo anterior.

## 2.1. Breve introducción al aprendizaje automático

El aprendizaje supervisado es una técnica de aprendizaje automático (*machine learning*) que utiliza dos conjuntos, el primero se compone de datos relacionados con las variables independientes y el segundo se constituye con las respectivas etiquetas (que conforman la variable objetivo o dependiente) y sirven para entrenar modelos de inteligencia artificial. Estos modelos aprenden las relaciones entre los dos conjuntos, lo que permite predecir salidas correctas basadas en datos nuevos. Esto es, los datos en el espacio de entrada $X$ (o espacio de características) están etiquetados, y la tarea de un algoritmo de aprendizaje supervisado es predecir las etiquetas correctas para datos desconocidos. Si $Y$ es un conjunto de etiquetas numéricas (continuas, discretas o binarias) la meta del aprendizaje supervisado es aproximar una función $F : X \to Y$ dado un *conjunto de entrenamiento* $\{(x_1, F(x_1)), \ldots (x_N, F(x_N))\}$, el cual es una colección de puntos en $X \times Y$ con los correspondientes valores de $F$. Los valores $F(x_i)$ se denominan *etiquetas de clase* cuando se trata de clasificación. Estos datos se pueden acomodar como las filas de una matriz de dimensión $N \times k + 1$, donde $k$ es el número de componentes de los vectores de entrenamiento, a esta matriz se le denomina *matriz de entrenamiento*. Una aproximación de la función $F$ puede ser determinada mediante un *modelo predictivo*, el cual es una función $f : X \to Y$ cuyos valores se denominan *predicciones*. Cuando un modelo $f$ depende de un conjunto de parámetros $\theta$, se denomina paramétrico, y se deben estimar los parámetros óptimos que proporcionan una buena aproximación a $F$.

Los modelos de aprendizaje supervisado son modelos predictivos que aproximan a $F$, se dividen principalmente en dos tipos: *clasificación* y *regresión*, aunque a su vez se dividen en paramétricos y no paramétricos.

**Definición 2.1.1** (**Problema de clasificación**). Dado un conjunto $D$ de $n$ vectores $d$ dimensionales y una etiqueta de clase con valores en $\{1, \ldots, A\}$, donde $A$ es el número de clases, el problema de clasificación se define como la búsqueda de un modelo $\mathcal{M}$ ajustado a los datos, el cual puede ser usado para predecir la etiqueta de clase de un nuevo registro $d$ - dimensional $x \notin D$.

La clasificación en machine learning utiliza los datos para organizarlos en categorías. Reconoce entidades específicas dentro del conjunto de datos e intenta establecer cómo deben etiquetarse o definirse esas entidades. Si el número de clases es $A = 2$, el problema se llama *problema de clasificación binaria*, en caso de que el número de clases sea mayor ($A > 2$), entonces el problema se llama *Problema de clasificación multiclase*. Los algoritmos de clasificación $\mathcal{M}$ más comunes son: *clasificadores lineales (caracterizados por crear un frontera de decisión lineal), máquinas de soporte vectorial (SVM), árboles de decisión y bosque aleatorio (utilizan un conjunto de reglas lógicas tipo "si-entonces"para asignar etiquetas a los datos) y k-vecinos más cercanos (fundamentados en la premisa de que los datos similares tienden a estar cerca unos de otros).*

La regresión una técnica que modela la relación entre una variable dependiente y una o más variables independientes. En los problemas de regresión, la salida es un valor continuo y los modelos predicen estos valores objetivo.

## 2.2. Caso linealmente separable

En el caso de clasificación, el objetivo es ajustar, a partir del conjunto de entrenamiento, un hiperplano que separe las dos clases, cuando estas son linealmente separables. A partir de dicho hiperplano se define una función de decisión que permite asignar una clase a nuevos elementos. La formulación del problema consiste, por tanto, en encontrar los parámetros óptimos de esta función de decisión, tarea que se aborda mediante herramientas de optimización convexa.

En el caso de regresión, la variable de respuesta es continua. En ambos contextos, clasificación y regresión, dicha variable recibe el nombre de variable objetivo y se denota por $y$. En esta investigación se considerará sólo la clasificación binaria, es decir, la variable objetivo toma sólo dos valores. Para dar paso a la formulación del modelo de máquina de soporte vectorial en este contexto, se considerarán las siguientes condiciones:

- Sea $S = \{x_1, x_2, \ldots, x_m\}$ un conjunto en el espacio vectorial $\mathbb{R}^n$ con $n \in \mathbb{N}$. Este conjunto de vectores será el conjunto de entrenamiento del modelo.

- A los vectores de entrenamiento se les asocia una variable objetivo, esta variable funciona como identificador de la clase a la que pertenecen. A esta entrada se le denomina etiqueta de la clase, la cual será denotada por $y_i$ donde $i \in \{1, ..., m\}$. Esta variable puede tomar valores $y_i \in \{-1, +1\}$, para toda $i \in \{1, ..., m\}$; esto, por conveniencia, permite clasificar solamente en dos clases.

- Ahora, sean los conjuntos

$$S_1 = \{x_j \mid x_j \in S \text{ y } y_j = 1\}, \qquad S_{-1} = \{x_k \mid x_k \in S \text{ y } y_k = -1\}. \tag{2.2.1}$$

  De ahora en adelante, se hará la suposición de que ambos conjuntos son no vacíos y que son linealmente separables.

Nótese que el espacio vectorial $\mathbb{R}^n$ es un espacio prehilbertiano con el producto interior ordinario, y que el conjunto de datos de entrenamiento se puede considerar a su vez como una matriz de tamaño $m \times n$, la cual claramente es conocida.

Se desea encontrar un hiperplano de tal manera que los datos de un lado (en el sentido de la Definición 1.2.24) se etiqueten como $y_i = +1$, mientras que los del otro lado se etiqueten como $y_i = -1$. Es deseable que el hiperplano orientado esté a la máxima distancia de las dos clases de puntos etiquetados situados a cada lado; es decir, la distancia entre el conjunto de puntos que conforman el hiperplano de separación y los conjuntos de puntos de cada uno de los lados sea máxima. Los puntos más cercanos a ambos lados son los que más influyen en la posición de este hiperplano de separación, por lo que se denominan vectores de soporte. Formalmente, los vectores de soporte son aquellos $x_1^*$ y $x_2^*$ que cumplen que

$$d(x_1^*, x_2^*) = \text{mín}(\{d(x_i, x_j) \mid y_i = 1 \text{ y } y_j = -1\}).$$

**Teorema 2.2.1.** Si los conjuntos $S_1$ y $S_{-1}$ son linealmente separables, entonces existe un hiperplano de separación.

*Demostración.* Dado que el conjunto de datos de entrenamiento $S$ es finito, los subconjuntos

$S_1$ y $S_{-1}$ son finitos y no son convexos. Si lo fueran, dados dos puntos dentro de $S_1$ o $S_{-1}$, podrían contener el segmento que los une; sin embargo, esto no es posible, pues la recta es un conjunto infinito de puntos.

Por otra parte, al ser linealmente separables $S_1$ y $S_{-1}$, sus envolventes convexas son disjuntas por el Teorema 1.2.31. Así también, las envolventes se encuentran en un espacio prehilbertiano, por lo que se satisfacen las hipótesis del Teorema 1.2.28; en consecuencia, existe un hiperplano de separación entre ambas envolventes convexas y, a su vez, separan a las clases por estar contenidas dentro de las envolventes. □

El hiperplano de separación derivado del teorema anterior, y como consecuencia del Teorema 1.2.28, es de la forma:

$$H = \{v \in \mathbb{R}^n : \langle w, v\rangle + b = 0\}, \tag{2.2.2}$$

donde $w$ es el vector de peso que determina la orientación del hiperplano y $b$ es el sesgo o desplazamiento del origen en el espacio de entrada. Luego de obtener los parámetros $w$ y $b$, la clasificación de una nueva entrada $z$, en el caso de la clasificación binaria, se realiza mediante la función de decisión $\phi$ dada por

$$\phi(z) = \text{signo}(\langle w, z\rangle + b).$$

Se puede observar que los datos de entrenamiento se clasifican correctamente mediante $\phi$, dado que

$$y_i(\langle w, x_i\rangle + b) \geq 0, \quad \text{para todo } i \in \{1, ..., m\},$$

debido a que $\langle w, x_i\rangle + b$ debe ser positivo cuando $y_i = +1$ y debe ser negativo cuando $y_i = -1$.

Los hiperplanos conformados por los puntos $x$ que satisfacen $\langle w, x\rangle + b = 1$ ó $\langle w, x\rangle + b = -1$ son llamados *hiperplanos canónicos* y la región entre estos hiperplanos canónicos es llamada la *banda de margen* o simplemente *margen* (véase la Figura 2.1).

Dados dos puntos $x_1$ y $x_2$ dentro de los hiperplanos canónicos, tales que $\langle w, x_1\rangle + b = 1$ y $\langle w, x_2\rangle + b = -1$, se realiza la diferencia entre estos, con lo que se tiene:

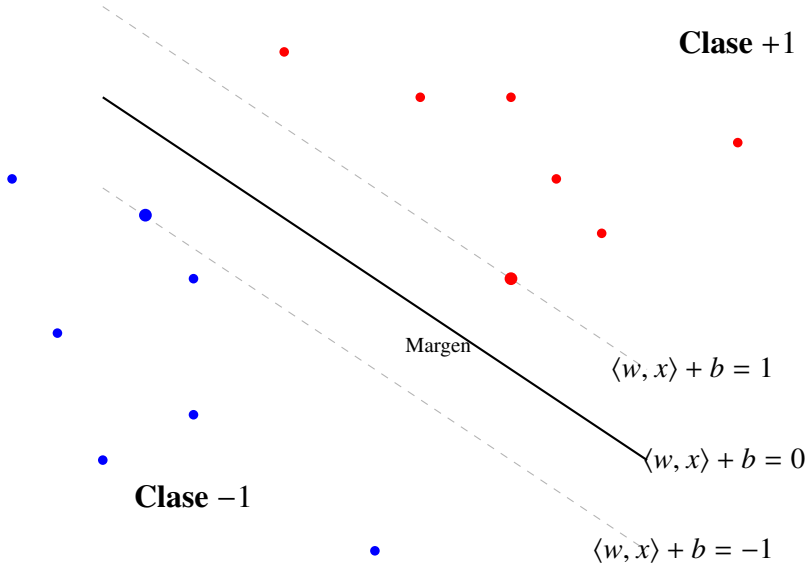


**Figura 2.1:** Hiperplano de separación (en negro), acompañado de los hiperplanos canónicos (en lineas punteadas).

$$\begin{aligned} \langle w, x_1 \rangle + b - (\langle w, x_2 \rangle + b) &= \text{1-(-1)}, \\ \langle w, x_1 \rangle - \langle w, x_2 \rangle &= 2, \\ \langle w, (x_1 - x_2) \rangle &= 2. \end{aligned}$$

Para el hiperplano separador $\langle w, x \rangle + b = 0$, el vector normal es

$$\frac{w}{||w||_2},$$

donde $||w||_2$ es la raíz cuadrada de $w^T w = \langle w, w \rangle$. Por lo tanto, la distancia entre los dos hiperplanos canónicos es igual a la proyección de $x_1 - x_2$ sobre el vector normal $w/||w||_2$, el cual está dado por

$$\left\langle (x_1 - x_2), \frac{w}{||w||_2} \right\rangle = \frac{1}{||w||_2} \langle (x_1 - x_2), w \rangle = \frac{2}{||w||_2}.$$

El margen se encuentra a la mitad de la distancia entre los dos hiperplanos canónicos, por ello, está dado por

$$\gamma = \frac{1}{||w||_2}.$$

El margen funcional de un punto $(x_i, y_i)$ se define como

$$\hat{\gamma}_i = y_i(\langle w, x_i \rangle + b), \qquad \hat{\gamma} = \min_i \hat{\gamma}_i.$$

Y el margen geométrico como

$$\gamma_i = \frac{y_i(\langle w, x_i \rangle + b)}{\|w\|_2}, \qquad \gamma = \min_i \gamma_i.$$

El problema de maximización del margen se reescribe como un problema de minimización, como lo establece la siguiente proposición.

**Proposición 2.2.2.** El problema de maximización del margen

$$\frac{1}{\|w\|_2},$$

sobre $w$ en el espacio de parámetros, es equivalente a minimizar la norma al cuadrado de $w$ dada por

$$\frac{1}{2}\|w\|_2^2.$$

*Demostración.* Sea el clasificador lineal definido por la función $\phi : \mathbb{R}^n \to \{1, -1\}$ dada por

$$\phi(x) = \text{signo}(\langle w, x \rangle + b).$$

Dado $\alpha > 0$, el par $(\alpha w, \alpha b)$ define el mismo hiperplano de decisión, pues se tiene que el hiperplano asociado a $(\alpha w, \alpha b)$ es

$$H_\alpha = \{x \in \mathbb{R}^n : \langle \alpha w, x \rangle + \alpha b = 0\},$$

pero

$$\begin{aligned}\langle \alpha w, x\rangle + \alpha b &= 0,\\ \alpha\langle w, x\rangle + \alpha b &= 0,\\ \alpha(\langle w, x\rangle + b) &= 0,\\ \langle w, x\rangle + b &= 0.\end{aligned}$$

Es decir $H_\alpha = \{x \in \mathbb{R}^n : \langle \alpha w, x\rangle + \alpha b = 0\} = \{x \in \mathbb{R}^n : \langle w, x\rangle + b = 0\}$, este último es el hiperplano de decisión.

Por otro lado, el margen geométrico es invariante bajo reescalamiento, es decir,

$$\hat{\gamma}_i \mapsto \alpha\hat{\gamma}_i, \qquad \gamma_i \mapsto \gamma_i.$$

Para fijar la escala, se impone la restricción

$$y_i(\langle w, x_i\rangle + b) \geq 1, \quad \text{para todo } i,$$

de modo que el mínimo margen funcional sea 1. Con esta normalización:

$$\gamma = \min_i \frac{y_i(\langle w, x_i\rangle + b)}{\|w\|_2} = \frac{1}{\|w\|_2}.$$

Ahora, el problema de maximizar el margen geométrico es equivalente a

$$\max \gamma = \max \frac{1}{\|w\|_2} \iff \min \|w\|_2.$$

Dado que $t \mapsto t$ y $t \mapsto \frac{1}{2}t^2$ son funciones estrictamente crecientes en $t \geq 0$, comparten el mismo conjunto de minimizadores:

$$\min \|w\|_2 \iff \min \tfrac{1}{2}\|w\|_2^2.$$

Bajo la normalización $y_i(\langle w, x_i\rangle + b) \geq 1$, maximizar el margen geométrico (equivalentemente $1/\|w\|_2$) es lo mismo que minimizar la norma de $w$, y por lo tanto, equivalente a minimizar

$$\frac{1}{2}\|w\|_2^2.$$

□

Lo que establece la Proposición 2.2.2 es que maximizar el margen es equivalente a

$$\text{mín}\ \frac{1}{2}\|w\|_2^2, \tag{2.2.3}$$

con respecto a los parámetros $w$ y $b$, sujeto a las restricciones

$$y_i(\langle w, x_i\rangle + b) \geq 1, \text{ para todo } i = 1, ..., m. \tag{2.2.4}$$

El conjunto de ecuaciones (2.2.3)-(2.2.4) es un problema de optimización en el cual se minimiza una *función objetivo*, sujeta a restricciones, formalmente se plantea el problema de optimización:

$$\begin{aligned} &\text{Minimizar} \quad \tfrac{1}{2}\|w\|^2 \\ &\text{sujeto a} \quad -y_i(\langle w, x_i\rangle + b) + 1 \leq 0, \quad i = 1, \ldots, m, \end{aligned} \tag{2.2.5}$$

Este es un problema de programación cuadrática, debido a que la función objetivo es una función cuadrática.

**Proposición 2.2.3.** Las funciones $l_i : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ dadas por

$$l_i(w, b) = -y_i(\langle w, x_i\rangle + b)$$

son funciones lineales con respecto a las variables $w$ y $b$. Las funciones $g_i : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ dadas por

$$g_i(w, b) = l_i(w, b) + 1$$

son afines.

*Demostración.* Sean $(w_1, b_1)$ y $(w_2, b_2)$ dos puntos de $\mathbb{R}^n \times \mathbb{R}$ y sean $\alpha$ y $\beta$ dos escalares en $\mathbb{R}$. Se toma

$$l_i(w, b) = -y_i(\langle w, x_i\rangle + b),$$

con lo que

$$\begin{aligned} l_i(\alpha(w_1, b_1) + \beta(w_2, b_2)) &= -y_i(\langle(\alpha w_1 + \beta w_2), x_i\rangle + (\alpha b_1 + \beta b_2)) \\ &= -y_i(\langle \alpha w_1, x_i\rangle + \langle \beta w_2, x_i\rangle + \alpha b_1 + \beta b_2) \\ &= -y_i((\langle \alpha w_1, x_i\rangle + \alpha b_1) + (\langle \beta w_2, x_i\rangle + \beta b_2)) \\ &= -y_i(\langle \alpha w_1, x_i\rangle + \alpha b_1) - y_i(\langle \beta w_2, x_i\rangle + \beta b_2) \\ &= \alpha(-y_i(\langle w_1, x_i\rangle + b_1)) + \beta(-y_i(\langle w_2, x_i\rangle + b_2)) \\ &= \alpha l_i(w_1, b_1) + \beta l_i(w_2, b_2). \end{aligned} \tag{2.2.6}$$

Con esto se comprueba que las funciones $l_i$ son lineales, para cada $i = 1, 2, \ldots, n$. Más aún

$$g_i(w, b) = l_i(w, b) + 1$$

son afines.

□

De ahora en adelante y por simplicidad, la función a minimizar $f : \mathbb{R}^n \to \mathbb{R}$ definida como $f(w) = \frac{1}{2}\|w\|_2^2$ será denotada simplemente como $\frac{1}{2}\|w\|_2^2$.

**Proposición 2.2.4.** La función

$$\frac{1}{2}\|w\|_2^2$$

es igual a la función

$$\frac{1}{2}w^T I w.$$

Con $I$ la matriz identidad sobre los reales de tamaño $n \times n$.

*Demostración.* Sea el vector $w^T = (w_1, w_2, \ldots, w_n)$. La expresión $\|w\|_2^2$ se reescribe como

$$\|w\|_2^2 = w_1^2 + w_2^2 + \cdots + w_n^2 = w^T \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} w = w^T I w.$$

Por lo tanto,

$$\frac{1}{2}\|w\|_2^2 = \frac{1}{2}w^T I w.$$

□

Formalmente, se obtiene el siguiente problema de optimización:

$$\begin{aligned} \text{Minimizar}_{w,b} \quad & \frac{1}{2}w^T I w \\ \text{sujeto a} \quad & -y_i(\langle w, x_i \rangle + b)) + 1 \leq 0. \end{aligned} \tag{2.2.7}$$

Dado que la matriz $I$ es una matriz semidefinida positiva, por el Teorema 1.2.12 se tiene que la función objetivo es convexa.

Visto como problema de optimización con restricciones, bajo la suposición de que se cumple la condición de Slater, esto es, existe $\bar{x}$ tal que $g_i(\bar{x}) < 0$ para todo $i$ y $h_j(\bar{x}) = 0$ para todo $j$, se puede encontrar una solución. Más aún, si se cumple la condición de Slater es equivalente a que los datos para entrenamiento sean linealmente separables, esto es:

**Proposición 2.2.5.** El problema primal de la SVM

$$\begin{aligned} &\text{Minimizar} \quad \tfrac{1}{2}\|w\|^2, \\ &\text{sujeto a} \quad -y_i(\langle w, x_i\rangle + b) + 1 \leq 0, \quad i = 1, \ldots, m, \end{aligned}$$

cumple con la condición de Slater si y sólo si los conjuntos $S_1$ y $S_{-1}$ son linealmente separables, es decir, los datos de entrenamiento son linealmente separables.

*Demostración.* Recuérdese que la condición de Slater requiere la existencia de un punto estrictamente factible, es decir, un par $(\bar{w}, \bar{b})$ tal que

$$y_i(\langle \bar{w}, x_i\rangle + \bar{b}) > 1, \quad \text{para todo } i = 1, \ldots, m.$$

**(i) Necesidad.** Si los datos no son linealmente separables, entonces no existe $(w, b)$ tal que $y_i(\langle w, x_i\rangle + b) \geq 1$ para todo $i$, y por lo tanto el conjunto factible es vacío. En consecuencia, no puede existir un punto estrictamente factible y la condición de Slater no se cumple.

**(ii) Suficiencia.** Supóngase que los datos son linealmente separables. Entonces existe un hiperplano $(w_0, b_0)$ y un margen $\gamma > 0$ tal que

$$y_i(\langle w_0, x_i\rangle + b_0) \geq \gamma > 0, \quad \text{para todo } i.$$

Defínase para $\varepsilon > 0$:

$$\bar{w} = \left(\tfrac{1}{\gamma} + \varepsilon\right) w_0, \quad \bar{b} = \left(\tfrac{1}{\gamma} + \varepsilon\right) b_0.$$

Así, para todo $i$:

$$y_i(\langle \bar{w}, x_i\rangle + \bar{b}) = \left(\tfrac{1}{\gamma} + \varepsilon\right) y_i(\langle w_0, x_i\rangle + b_0) \geq \left(\tfrac{1}{\gamma} + \varepsilon\right)\gamma = 1 + \varepsilon\gamma > 1.$$

Por lo tanto, $(\bar{w}, \bar{b})$ es estrictamente factible y la condición de Slater se cumple. □

La Proposición 2.2.5 da paso a que el problema de optimización sea resoluble sólo si los datos son linealmente separables; por esta razón, a este problema se le conoce como “de margen

duro". Un ejemplo se observa en la Figura 2.2.

Bajo la suposición de que los datos son linealmente separables y, como consecuencia del Teorema 1.2.38, se formula la función de Lagrange:

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}\langle w, w\rangle - \sum_{i=1}^{m} \alpha_i (y_i(\langle w, x_i\rangle + b) + 1), \tag{2.2.8}$$

donde $\alpha_i$ son multiplicadores de Lagrange, y por lo tanto, $\alpha_i \geq 0$. El cual consiste en la suma de la función objetivo y las $m$ restricciones multiplicadas por su respectivo multiplicador de Lagrange.

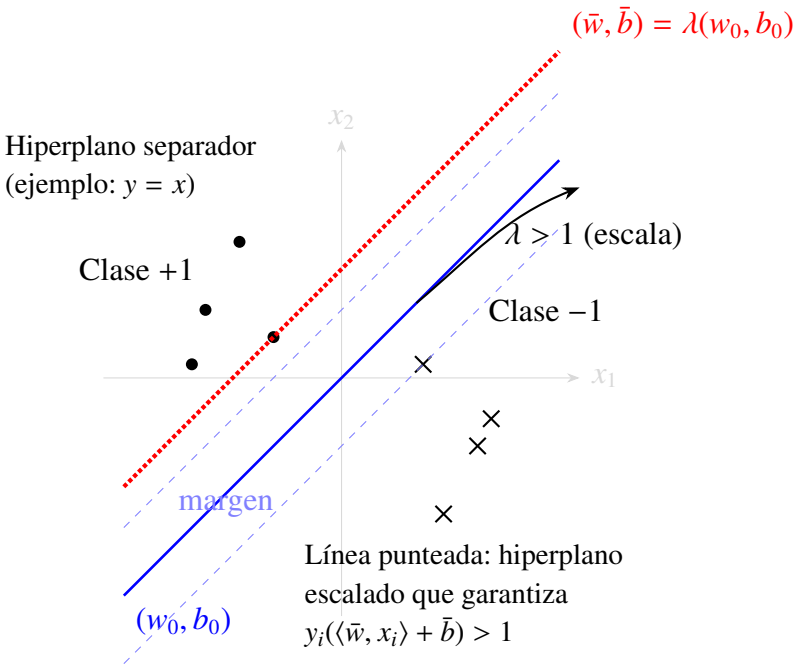


**Figura 2.2:** Esquema: hiperplano separador $(w_0, b_0)$, en azul, y su versión escalada $(\bar{w}, \bar{b}) = \lambda(w_0, b_0)$, punteada en rojo, que incrementa el margen hasta superar 1.

Luego, se desea hallar el ínfimo del Lagrangiano, el cual se encuentra de forma explícita, como se demuestra en la siguiente proposición.

**Proposición 2.2.6.** Si los datos de entrenamiento son linealmente separables, entonces la solución del problema

$$\text{Minimizar } \mathcal{L}(w, b, \alpha) = \frac{1}{2}\langle w, w\rangle - \sum_{i=1}^{m} \alpha_i(y_i(\langle w, x_i\rangle + b) + 1), \tag{2.2.9}$$

$$\text{sujeto a } \alpha_i \geq 0, \tag{2.2.10}$$

son aquellos valores $\alpha$ tales que $\sum_{i=1}^{m} \alpha_i y_i = 0$. Además, el valor del parámetro $w$ del problema primal es

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i.$$

*Demostración.* Al calcular las derivadas parciales con respecto a $b$ y $w$ e igualar a cero, se tiene

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial b} &= \frac{\partial}{\partial b}\left(\frac{1}{2}\langle w, w\rangle - \sum_{i=1}^{m} \alpha_i(y_i(\langle w, x_i\rangle + b) + 1)\right) \\
&= -\frac{\partial}{\partial b}(\sum_{i=1}^{m} \alpha_i(y_i(\langle w, x_i\rangle + b) + 1)) \\
&= -\sum_{i=1}^{m} \frac{\partial}{\partial b}(\alpha_i(y_i(\langle w, x_i\rangle + b) + 1) \\
&= -\sum_{i=1}^{m} \alpha_i y_i.
\end{aligned}$$

Por lo tanto,

$$\frac{\partial \mathcal{L}}{\partial b} = 0,$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0.$$

Se calcula la derivada parcial de $\mathcal{L}$ con respecto a $w$:

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{\partial}{\partial w}(\frac{1}{2}\langle w, w\rangle - \sum_{i=1}^{m} \alpha_i(y_i(\langle w, x_i\rangle + b) + 1))$$

$$
\begin{aligned}
&= \frac{\partial}{\partial w}(\frac{1}{2}\langle w, w\rangle) - \frac{\partial}{\partial w}(\sum_{i=1}^{m} \alpha_i(y_i(\langle w, x_i\rangle + b) + 1)) \\
&= w - \sum_{i=1}^{m} \frac{\partial}{\partial w}(\alpha_i(y_i(\langle w, x_i\rangle + b) + 1)) \\
&= w - \sum_{i=1}^{m} \alpha_i y_i \frac{\partial}{\partial w}(\langle w, x_i\rangle + b) \\
&= w - \sum_{i=1}^{m} \alpha_i y_i x_i.
\end{aligned}
$$

En consecuencia

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w} &= 0, \\
w - \sum_{i=1}^{m} \alpha_i y_i x_i &= 0. \qquad (2.2.11)
\end{aligned}
$$

Más aún, al despejar $w$ de la ecuación (2.2.11) se obtiene

$$
w = \sum_{i=1}^{m} \alpha_i y_i x_i. \qquad (2.2.12)
$$

□

Al sustituir $w$ en la ecuación (2.2.8), se obtiene la formulación de la función $\mathcal{W} : \mathbb{R}^m \to \mathbb{R}$, la cual se utiliza para la formulación dual del problema:

$$
\mathcal{W}(\alpha) = \frac{1}{2}\left\langle \sum_{i=1}^{m} \alpha_i y_i x_i \, , \sum_{j=1}^{m} \alpha_j y_j x_j \right\rangle - \sum_{j=1}^{m} \alpha_i \left( y_i \left( \left\langle \sum_{j=1}^{m} \alpha_j y_j x_j \, , x_i \right\rangle + b \right) + 1 \right)
$$

$$
\begin{aligned}
&= \frac{1}{2}\sum_{i=1}^{m}\alpha_i y_i \sum_{j=1}^{m}\alpha_j y_j \langle x_i\ ,\ x_j\rangle - \sum_{i=1}^{m}\alpha_i y_i \sum_{j=1}^{m}\alpha_j y_j \langle x_j, x_i\rangle - \sum_{i=1}^{m}\alpha_i y_i b + \sum_{i=1}^{m}\alpha_i \\
&= \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i y_i \alpha_j y_j \langle x_i\ , x_j\rangle - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i y_i \alpha_j y_j \langle x_j\ , x_i\rangle - b\sum_{i=1}^{m}\alpha_i y_i + \sum_{i=1}^{m}\alpha_i \\
&= \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i \alpha_j y_i y_j \langle x_i\ , x_j\rangle.
\end{aligned}
$$

En consecuencia,

$$
\mathcal{W}(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i \alpha_j y_i y_j \langle x_i\ , x_j\rangle. \tag{2.2.13}
$$

EL problema dual, de acuerdo con el Teorema 1.2.38, consiste en maximizar la expresión $\mathcal{W}(\alpha)$ con respecto a las variables $\alpha_i$, sujeto a las restricciones:

$$
\alpha_i \geq 0, \qquad \textstyle\sum_{i=1}^{m}\alpha_i y_i = 0.
$$

Formalmente, se obtiene el siguiente problema de optimización:

$$
\text{Maximizar} \qquad \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i \alpha_j y_i y_j \langle x_i\ , x_j\rangle \tag{2.2.14}
$$

$$
\text{sujeto a} \qquad \sum_{i=1}^{m}\alpha_i y_i = 0, \tag{2.2.15}
$$

$$
\alpha_i \geq 0, \quad \text{para toda } i = 1, 2, \ldots, m. \tag{2.2.16}
$$

El problema de optimización conformado por las ecuaciones (2.2.14)-(2.2.16) es un problema de optimización cuadrático, el cual se resuelve por medio de un método que se describe a continuación, llamado método del gradiente ascendente; se hace uso de este algoritmo debido a que la función $\mathcal{W}$ es cóncava; como consecuencia, converge a un máximo global.

**Proposición 2.2.7.** La función $\mathcal{W} : \mathbb{R}^m \to \mathbb{R}$ dada por

$$\mathcal{W}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i \, , x_j \rangle.$$

es cóncava.

*Demostración.* Para demostrar que la función $\mathcal{W}$ es cóncava, se calcula la matriz Hessiana y se comprobará que es semidefinida negativa. La matriz Hessiana, por definición, tiene la siguiente forma

$$H(\mathcal{W}) = \begin{pmatrix} \frac{\partial^2 \mathcal{W}}{\partial x_1^2} & \frac{\partial^2 \mathcal{W}}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 \mathcal{W}}{\partial x_1 \partial x_m} \\ \frac{\partial^2 \mathcal{W}}{\partial x_2 \partial x_1} & \frac{\partial^2 \mathcal{W}}{\partial^2 x_2} & \cdots & \frac{\partial^2 \mathcal{W}}{\partial x_2 \partial x_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \mathcal{W}}{\partial x_m \partial x_1} & \frac{\partial^2 \mathcal{W}}{\partial x_m \partial x_2} & \cdots & \frac{\partial^2 \mathcal{W}}{\partial^2 x_m} \end{pmatrix}. \tag{2.2.17}$$

Las primeras derivadas parciales de la función $\mathcal{W}$ son

$$\begin{aligned} \frac{\partial \mathcal{W}(\alpha)}{\partial \alpha_k} &= \frac{\partial}{\partial \alpha_k} \left( \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i \, , x_j \rangle \right) \\ &= \frac{\partial}{\partial \alpha_k} \Big( \alpha_1 + ... + \alpha_k + ... + \alpha_m - \frac{1}{2} \Big( \alpha_1^2 y_1^2 \langle x_1, x_1 \rangle + ... + \alpha_1 \alpha_k y_1 y_k \langle x_1, x_k \rangle + \\ &\quad ... + \alpha_1 \alpha_m y_1 y_m \langle x_1, x_m \rangle + ... + \alpha_k \alpha_1 y_k y_1 \langle x_k, x_1 \rangle + ... + \alpha_k^2 y_k^2 \langle x_k, x_k \rangle + \\ &\quad ... + \alpha_k \alpha_m y_k y_m \langle x_k, x_m \rangle + ... + \alpha_m \alpha_1 y_m y_1 \langle x_m, x_1 \rangle + ... \\ &\quad ... + \alpha_m^2 y_m^2 \langle x_m, x_m \rangle \Big) \Big) \\ &= 1 - \frac{1}{2} \Big( \alpha_1 y_1 y_k \langle x_1, x_k \rangle + \alpha_2 y_2 y_k \langle x_2, x_k \rangle + ... + \alpha_m y_m y_k \langle x_m, x_k \rangle + \\ &\quad + \alpha_m \alpha_k y_m y_k \langle x_m, x_k \rangle + ... + \alpha_1 y_k y_1 \langle x_k, x_1 \rangle + \alpha_2 y_k y_2 \langle x_k, x_2 \rangle + ... \\ &\quad + \alpha_m y_k y_m \langle x_k, x_m \rangle \Big) \\ &= 1 - \frac{2}{2} \Big( \alpha_1 y_1 y_k \langle x_1, x_k \rangle + \alpha_2 y_2 y_k \langle x_2, x_k \rangle + ... + \alpha_m y_m y_k \langle x_m, x_k \rangle \Big). \end{aligned}$$

De lo cual resulta

$$\frac{\partial \mathcal{W}(\alpha)}{\partial \alpha_k} = 1 - y_k \sum_{i=1}^{m} \alpha_i y_i \langle x_i, x_k \rangle. \tag{2.2.18}$$

Para las componentes de la diagonal, se deriva nuevamente respecto a la misma variable:

$$\begin{aligned}
\frac{\partial^2 \mathcal{W}}{\partial \alpha_k^2} &= \frac{\partial}{\partial \alpha_k}\left(1 - y_k \sum_{i=1}^{m} \alpha_i y_i \langle x_i, x_k \rangle\right) \\
&= \frac{\partial}{\partial \alpha_k}\Big(1 - y_k(\alpha_1 y_1 \langle x_1, x_k \rangle + ... + \alpha_m y_m)\langle x_m, x_k \rangle)\Big) \\
&= -y_k^2 \langle x_k, x_k \rangle \\
&= -\langle x_k, x_k \rangle.
\end{aligned}$$

El resto de las entradas de la matriz se pueden calcular al derivar la ecuación (2.2.18) con respecto a dos distintas variables:

$$\begin{aligned}
\frac{\partial^2 \mathcal{W}}{\partial \alpha_k \partial \alpha_h} &= \frac{\partial}{\partial \alpha_h}\left(1 - y_k \sum_{i=1}^{m} \alpha_i y_i \langle x_i, x_k \rangle\right) \\
&= \frac{\partial}{\partial \alpha_h}\Big(1 - y_k(\alpha_1 y_1 \langle x_1, x_k \rangle + ... + \alpha_m y_m)\langle x_m, x_k \rangle)\Big) \\
&= -y_k y_h \langle x_h, x_k \rangle \\
&= -y_k y_h \langle x_k, x_h \rangle.
\end{aligned}$$

De donde se sigue que la matriz Hessiana de la función $\mathcal{W}$ es

$$H(\mathcal{W}) = \begin{pmatrix}
-\langle x_1, x_1 \rangle & -y_1 y_2 \langle x_2, x_1 \rangle & \dots & -y_1 y_m \langle x_1, x_m \rangle \\
-y_1 y_2 \langle x_1, x_2 \rangle & -\langle x_2, x_2 \rangle & \dots & -y_2 y_m \langle x_2, x_m \rangle \\
\vdots & \vdots & \ddots & \vdots \\
-y_1 y_m \langle x_m, x_1 \rangle & -y_2 y_m \langle x_m, x_2 \rangle & \dots & -\langle x_m, x_m \rangle
\end{pmatrix}. \tag{2.2.19}$$

Ahora, si se toma

$$Q = \begin{pmatrix} \langle x_1, x_1 \rangle & y_1 y_2 \langle x_1, x_2 \rangle & \dots & y_1 y_m \langle x_1, x_m \rangle \\ y_1 y_2 \langle x_2, x_1 \rangle & \langle x_2, x_2 \rangle & \dots & y_2 y_m \langle x_2, x_m \rangle \\ \vdots & \vdots & \ddots & \vdots \\ y_1 y_m \langle x_m, x_1 \rangle & y_2 y_m \langle x_m, x_2 \rangle & \dots & \langle x_m, x_m \rangle \end{pmatrix}, \tag{2.2.20}$$

la matriz Hessiana queda como

$$H(\mathcal{W}) = -Q. \tag{2.2.21}$$

La matriz $Q$ es la matriz de Gram de los vectores

$$v_i = y_i x_i \in \mathbb{R}^n,$$

debido a que

$$Q_{ij} = y_i y_j \langle x_i, x_j \rangle = \langle v_i, v_j \rangle.$$

La matriz $Q$ es semidefinida positiva, en efecto, sea $z \in \mathbb{R}^m$, se tiene

$$\begin{aligned} z^t Q z &= \sum_{i=1}^{m} \sum_{j=1}^{m} z_i z_j \langle v_i, v_j \rangle \\ &= \left\langle \sum_{i=1}^{m} z_i v_i, \sum_{j=1}^{m} z_j v_j \right\rangle \\ &= \left\| \sum_{i=1}^{m} z_i v_i \right\|^2 \geq 0. \end{aligned}$$

Como $Q$ es semidefinida positiva, entonces la matriz $-Q$ es semidefinida negativa. En consecuencia, la función $\mathcal{W}$ es cóncava. □

Para hallar el óptimo, se debe calcular en un primer paso el gradiente de $\mathcal{W}$; para hacerlo, se calculan las derivadas parciales de la función

$$\mathcal{W}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i \, , x_j \rangle$$

con respecto a cada una de las componentes $\alpha_k$ del vector $\alpha$, las cuales se calcularon en la demostración de la Proposición 2.2.7, por lo que el gradiente de la función $\mathcal{W}$ es

$$\nabla \mathcal{W} = \left( \frac{\partial W}{\partial \alpha_1}, \frac{\partial W}{\partial \alpha_2}, \cdots, \frac{\partial W}{\partial \alpha_m} \right), \tag{2.2.22}$$

donde

$$\frac{\partial \mathcal{W}}{\partial \alpha_k} = 1 - y_k \sum_{i=1}^{m} \alpha_i y_i \langle x_i, x_k \rangle, \qquad \text{para cada } k = 1, 2, ..., m. \tag{2.2.23}$$

El método del gradiente ascendente calcula de manera iterativa el siguiente punto por medio del gradiente del punto actual; este punto se escala con una razón de aprendizaje $\gamma \in \mathbb{R}^+$; es decir, el gradiente ascendente se calcula de la siguiente manera.

$$(\alpha_1, \, ..., \, \alpha_m) = (\alpha_1, \, ..., \alpha_m) + \gamma \nabla \mathcal{W}(\alpha_1, \, ..., \alpha_m). \tag{2.2.24}$$

Si el valor de paso $\gamma$ es relativamente grande, el algoritmo puede oscilar y no converger; en caso de ser muy pequeño, converge, pero muy lentamente. La solución inicial puede elegirse como el vector de ceros, el cual también es una solución factible para $\alpha$. Un problema que puede presentarse es que las restricciones

$$\alpha_i \geq 0 \qquad \text{y} \qquad \sum_{i=1}^{m} \alpha_i y_i = 0, \tag{2.2.25}$$

pueden ser violadas después de actualizar $(\alpha_1, \, ..., \alpha_m)$ por $(\alpha_1, \, ..., \, \alpha_m) + \gamma \nabla \mathcal{W}(\alpha_1, \, ..., \alpha_m)$, por lo tanto, cualquier valor negativo $\alpha_i$ se reemplaza por cero.

**Proposición 2.2.8.** Sean $\alpha_i$, $i = 1, \ldots, m$, los multiplicadores de Lagrange obtenidos por el método de descenso de gradiente y $P = \{i : \alpha_i > 0\}$. Si los datos de entrenamiento son linealmente separables, entonces el valor del parámetro $b$ de la ecuación (2.2.2) es el promedio de los $b_r$

$$b_r = \frac{1}{y_r} - \left\langle \sum_{i=1}^{m} \alpha_i y_i x_i \ , x_r \right\rangle, \tag{2.2.26}$$

donde $r \in P$. Por lo tanto,

$$b^* = \frac{\sum_{r \in P} b_r}{|P|}. \tag{2.2.27}$$

*Demostración.* Considérense las restricciones de la formulación primal del problema de optimización dado en (2.2.7). Luego, se toman solamente los datos de entrenamiento que tengan asociado un operador de Lagrange $\alpha_r > 0$. Así también considérese el valor de $w$ resultado de la Proposición 2.2.6. Con estos valores se despeja a $b$:

$$\begin{aligned}
y_r(\langle w, x_r \rangle + b_r) &= 1, && \text{para todo } r : \alpha_r > 0,\\
y_r\left(\left\langle \sum_{i=1}^{m} \alpha_i y_i x_i, x_r \right\rangle + b_r\right) &= 1, && \text{para todo } r : \alpha_r > 0,\\
\left\langle \sum_{i=1}^{m} \alpha_i y_i x_i \ , x_r \right\rangle + b_r &= \frac{1}{y_r}, && \text{para todo } r : \alpha_r > 0,\\
b_r &= \frac{1}{y_r} - \left\langle \sum_{i=1}^{m} \alpha_i y_i x_i \ , x_r \right\rangle, && \text{para todo } r : \alpha_r > 0.
\end{aligned}$$

Se toma el promedio de todos los $b_r$, es decir,

$$b^* = \frac{\sum_{r \in P} b_r}{|P|}.$$

Con lo que se demuestra la proposición. □

Finalmente, si los datos de entrenamiento son linealmente separables, el resultado deseado es la función de clasificación, la cual, para una nueva entrada $z \in \mathbb{R}^n$, la clase está determinada por la función de decisión que se define como

$$\phi(z) = \text{signo}(\langle w \, , z \rangle + b)$$

$$= \text{signo}\left(\left\langle \sum_{i=1}^{m} \alpha_i y_i x_i \ , z \right\rangle + b^*\right). \tag{2.2.28}$$

A continuación se muestra un ejemplo sencillo que ilustra el funcionamiento de una máquina de soporte vectorial (SVM) en un problema de clasificación binaria linealmente separable. Considérese el siguiente conjunto de puntos en $\mathbb{R}^2$:

$$\text{Clase positiva } (y = +1): \ (2,3),\ (3,3),\ (3,4),$$

$$\text{Clase negativa } (y = -1): \ (0,1),\ (1,1),\ (1,2).$$

La Figura 2.3 muestra la ubicación de los puntos en el plano.

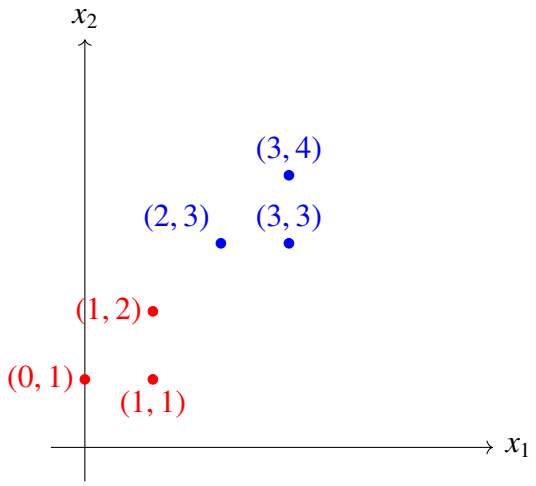


**Figura 2.3:** Conjunto de datos de entrenamiento.

La función de decisión está dada por

$$\phi(x) = \text{signo}(\langle w, x\rangle + b).$$

El problema de optimización está dado por

$$\begin{aligned} &\text{Minimizar} \quad \tfrac{1}{2}\|w\|^2 \\ &\text{sujeto a} \quad -y_i(\langle w, x_i\rangle + b) + 1 \le 0, \quad i = 1, \ldots, 6. \end{aligned} \tag{2.2.29}$$

El problema de optimización dual es de la forma:

$$\text{Maximizar} \qquad \mathcal{W}(\alpha) = \sum_{i=1}^{6} \alpha_i - \frac{1}{2} \sum_{i=1}^{6} \sum_{j=1}^{6} \alpha_i \alpha_j y_i y_j \langle x_i \, , x_j \rangle \tag{2.2.30}$$

$$\text{sujeto a} \qquad \sum_{i=1}^{6} \alpha_i y_i = 0, \tag{2.2.31}$$

$$\alpha_i \geq 0, \quad \text{para toda } i = 1, 2, \ldots, 6. \tag{2.2.32}$$

El gradiente de $\mathcal{W}$ es

$$\nabla \mathcal{W} = \left( \frac{\partial W}{\partial \alpha_1}, \frac{\partial W}{\partial \alpha_2}, \cdots, \frac{\partial W}{\partial \alpha_6} \right),$$

donde

$$\frac{\partial \mathcal{W}}{\partial \alpha_k} = 1 - y_k \sum_{i=1}^{6} \alpha_i y_i \langle x_i, x_k \rangle, \qquad \text{para cada } k = 1, 2, ..., 6.$$

Si se enumeran los datos como

$$x_1 = (2, 3), \qquad x_2 = (0, 1),$$
$$x_3 = (3, 3), \qquad x_4 = (1, 1),$$
$$x_5 = (3, 4), \qquad x_6 = (1, 2),$$

las derivadas parciales $\frac{\partial \mathcal{W}}{\partial \alpha_k}$ para $k = 1, ..., 6$ quedan de la forma:

$$\begin{aligned} \frac{\partial \mathcal{W}}{\partial \alpha_1} &= 1 - 1(\alpha_1 \cdot 1\langle (2,3),(2,3) \rangle + \alpha_2 \cdot (-1)\langle (0,1),(2,3) \rangle + \alpha_3 \cdot (1)\langle (3,3),(2,3) \rangle \\ &\qquad + \alpha_4 \cdot (-1)\langle (1,1),(2,3) \rangle + \alpha_5 \cdot (1)\langle (3,4),(2,3) \rangle + \alpha_6 \cdot (-1)\langle (1,2),(2,3) \rangle) \\ &= 1 - (13\alpha_1 - 3\alpha_2 + 15\alpha_3 - 5\alpha_4 + 18\alpha_5 - 8\alpha_6). \end{aligned}$$

$$\frac{\partial \mathcal{W}}{\partial \alpha_2} = 1 - (-1)(\alpha_1 \cdot 1\langle (2,3),(0,1) \rangle + \alpha_2 \cdot (-1)\langle (0,1),(0,1) \rangle + \alpha_3 \cdot (1)\langle (3,3),(0,1) \rangle$$

$$+ \alpha_4 \cdot (-1)\langle(1,1),(0,1)\rangle + \alpha_5 \cdot (1)\langle(3,4),(0,1)\rangle + \alpha_6 \cdot (-1)\langle(1,2),(0,1)\rangle)$$
$$= 1 + (3\alpha_1 - \alpha_2 + 3\alpha_3 - \alpha_4 + 4\alpha_5 - 2\alpha_6).$$

$$\frac{\partial \mathcal{W}}{\partial \alpha_3} = 1 - (1)(\alpha_1 \cdot 1\langle(2,3),(3,3)\rangle + \alpha_2 \cdot (-1)\langle(0,1),(3,3)\rangle + \alpha_3 \cdot (1)\langle(3,3),(3,3)\rangle$$
$$+ \alpha_4 \cdot (-1)\langle(1,1),(3,3)\rangle + \alpha_5 \cdot (1)\langle(3,4),(3,3)\rangle + \alpha_6 \cdot (-1)\langle(1,2),(3,3)\rangle)$$
$$= 1 - (15\alpha_1 - 3\alpha_2 + 18\alpha_3 - 6\alpha_4 + 21\alpha_5 - 9\alpha_6).$$

$$\frac{\partial \mathcal{W}}{\partial \alpha_4} = 1 - (-1)(\alpha_1 \cdot 1\langle(2,3),(1,1)\rangle + \alpha_2 \cdot (-1)\langle(0,1),(1,1)\rangle + \alpha_3 \cdot (1)\langle(3,3),(1,1)\rangle$$
$$+ \alpha_4 \cdot (-1)\langle(1,1),(1,1)\rangle + \alpha_5 \cdot (1)\langle(3,4),(1,1)\rangle + \alpha_6 \cdot (-1)\langle(1,2),(1,1)\rangle)$$
$$= 1 + (5\alpha_1 - \alpha_2 + 6\alpha_3 - 2\alpha_4 + 7\alpha_5 - 3\alpha_6).$$

$$\frac{\partial \mathcal{W}}{\partial \alpha_5} = 1 - (1)(\alpha_1 \cdot 1\langle(2,3),(3,4)\rangle + \alpha_2 \cdot (-1)\langle(0,1),(3,4)\rangle + \alpha_3 \cdot (1)\langle(3,3),(3,4)\rangle$$
$$+ \alpha_4 \cdot (-1)\langle(1,1),(3,4)\rangle + \alpha_5 \cdot (1)\langle(3,4),(3,4)\rangle + \alpha_6 \cdot (-1)\langle(1,2),(3,4)\rangle)$$
$$= 1 - (18\alpha_1 - 4\alpha_2 + 21\alpha_3 - 7\alpha_4 + 25\alpha_5 - 11\alpha_6).$$

$$\frac{\partial \mathcal{W}}{\partial \alpha_6} = 1 - (-1)(\alpha_1 \cdot 1\langle(2,3),(1,2)\rangle + \alpha_2 \cdot (-1)\langle(0,1),(1,2)\rangle + \alpha_3 \cdot (1)\langle(3,3),(1,2)\rangle$$
$$+ \alpha_4 \cdot (-1)\langle(1,1),(1,2)\rangle + \alpha_5 \cdot (1)\langle(3,4),(1,2)\rangle + \alpha_6 \cdot (-1)\langle(1,2),(1,2)\rangle)$$
$$= 1 + (8\alpha_1 - 2\alpha_2 + 9\alpha_3 - 3\alpha_4 + 11\alpha_5 - 5\alpha_6).$$

El método de gradiente ascendente inicia con $(\alpha_1^0, ..., \alpha_6^0) = (0, ..., 0)$ y con $\gamma = 1$ se obtiene el vector solución:

$$\alpha^\star = (1, 0, 0, 0, 0, 1). \tag{2.2.33}$$

Del vector solución (2.2.33), se tiene que los únicos valores que son mayores que 0 son $\alpha_1 = \alpha_6 = 1$, por lo que se utiliza la Proposición 2.2.6 para hallar el valor de $w$:

$$\begin{aligned} w &= \sum_{i=1}^{6} \alpha_i y_i x_i \\ &= (\alpha_1 y_1 x_1 + \alpha_6 y_6 x_6) \\ &= 1(1)(2,3) + 1(-1)(1,2) \\ &= (1,1). \end{aligned}$$

Para hallar el valor del parámetro $b$ del hiperplano de separación se utiliza la Proposición 2.2.8. Nótese que los valores de los multiplicadores de Lagrange son los $\alpha_1 = \alpha_6 = 1 > 0$, por lo que se calculan los $b_r$ respectivos:

$$\begin{aligned} b_1 &= \frac{1}{1} - \left\langle \sum_{i=1}^{6} \alpha_i y_i x_i, x_1 \right\rangle \\ &= 1 - \langle (1,1), (2,3) \rangle \\ &= 1 - 5 \\ &= -4, \end{aligned}$$

$$\begin{aligned} b_6 &= \frac{1}{-1} - \left\langle \sum_{i=1}^{6} \alpha_i y_i x_i, x_6 \right\rangle \\ &= -1 - \langle (1,1), (1,2) \rangle \\ &= -1 - 3 \\ &= -4. \end{aligned}$$

Ahora se calcula el promedio de estos dos elementos.

$$b^* = \frac{(-4) + (-4)}{2} = -4. \tag{2.2.34}$$

Por lo tanto, el hiperplano de separación es $H = \{x \in \mathbb{R}^2 : \langle (1,1), x\rangle - 4 = 0\}$.

Para verificar que la SVM funciona correctamente, considere la evaluación para la clase positiva:

$$\langle (1,1),(2,3)\rangle - 4 = 2+3-4 = 1 > 0, \qquad \langle (1,1),(3,3)\rangle - 4 = 3+3-4 = 2 > 0,$$

$$\langle (1,1),(3,4)\rangle - 4 = 3+4-4 = 3 > 0.$$

Para la clase negativa:

$$\langle (1,1),(0,1)\rangle - 4 = 0+1-4 = -3 < 0, \qquad \langle (1,1),(1,1)\rangle - 4 = 1+1-4 = -2 < 0,$$

$$\langle (1,1),(1,2)\rangle - 4 = 1+2-4 = -1 < 0.$$

Con lo que se observa que el hiperplano separa correctamente ambas clases como se muestra en la Figura 2.4.

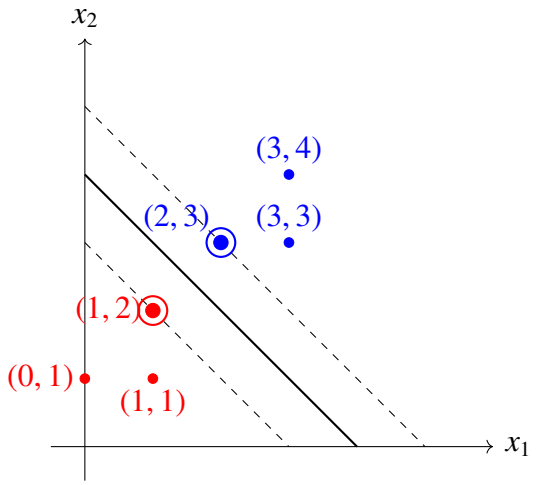


**Figura 2.4:** Hiperplano óptimo y puntos de soporte en una SVM separable.

Los puntos de soporte son aquellos que satisfacen $|\langle w, x\rangle + b| = 1$ y que determinan el margen:

$$x_1 + x_2 - 4 = 1,$$

$$x_1 + x_2 - 4 = -1.$$

En este ejemplo, los puntos de soporte son

$$(2, 3) \quad \text{y} \quad (1, 2).$$

Por lo que la función de decisión para este conjunto de datos es

$$\phi(z) = \text{signo}(z_1 + z_2 - 4), \tag{2.2.35}$$

para una nueva entrada $z = (z_1, z_2) \in \mathbb{R}^2$.

Para mostrar la clasificación se toman nuevos vectores de entrada en $\mathbb{R}^2$:

$$z^{(1)} = (4, 3)$$

y

$$z^{(2)} = (1, 0).$$

Y las respectivas etiquetas se calculan como

$$\begin{aligned}
\phi(z^{(1)}) &= \text{signo}(\langle (1, 1), (4, 3) \rangle - 4) \\
&= \text{signo}(7 - 4) \\
&= \text{signo}(3) \\
&= +1,
\end{aligned}$$

$$\begin{aligned}
\phi(z^{(2)}) &= \text{signo}(\langle (1, 1), (1, 0) \rangle - 4) \\
&= \text{signo}(1 - 4) \\
&= \text{signo}(-3) \\
&= -1.
\end{aligned}$$

Por lo tanto, $z^{(1)}$ pertenece a la clase $+1$, mientras que $z^{(2)}$ pertenece a la clase $-1$, esto se puede observar en la Figura 2.5.

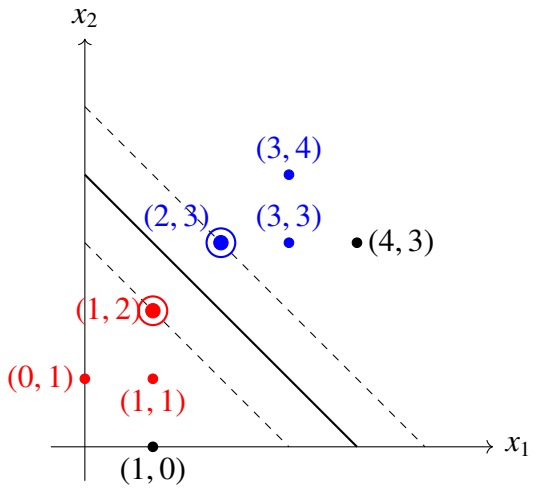


**Figura 2.5:** Clasificación de los puntos (4,3) y (1,0).

Este ejemplo ilustra cómo la SVM determina el hiperplano óptimo y cómo utiliza los puntos de soporte para realizar la clasificación.

## 2.3. Caso no linealmente separable

En la práctica, la mayoría de los conjuntos de entrenamiento no son linealmente separables, es decir, no se puede garantizar la existencia de un hiperplano que separe totalmente las clases 1 y $-1$. Es por ello que se desarrollaron diferentes técnicas para compensar el peso de los datos que no se encuentran en la clase correcta. Una de ellas es la adaptación de la teoría expuesta en las secciones anteriores a una que penaliza los datos que se encuentran en la clase incorrecta mediante una ponderación conocida como error, lo que permite separar las dos clases con el mismo principio del margen de separación.

### Método de margen suave

Muchos conjuntos de datos en la vida real contienen ruido, es decir, datos que no pertenecen a ninguna de las clases o que están en ambas, lo cual provoca una clasificación deficiente. Los

valores atípicos pueden influir de forma inadecuada en la posición del hiperplano de separación. Los efectos de los valores atípicos pueden reducirse introduciendo un *margen suave*. Sin embargo, la función de decisión es similar a la del caso linealmente separable:

$$\phi(x) = \text{signo}(\langle w, x\rangle + b). \tag{2.3.1}$$

En este estudio se introduce la siguiente restricción a los operadores de Lagrange:

$$0 \leq \alpha_i \leq C, \tag{2.3.2}$$

con $C$ una constante de regularización (véase la igualdad (2.3.17). El valor apropiado del parámetro $C$ puede ser encontrado por medio de un *estudio de validación*, sin embargo, este tema se sale del objetivo de la investigación.

La constante $C$ cuantifica el compromiso entre el ajuste a los datos de entrenamiento y la complejidad del hiperplano de separación. En el límite $C \to \infty$, el hiperplano óptimo coincide con aquel que separa completamente los datos, siempre que tal separación exista. Para valores finitos de $C$, el problema de clasificación se conoce como de *margen suave*, pues permite que algunos datos queden mal clasificados o dentro del margen. El parámetro $C$ es ajustable: valores grandes de $C$ asignan mayor penalización a los errores de clasificación, priorizando la correcta clasificación de los datos de entrenamiento, mientras que valores pequeños de $C$ conducen a hiperplanos más flexibles que toleran violaciones del margen con el fin de mejorar la generalización. En particular, valores finitos de $C$ resultan adecuados cuando los datos no son linealmente separables.

La justificación para usar técnicas de margen suave surge dentro de la teoría del aprendizaje estadístico y puede ser considerada como "relajación de las restricciones de margen duro". Se permiten algunos datos dentro del margen suave o incluso en el lado equivocado del hiperplano durante el entrenamiento; esta última posibilidad significa que se tiene un error de entrenamiento distinto de cero. Los puntos dentro de la banda de margen o en el lado equivocado del hiperplano son llamados *errores de margen*. Así también, como se ha señalado anteriormente,

los valores atípicos pueden influir de forma indebida en la posición del plano de separación. Los efectos de los valores atípicos pueden reducirse introduciendo un *margen suave*.

En el método de margen suave se introduce una variable de holgura $\xi = (\xi_1, \ldots, \xi_m)$, con $\xi_i \geq 0$, para todo $i = 1, 2, \ldots, m$, dentro de las restricciones del problema de optimización principal, es decir:

$$y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i, \text{para todo } i = 1, ..., m. \tag{2.3.3}$$

Ahora, la tarea consiste en minimizar la suma de errores agregada a la función objetivo original:

$$\text{Minimizar} \quad \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{m}\xi_i, \tag{2.3.4}$$

donde la constante $C$ determina la compensación entre la maximización del margen y la minimización del error de entrenamiento. Formalmente, se tiene el siguiente problema de optimización

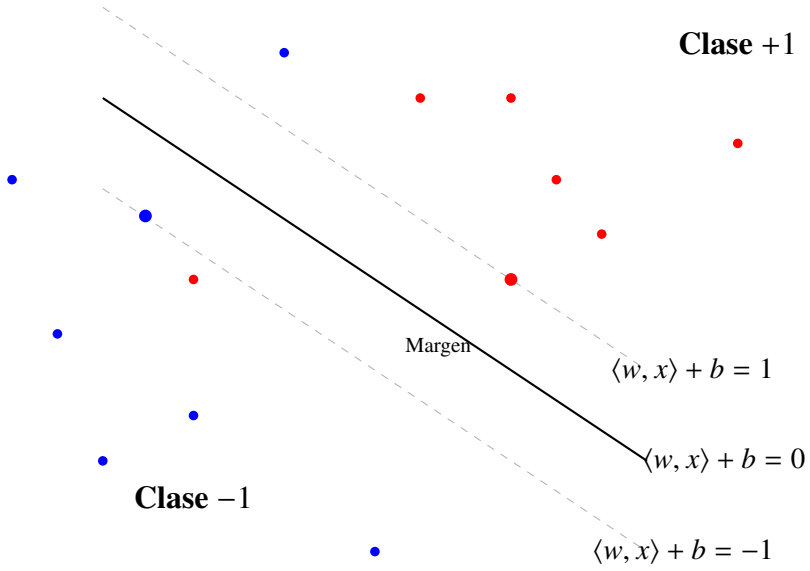


**Figura 2.6:** Hiperplano de separación (en negro), acompañado de los hiperplanos canónicos (en lineas punteadas). Obsérvese que hay un punto de cada clase en la contraria que rebasa los márgenes.

cuadrática:

$$\begin{aligned} \text{Minimizar} \quad & \frac{1}{2} w^T I w + C \sum_{i=1}^{m} \xi_i, \\ \text{sujeto a} \quad & -y_i(\langle w, x_i \rangle + b)) + 1 - \xi_i \leq 0, \\ & -\xi_i \leq 0, \\ & i = 1, \ldots, m. \end{aligned} \tag{2.3.5}$$

Si $\xi_i > 0$, se tiene un error de margen. Algo interesante es que este problema cumple con la condición de Slater, esto es, hace que los conjuntos que no son linealmente separables en cierto sentido logren serlo:

**Proposición 2.3.1.** El problema primal de la SVM con margen suave

$$\begin{aligned} \text{Minimizar} \quad & \frac{1}{2} w^T I w + C \sum_{i=1}^{m} \xi_i, \\ \text{sujeto a} \quad & -y_i(\langle w, x_i \rangle + b)) + 1 - \xi_i \leq 0, \\ & -\xi_i \leq 0, \\ & i = 1, \ldots, m. \end{aligned}$$

cumple con la condición de Slater.

*Demostración.* Debido a que las variables de holgura $\xi_i \geq 0$, basta elegir $w = 0$, $b = 0$ y $\xi_i > 1$ para que todas las restricciones se satisfagan de manera estricta, en efecto:

$$-y_i(\langle 0, x_i \rangle + 0)) + 1 - \xi_i = 1 - \xi_i < 0.$$

□

Esto garantiza la dualidad fuerte, incluso cuando los datos no son linealmente separables.

**Proposición 2.3.2.** Las restricciones del problema (2.3.5) son funciones afines respecto a las variables $w$, $b$ y $\xi$.

*Demostración.* Se denotan los vectores

$$z = (w,\ b,\ \xi,)^T \in \mathbb{R}^{n+1+m},$$

donde $\xi = (\xi_1\ ...\xi_n)^T$. Luego, se toman las funciones:

$$g_i(w, b, \xi) = -y_i(\langle w, x_i\rangle + b)) + 1 - \xi_i.$$

Y se define el vector:

$$q_i = (-y_i x_i,\ -y_i,\ \ e_i)^T \in \mathbb{R}^{n+1+m}, \tag{2.3.6}$$

donde $e_i$ es el vector de la base canónica del espacio $\mathbb{R}^m$. Las funciones $g_i(w, b, \xi)$ se pueden reescribir de la forma

$$\langle q_i, z\rangle + 1, \qquad \text{para cada } i = 1, \ldots, m. \tag{2.3.7}$$

Donde $\langle q_i, z\rangle$ es una aplicación lineal, por ende, una función afín. Para las restricciones $-\xi_i \leq 0$ se define el vector:

$$\bar{q}_i = (0_n,\ \ 0,\ -e_i)^T. \tag{2.3.8}$$

Por lo que las restricciones se pueden expresar como:

$$-\xi_i = \langle \bar{q}_i, z\rangle + 0, \tag{2.3.9}$$

siendo una función afín. □

El problema (2.3.5) se formula en un sentido dual al sumarle a la función objetivo las restricciones, con lo que se obtiene una función de Lagrange primaria:

$$\mathcal{L}(w, b, \alpha, \lambda) = \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i[y_i(\langle w, x_i\rangle + b) - 1 + \xi_i] - \sum_{i=1}^{m}\lambda_i\xi_i. \tag{2.3.10}$$

Con multiplicadores de Lagrange $\alpha_i \geq 0$ y $\lambda_i \geq 0$. Luego, las derivadas con respecto a $w$, $b$ y $\xi$ son

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial b} &= \frac{\partial}{\partial b}\left(\frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i[y_i(\langle w, x_i\rangle + b) - 1 + \xi_i] - \sum_{i=1}^{m}\lambda_i\xi_i\right)\\ &= -\frac{\partial}{\partial b}\left(\sum_{i=1}^{m}\alpha_i[y_i(\langle w, x_i\rangle + b) - 1 + \xi_i]\right)\\ &= -\sum_{i=1}^{m}\frac{\partial}{\partial b}(\alpha_i[y_i(\langle w, x_i\rangle + b) - 1 + \xi_i]\\ &= -\sum_{i=1}^{m}\alpha_i y_i.\end{aligned}$$

Al igualar la derivada a cero se obtiene que

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial b} &= 0,\\ \sum_{i=1}^{m}\alpha_i y_i &= 0.\end{aligned} \tag{2.3.11}$$

De esta manera, es posible hallar el valor del parámetro $w$:

**Proposición 2.3.3.** El valor $w$ de la función (2.3.1) es de la forma:

$$w = \sum_{i=1}^{m}\alpha_i y_i x_i. \tag{2.3.12}$$

*Demostración.* Se calcula la derivada parcial de $\mathcal{L}$ mostrada en (2.3.10) con respecto a $w$:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial w} &= \frac{\partial}{\partial w}\left(\frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i[y_i(\langle w, x_i\rangle + b) - 1 + \xi_i] - \sum_{i=1}^{m}\lambda_i\xi_i\right)\\ &= \frac{\partial}{\partial w}\left(\frac{1}{2}\|w\|_2^2\right) + \frac{\partial}{\partial w}\left(C\sum_{i=1}^{m}\xi_i\right) - \frac{\partial}{\partial w}\left(\sum_{i=1}^{m}\alpha_i[y_i(\langle w, x_i\rangle + b) - 1 + \xi_i]\right)\\ &\quad - \frac{\partial}{\partial w}\left(\sum_{i=1}^{m}\lambda_i\xi_i\right)\end{aligned}$$

$$
\begin{aligned}
&= w - \sum_{i=1}^{m} \frac{\partial}{\partial w}(\alpha_i[y_i(\langle w, x_i\rangle + b) - 1 + \xi_i]) \\
&= w - \sum_{i=1}^{m} \alpha_i y_i \frac{\partial}{\partial w}(\langle w, x_i\rangle + b) \\
&= w - \sum_{i=1}^{m} \alpha_i y_i x_i.
\end{aligned}
$$

Se iguala la derivada a cero y se despeja el valor de $w$ como

$$
w = \sum_{i=1}^{m} \alpha_i y_i x_i.
$$

□

Observe que el valor de $w$ mostrado en la Proposición 2.3.3 para el caso linealmente separable es de la misma forma que en el caso no linealmente separable mostrado en la Proposición 2.2.6.

Ahora, se calcula la derivada con respecto a $\xi$:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \xi} &= \frac{\partial}{\partial \xi}\left(\frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i[y_i(\langle w, x_i\rangle + b) - 1 + \xi_i] - \sum_{i=1}^{m}\lambda_i\xi_i\right) \\
&= \frac{\partial}{\partial \xi}\left(\frac{1}{2}\|w\|_2^2\right) + \frac{\partial}{\partial \xi}\left(C\sum_{i=1}^{m}\xi_i\right) - \frac{\partial}{\partial \xi}\left(\sum_{i=1}^{m}\alpha_i[y_i(\langle w, x_i\rangle + b) - 1 + \xi_i]\right) \\
&\quad - \frac{\partial}{\partial \xi}\left(\sum_{i=1}^{m}\lambda_i\xi_i\right) \\
&= \frac{\partial}{\partial \xi}\left(C\sum_{i=1}^{m}\xi_i\right) - \frac{\partial}{\partial \xi}\left(\sum_{i=1}^{m}\alpha_i[y_i(\langle w, x_i\rangle + b) - 1 + \xi_i]\right) - \frac{\partial}{\partial \xi}\left(\sum_{i=1}^{m}\lambda_i\xi_i\right).
\end{aligned}
$$

Denótese $\overrightarrow{1}$ como el vector de tamaño $m$ donde todas sus entradas son el valor 1. Y se tiene

$$
\frac{\partial \mathcal{L}}{\partial \xi} = \frac{\partial}{\partial \xi}\left(C\langle \overrightarrow{1}, \xi\rangle\right) - \sum_{i=1}^{m}\frac{\partial}{\partial \xi}(\alpha_i[y_i(\langle w, x_i\rangle + b) - 1 + \xi_i]) - \frac{\partial}{\partial \xi}(\langle \lambda, \xi\rangle).
$$

El símbolo $\overrightarrow{C}$ denota el vector de tamaño $m$ donde todas sus entradas son el valor $C$, con esto, la igualdad queda

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \xi} &= \overrightarrow{C} - \frac{\partial}{\partial \xi}(\langle \alpha, \xi \rangle) - \lambda \\ &= \overrightarrow{C} - \alpha - \lambda.\end{aligned}$$

Posteriormente, se iguala la derivada a cero:

$$\frac{\partial \mathcal{L}}{\partial \xi} = 0,$$

es decir:

$$\begin{aligned}\overrightarrow{C} - \alpha - \lambda &= 0, \\ \lambda &= \overrightarrow{C} - \alpha,\end{aligned} \tag{2.3.13}$$

de aquí que

$$\lambda_i = C - \alpha_i, \text{ para cada } i = 1, \ldots, m. \tag{2.3.14}$$

Se sustituye el valor de $w$ en la función (2.3.10) para formular la función Lagrangiana dual:

$$\begin{aligned}\mathcal{W}(\alpha) = &\frac{1}{2}\left\langle \sum_{i=1}^{m} \alpha_i y_i x_i, \sum_{i=1}^{m} \alpha_i y_i x_i \right\rangle + C \sum_{i=1}^{m} \xi_i \\ &- \sum_{i=1}^{m} \alpha_i \left[ y_i \left( \left\langle \sum_{j=1}^{m} \alpha_j y_j x_j, x_i \right\rangle + b \right) - 1 + \xi_i \right] - \sum_{i=1}^{m} \lambda_i \xi_i \\ = &\frac{1}{2} \sum_{i=1}^{m} \alpha_i y_i \sum_{j=1}^{m} \alpha_j y_j \langle x_i, x_j \rangle + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i y_i \sum_{j=1}^{m} \alpha_j y_j \langle x_j, x_i \rangle \\ &+ \sum_{i=1}^{m} b \alpha_i y_i + \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \xi_i \alpha_i - \sum_{i=1}^{m} \lambda_i \xi_i\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i y_i \alpha_j y_j \langle x_i, x_j\rangle + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i y_i \alpha_j y_j \langle x_i, x_j\rangle \\
&\quad + b\sum_{i=1}^{m}\alpha_i y_i + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\xi_i\alpha_i - \sum_{i=1}^{m}\lambda_i\xi_i \\
&= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i y_i \alpha_j y_j \langle x_i, x_j\rangle + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\xi_i\alpha_i - \sum_{i=1}^{m}\lambda_i\xi_i \\
&= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i y_i \alpha_j y_j \langle x_i, x_j\rangle + \sum_{i=1}^{m}\alpha_i + C(\xi_1 + \xi_2 + \cdots + \xi_m) \\
&\quad - (\xi_1\alpha_1 + \cdots + \xi_m\alpha_m) - (\lambda_1\xi_1 + \cdots + \lambda_m\xi_m,) \\
&= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i y_i \alpha_j y_j \langle x_i, x_j\rangle + \sum_{i=1}^{m}\alpha_i \\
&\quad + [(C - \alpha_1 - \lambda_1)\xi_1 + \cdots + (C - \alpha_m - \alpha_m - \lambda_m)\xi_m] \\
&= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i y_i \alpha_j y_j \langle x_i, x_j\rangle + \sum_{i=1}^{m}\alpha_i + \sum_{i=1}^{m}(C - \alpha_i - \lambda_i)\xi_i \\
&= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i y_i \alpha_j y_j \langle x_i, x_j\rangle + \sum_{i=1}^{m}\alpha_i.
\end{aligned}
$$

Por lo tanto, la función dual lagrangiana queda formulada como

$$
\mathcal{W}(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \langle x_i\ , x_j\rangle. \tag{2.3.15}
$$

Nótese que la ecuación (2.3.15) es la misma que la función dual lagrangiana del caso linealmente separable (ver (2.2.13) para comparar). Por la Proposición 2.2.7, se tiene que la función $\mathcal{W}$ es cóncava, por lo que se puede utilizar el método del gradiente ascendente para encontrar los valores del vector $\alpha$.

Al sustituir (2.3.14) en $\lambda_i \geq 0$ se obtiene la restricción:

$$
C \geq \alpha_i. \tag{2.3.16}
$$

Dado que $\alpha_i \geq 0$ y en conjunto con la desigualdad (2.3.16), se concluye que

$$0 \leq \alpha_i \leq C, \qquad \text{para cada } i = 1, 2, \ldots, m. \tag{2.3.17}$$

Los puntos con valores $\alpha_i = 0$ o $\alpha_i = C$ se denominarán *límites* y aquellos $0 < \alpha_i < C$ se denominarán *no límites* . De las condiciones KKT del Teorema 1.2.38, se incluyen las dos ecuaciones siguientes:

$$\lambda_i \xi_i = 0, \tag{2.3.18}$$

$$(y_i(\langle w, x_i \rangle + b) - 1 + \xi_i) = 0. \tag{2.3.19}$$

El valor del sesgo $b$ de la función de decisión (2.3.1) se calcula de la manera establecida en la siguiente proposición.

**Proposición 2.3.4.** Sea $\{\alpha_i\}_{i=1}^m$ una solución óptima del problema dual de la SVM de margen suave, y sea

$$M = \{\, k : 0 < \alpha_k < C \,\}$$

el conjunto de índices correspondientes a los vectores de soporte que satisfacen las condiciones de KKT con igualdad estricta. Para cada $k \in M$ defínase

$$b_k \;=\; y_k \;-\; \sum_{i=1}^{m} \alpha_i y_i \langle x_i, x_k \rangle.$$

Entonces el sesgo óptimo $b$ en la función de decisión (2.3.1) está dado por el promedio de estos valores, es decir,

$$b^* \;=\; \frac{1}{|M|} \sum_{k \in M} b_k.$$

*Demostración.* Se seleccionan los puntos no límites con valores $\alpha_k$, es decir, $0 < \alpha_k < C$. De la igualdad (2.3.14) y dado que $\alpha_k < C$ se deduce que $\lambda_k > 0$, aunado a esto, de la restricción (2.3.18), se obtiene $\xi_k = 0$. Sustituyendo los valores encontrados en la restricción (2.3.19), se tiene

$$y_i(\langle w, x_i \rangle + b) - 1 = 0. \tag{2.3.20}$$

Finalmente, dado que $y_k^2 = 1$ y $w = \sum_{i=1}^{m} \alpha_i y_i x_i$, se sustituye en la igualdad (2.3.20):

$$\begin{aligned} y_k \left( \left\langle \sum_{i=1}^{m} \alpha_i y_i x_i, x_k \right\rangle + b_k \right) - 1 &= 0, \\ y_k \left( \sum_{i=1}^{m} \alpha_i y_i \langle x_i, x_k \rangle + b_k \right) &= 1, \\ y_k \left( \sum_{i=1}^{m} \alpha_i y_i \langle x_i, x_k \rangle + b_k \right) &= y_k^2, \\ \sum_{i=1}^{m} \alpha_i y_i \langle x_i, x_k \rangle + b_k &= y_k, \\ b_k &= y_k - \sum_{i=1}^{m} \alpha_i y_i \langle x_i, x_k \rangle. \end{aligned}$$

Por lo tanto, el parámetro $b$ es el promedio sobre todos los $b_k$ donde $k \in M = \{k : 0 < \alpha_k < C\}$, esto es,

$$b^* = \frac{\sum_{k \in M} b_k}{|M|}. \tag{2.3.21}$$

□

Finalmente, si los datos de entrenamiento no son linealmente separables, para una nueva entrada $z \in \mathbb{R}^n$, la clase está determinada por la función de decisión que se define como

$$\begin{aligned} \phi(z) &= \text{signo}(\langle w^*, z \rangle + b^*) \\ &= \text{signo}\left( \left\langle \sum_{i=1}^{m} \alpha_i y_i x_i \, , z \right\rangle + b^* \right). \end{aligned} \tag{2.3.22}$$

A manera de ejemplo, considérese el siguiente conjunto de datos en $\mathbb{R}^2$:

| $i$ | $x_i$ | $y_i$ |
|---|---|---|
| 1 | $(0,0)$ | $+1$ |
| 2 | $(1,0)$ | $+1$ |
| 3 | $(0,1)$ | $+1$ |
| 4 | $(1,1)$ | $-1$ |
| 5 | $(2,1)$ | $-1$ |
| 6 | $(1,2)$ | $-1$ |

Este conjunto no es linealmente separable (Figura 2.7):

**Proposición 2.3.5.** El conjunto de datos en $\subset \mathbb{R}^2 \times \{-1,+1\}$ dado por

$$\{(x_i,y_i)\}_{i=1}^{6} = \{((0,0),+1),\ ((1,0),+1),\ ((0,1),+1),\ ((1,1),-1),\ ((2,1),-1),\ ((1,2),-1)\}$$

no es linealmente separable.

*Demostración.* Supóngase, por contradicción, que el conjunto es linealmente separable. Entonces existen un vector $w = (w_1, w_2) \in \mathbb{R}^2$ y un escalar $b \in \mathbb{R}$ tales que

$$y_i(\langle w, x_i\rangle + b) > 0 \quad \text{para todo } i = 1,\dots,6.$$

Para los puntos con etiqueta $+1$ se tiene

$$\langle w,(0,0)\rangle + b > 0,$$
$$\langle w,(1,0)\rangle + b > 0,$$
$$\langle w,(0,1)\rangle + b > 0.$$

De la primera desigualdad se deduce que $b > 0$, y de las dos restantes que

$$w_1 + b > 0, \qquad w_2 + b > 0.$$

Por otro lado, para los puntos con etiqueta $-1$ debe cumplirse

$$\langle w, x_i \rangle + b < 0.$$

En particular,

$$\langle w, (1,1) \rangle + b = w_1 + w_2 + b < 0.$$

Sin embargo, al sumar las desigualdades $w_1 + b > 0$ y $w_2 + b > 0$ se obtiene

$$w_1 + w_2 + 2b > 0.$$

Como además $b > 0$, se sigue que

$$w_1 + w_2 + b > 0,$$

lo cual contradice la condición necesaria para clasificar correctamente al punto $(1,1)$ con etiqueta $-1$.

Esta contradicción muestra que no existen $w \in \mathbb{R}^2$ y $b \in \mathbb{R}$ capaces de separar linealmente ambas clases. Por lo tanto, el conjunto de datos no es linealmente separable. □

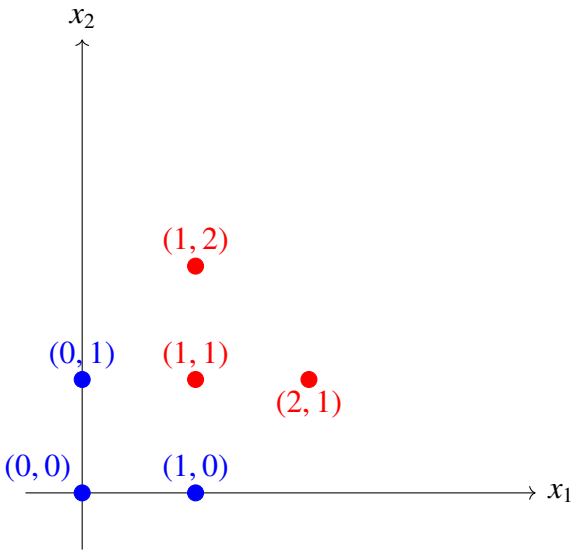


**Figura 2.7:** Gráfica de puntos de entrenamiento.

El problema de optimización para la SVM utilizando margen suave aplicado a este problema

es

$$\begin{aligned} \text{Minimizar} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{6}\xi_i \\ \text{sujeto a} \quad & y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i, \quad i = 1, \ldots, 6, \\ & \xi_i \geq 0. \end{aligned}$$

El parámetro $C > 0$ controla la relación entre la amplitud del margen y la penalización por errores, esto es, determina cuánto se penalizan las violaciones al margen, este parámetro $C$ controla el equilibrio entre maximizar el margen y minimizar los errores de clasificación.

La Lagrangiana asociada es la función $\mathcal{L} : \mathbb{R}^2 \times \mathbb{R} \times \mathbb{R}^6 \times \mathbb{R}^6 \times \mathbb{R}^6 \to \mathbb{R}$ definida como

$$\begin{aligned} \mathcal{L}(w, b, \xi, \alpha, \mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{6}\xi_i \\ - \sum_{i=1}^{6}\alpha_i(y_i(\langle w, x_i\rangle + b) - 1 + \xi_i) - \sum_{i=1}^{6}\mu_i\xi_i. \end{aligned}$$

El problema dual se formula como

$$\begin{aligned} \text{Maximizar} \quad & \mathcal{W}(\alpha) = \sum_{i=1}^{6}\alpha_i - \frac{1}{2}\sum_{i=1}^{6}\sum_{j=1}^{6}\alpha_i\alpha_j y_i y_j\langle x_i, x_j\rangle \\ \text{sujeto a} \quad & \sum_{i=1}^{6}\alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq C. \end{aligned}$$

Ahora, defínase la matriz $Q \in \mathbb{R}^{6\times 6}$ donde sus entradas están dadas por:

$$Q_{ij} = y_i y_j\langle x_i, x_j\rangle,$$

utilizando esta matriz, la función $\mathcal{W} : \mathbb{R}^6 \to \mathbb{R}$ dual puede escribirse como

$$\mathcal{W}(\alpha) = \mathbf{1}^t\alpha - \frac{1}{2}\alpha^t Q\alpha.$$

El gradiente de $\mathcal{W}$ está dado por:

$$\nabla\mathcal{W}(\alpha) = \mathbf{1} - Q\alpha.$$

donde la matriz $Q$ es de la siguiente forma:

$$Q = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & -2 & -1 \\ 0 & 0 & 1 & -1 & -1 & -2 \\ 0 & -1 & -1 & 2 & 3 & 3 \\ 0 & -2 & -1 & 3 & 5 & 4 \\ 0 & -1 & -2 & 3 & 4 & 5 \end{pmatrix}.$$

El método de gradiente ascendente se define por la iteración:

$$\alpha^{(k+1)} = \alpha^{(k)} + \eta(\mathbf{1} - Q\alpha^{(k)}),$$

donde $\eta = 1$ es el tamaño de paso.

Para iniciar el método del gradiente ascendente se toma

$$\alpha^{(0)} = 0 \in \mathbb{R}^6,$$

En la primera iteración se obtiene:

$$\alpha^{(1)} = \eta\mathbf{1}.$$

Tras varias iteraciones, el algoritmo converge a:

$$\alpha^\star = (C, 0, 0, C, 0, 0)^t,$$

el cual satisface:

$$\sum_{i=1}^{6} \alpha_i^\star y_i = C - C = 0.$$

Para calcular el valor del parámetro $w$ del hiperplano de separación se utiliza la Proposición 2.3.3 y con los valores $x_1 = (0, 0)$ y $x_4 = (1, 1)$ el valor de $w$ está dado por:

$$w = \sum_{i=1}^{6} \alpha_i y_i x_i,$$

se obtiene:

$$w = C((0, 0) - (1, 1)) = (-C, -C).$$

El vector $w$ se puede normalizar al dividir entre $-C$ (esto sin pérdida de generalidad), ya que el signo es irrelevante para el hiperplano, por lo que:

$$w = (1, 1).$$

Para el cálculo de $b$ se utiliza la Proposición 2.3.4. Dado que los índices de los valores $\alpha_i > 0$ son $i = 1, 4$, entonces el valor de $b_1$ y $b_4$ son

$$\begin{aligned} b_1 &= 1 - C(\langle x_1, x_1\rangle + \langle x_4, x_1\rangle) \\ &= 1 - C(\langle (0, 0), (0, 0)\rangle + \langle (1, 1), (0, 0)\rangle) \\ &= 1. \end{aligned}$$

$$\begin{aligned} b_4 &= -1 - C(\langle x_1, x_4\rangle + \langle x_4, x_4\rangle) \\ &= -1 - C(\langle (0, 0), (1, 1)\rangle + \langle (1, 1), (1, 1)\rangle) \\ &= -1 - 2C. \end{aligned}$$

así, el valor óptimo del parámetro $b$ es:

$$b^* = \frac{1}{2}(1 - 1 - 2C) = -C$$

Tomando el valor de $C = 1$, se tiene que $b^* = -1$. Gráficamente se puede ver en la Figura 2.8.

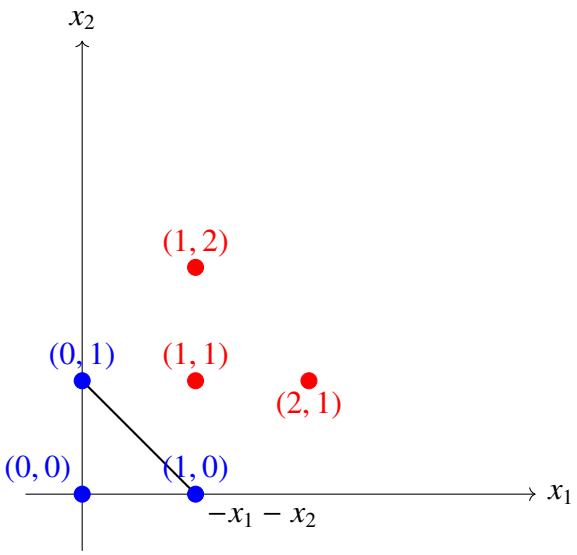


**Figura 2.8:** Gráfica de puntos de entrenamiento y el hiperplano de separación. Observe que ahora dos puntos pertenecen al hiperplano debido a que se permiten violaciones del margen.

Por lo que la función de decisión final está dada por:

$$\phi(x) = \text{signo}(x_1 + x_2 - 1).$$

Con lo anterior, se ha resuelto completamente el problema dual de la SVM con margen suave para un conjunto de datos no linealmente separables, obteniendo explícitamente los multiplicadores de Lagrange, el vector normal $w$ y el término independiente $b$.

## 2.4. Funciones Kernel y aumento de dimensionalidad

En la sección anterior se estableció un método basado en permitir que ciertos datos se encuentren en la clase incorrecta; sin embargo, existen casos en los que una clase se encuentra tan inmersa dentro de otra que esta técnica ya no es factible. Una estrategia para separarlas es introducir funciones (denominadas *kernels*) que mapeen estos datos a otro espacio de dimensión mayor, en el que se permita trabajar con los datos con el algoritmo en su forma lineal sin perder alguna de las características del espacio de entrada. Es decir, que en un espacio de dimensión

mayor los datos se vuelven linealmente separables; esto lo establece el teorema llamado de "Separabilidad mediante aumento de dimensionalidad", que es uno de los teoremas principales de esta sección (Teorema 2.4.23).

Para obtener una representación alterna de los datos, se mapean en un espacio de dimensionalidad diferente, denominado *espacio de características*, mediante una sustitución:

$$\langle x_i, x_j \rangle \to \langle \psi(x_i), \psi(x_j) \rangle,$$

donde la función $\psi : \mathbb{R}^n \to \mathbb{R}^p$, con $p \gg n$, es la función de mapeo. Una razón para realizar este mapeo es que los datos utilizados pueden no ser linealmente separables en el espacio de entrada, ya que puede que no se encuentre un hiperplano orientado que separe las dos clases de datos.

Antes de formular el teorema de separabilidad, se mostrarán algunos teoremas básicos en su demostración; se inicia con el Teorema de Cover, el cual fue mostrado por Thomas M. Cover en [14].

**Teorema 2.4.1** (Teorema de Cover, 1965)**.** Sea $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ un conjunto de $m$ puntos en $\mathbb{R}^d$ en posición general, es decir, ningún subconjunto de $d+1$ puntos está contenido en un hiperplano de dimensión $d-1$. Supóngase que cada punto puede pertenecer a una de dos clases, etiquetadas como $+1$ ó $-1$. El número máximo de dicotomías (particiones en dos clases) de $S$ que son linealmente separables es:

$$\Gamma(m, d) = \begin{cases} 2^m & \text{si } m \leq d+1; \\ 2\displaystyle\sum_{k=0}^{d} \binom{m-1}{k} & \text{si } m > d+1, \end{cases}$$

donde, por convención, $\binom{n}{k} = 0$ cuando $k > n$.

Equivalentemente, para todo $m, d \geq 1$:

$$\Gamma(m, d) = 2 \sum_{k=0}^{\text{mín}(d,m-1)} \binom{m-1}{k}.$$

En particular, cuando $d$ crece, la probabilidad de que una clasificación aleatoria de los puntos sea linealmente separable tiende rápidamente a 1.

*Demostración.* La demostración se estructura en varios pasos:

- Paso 1: Espacio aumentado y notación.

  Se define el espacio aumentado $\mathbb{R}^{d+1}$ mediante la transformación:

  $$\tilde{\mathbf{x}}_i = \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix} \in \mathbb{R}^{d+1}, \quad i = 1, \ldots, m.$$

  Un hiperplano en $\mathbb{R}^d$ con ecuación $\langle w, x \rangle + b = 0$ se convierte en un hiperplano que pasa por el origen en $\mathbb{R}^{d+1}$:

  $$\tilde{\mathbf{w}}^t \tilde{\mathbf{x}} = 0, \quad \text{donde } \tilde{\mathbf{w}} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}.$$

  Una dicotomía $y = (y_1, \ldots, y_m) \in \{+1, -1\}^m$ es linealmente separable si existe $\tilde{\mathbf{w}} \in \mathbb{R}^{d+1}$ tal que:

  $$y_i(\langle \tilde{\mathbf{w}}, \tilde{\mathbf{x}}_i \rangle) > 0, \quad \text{para toda } i = 1, \ldots, m.$$

  La condición de posición general en $\mathbb{R}^d$ implica que los puntos $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_m$ en $\mathbb{R}^{d+1}$ están en posición general: ningún subconjunto de $d + 2$ puntos es linealmente dependiente.

- Paso 2: Relación de recurrencia

  Para $m \geq 1$ y $d \geq 0$, defínase $C(m, d)$ como el número máximo de dicotomías linealmente separables de $m$ puntos en posición general en $\mathbb{R}^d$.

**Lema 2.4.2** (Recurrencia fundamental)**.** Para $m \geq 2$ y $d \geq 1$, se cumple:

$$\Gamma(m, d) = \Gamma(m-1, d) + \Gamma(m-1, d-1).$$

*Demostración.* Considere $m$ puntos $\mathbf{x}_1, \ldots, \mathbf{x}_m$ en posición general en $\mathbb{R}^d$, es decir, no son colineales. Fíjese los primeros $m-1$ puntos y supóngase que se añade $\mathbf{x}_m$.

Sea $\mathcal{D}_{m-1}$ el conjunto de dicotomías linealmente separables de $\{\mathbf{x}_1, \ldots, \mathbf{x}_{m-1}\}$. Para cada $y \in \mathcal{D}_{m-1}$, existe un hiperplano separador $\tilde{\mathbf{w}}$ en el espacio aumentado.

Al añadir $\mathbf{x}_m$, hay dos casos:

(I) Dicotomías no críticas: Existe un hiperplano separador $\tilde{\mathbf{w}}$ para $y$ tal que $\langle \tilde{\mathbf{w}}, \tilde{\mathbf{x}}_m \rangle \neq 0$. En este caso, se puede extender la dicotomía a los $m$ puntos de modo que se asigna $y_m = +1$ ó $y_m = -1$, se ajusta ligeramente $\tilde{\mathbf{w}}$ si es necesario. Cada dicotomía no crítica produce dos extensiones.

(II) Dicotomías críticas: Todo hiperplano separador $\tilde{\mathbf{w}}$ para $y$ satisface $\tilde{\mathbf{w}}^t \tilde{\mathbf{x}}_m = 0$. Esto significa que $\mathbf{x}_m$ está en el hiperplano separador. En este caso, el signo de $y_m$ está determinado por la orientación del hiperplano. Cada dicotomía crítica produce una extensión.

El número de dicotomías críticas de $m-1$ puntos es exactamente $\Gamma(m-1, d-1)$. Para ver esto, se implementa la restricción $\tilde{\mathbf{w}}^t \tilde{\mathbf{x}}_m = 0$. Esto define un subespacio de dimensión $d$ en $\mathbb{R}^{d+1}$. Si se proyectan ortogonalmente los puntos $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_{m-1}$ sobre este subespacio, se obtienen puntos en $\mathbb{R}^d$ en posición general. Las dicotomías críticas corresponden a dicotomías linealmente separables de estos $m-1$ puntos en $\mathbb{R}^{d-1}$ (en el espacio original).

Si hay $\Gamma_{\text{crit}} = \Gamma(m-1, d-1)$ dicotomías críticas y $\Gamma_{\text{no-crit}} = \Gamma(m-1, d) - \Gamma(m-1, d-1)$ dicotomías no críticas, entonces:

$$\Gamma(m, d) = \Gamma_{\text{no-crit}} \cdot 2 + \Gamma_{\text{crit}} \cdot 1 = 2[\Gamma(m-1, d) - \Gamma(m-1, d-1)] + \Gamma(m-1, d-1).$$

Es decir

$$\Gamma(m, d) = 2\Gamma(m-1, d) - \Gamma(m-1, d-1).$$

Sin embargo, esta es una forma alternativa de la recurrencia. Para obtener $\Gamma(m, d) = \Gamma(m-1, d) + \Gamma(m-1, d-1)$, se tiene que:

$$\Gamma(m, d) - \Gamma(m-1, d) = \Gamma(m-1, d) - \Gamma(m-1, d-1).$$

Se define $D(m, d) = \Gamma(m, d) - \Gamma(m-1, d)$, se obtiene $D(m, d) = D(m-1, d)$. Se itera esta recurrencia:

$$D(m, d) = D(m-1, d) = \cdots = D(1, d).$$

Pero $\Gamma(1, d) = 2$ (un solo punto siempre es separable) y $\Gamma(0, d) = 1$ (conjunto vacío), luego $D(1, d) = 1$. Por tanto:

$$\Gamma(m, d) - \Gamma(m-1, d) = 1.$$

Lo cual es incorrecto.

La recurrencia correcta se obtiene directamente del argumento geométrico: las dicotomías de $m$ puntos se dividen en:

- Aquellas donde $\mathbf{x}_m$ se puede separar como $+1$: corresponden a dicotomías de $m-1$ puntos en $\mathbb{R}^d$.
- Aquellas donde $\mathbf{x}_m$ se puede separar como $-1$: también corresponden a dicotomías de $m-1$ puntos en $\mathbb{R}^d$.

Pero esto contaría cada dicotomía dos veces. La corrección es: fíjese la etiqueta de $\mathbf{x}_m$. Si $y_m = +1$, los $m-1$ puntos restantes deben ser separable con un hiperplano que deje $\mathbf{x}_m$ en el lado positivo. Esto equivale a dicotomías de $m-1$ puntos en $\mathbb{R}^d$ que son extendibles con $y_m = +1$, que son $\Gamma(m-1, d)$. Similar para $y_m = -1$. Pero estas dos clases se solapan exactamente en las dicotomías críticas, que se cuentan en ambos. Se aplica el principio

de inclusión-exclusión:

$$\Gamma(m,d) = \Gamma(m-1,d) + \Gamma(m-1,d) - \Gamma(m-1,d-1) = 2\Gamma(m-1,d) - \Gamma(m-1,d-1).$$

Esta es la recurrencia correcta. Se verifica que con la fórmula cerrada se satisface $\Gamma(m,d) = \Gamma(m-1,d) + \Gamma(m-1,d-1)$. □

- Paso 3: Solución de la recurrencia.

**Lema 2.4.3** (Fórmula cerrada)**.** La solución de la recurrencia $\Gamma(m,d) = 2\Gamma(m-1,d) - \Gamma(m-1,d-1)$ con condiciones iniciales:

$$\Gamma(1,d) = 2 \quad \text{para } d \geq 0, \qquad \Gamma(m,0) = 2 \quad \text{para } m \geq 1,$$

es:

$$\Gamma(m,d) = 2\sum_{k=0}^{d} \binom{m-1}{k}, \quad m \geq 1, d \geq 0.$$

*Demostración.* Se realiza la inducción en $m$ y $d$.

Caso base: Para $m = 1$,

$$2\sum_{k=0}^{d} \binom{0}{k} = 2\binom{0}{0} = 2 = \Gamma(1,d).$$

Para $d = 0$:

$$2\sum_{k=0}^{0} \binom{m-1}{0} = 2 = \Gamma(m,0).$$

Paso inductivo: Supóngase que la fórmula es válida para $m-1$ y para $d-1$. La recurrencia

queda como

$$\begin{aligned}
2\Gamma(m-1,d)-\Gamma(m-1,d-1) &= 2\left[2\sum_{k=0}^{d}\binom{m-2}{k}\right]-\left[2\sum_{k=0}^{d-1}\binom{m-2}{k}\right] \\
&= 2\left[2\sum_{k=0}^{d}\binom{m-2}{k}-\sum_{k=0}^{d-1}\binom{m-2}{k}\right] \\
&= 2\left[\binom{m-2}{0}+\sum_{k=1}^{d}\binom{m-2}{k}+\sum_{k=0}^{d-1}\binom{m-2}{k}\right] \\
&= 2\left[\binom{m-2}{0}+\sum_{k=1}^{d}\left(\binom{m-2}{k}+\binom{m-2}{k-1}\right)\right].
\end{aligned}$$

Se usa la identidad binomial $\binom{m-2}{k}+\binom{m-2}{k-1}=\binom{m-1}{k}$ y se obtiene:

$$\begin{aligned}
2\Gamma(m-1,d)-\Gamma(m-1,d-1) &= 2\left[\binom{m-2}{0}+\sum_{k=1}^{d}\binom{m-1}{k}\right] \\
&= 2\left[\binom{m-1}{0}+\sum_{k=1}^{d}\binom{m-1}{k}\right] \quad \text{pues } \binom{m-2}{0}=1=\binom{m-1}{0} \\
&= 2\sum_{k=0}^{d}\binom{m-1}{k}=\Gamma(m,d).
\end{aligned}$$

Esto completa la inducción. □

- Paso 4: Interpretación para $m \leq d+1$

  Cuando $m \leq d+1$, todos los $\binom{m-1}{k}$ para $k \leq d$ son no nulos. Si se usa la identidad:

$$\sum_{k=0}^{m-1}\binom{m-1}{k}=2^{m-1},$$

se obtiene

$$\Gamma(m,d)=2\sum_{k=0}^{d}\binom{m-1}{k}=2\sum_{k=0}^{m-1}\binom{m-1}{k}=2\cdot 2^{m-1}=2^{m},$$

pues para $k > m-1$, $\binom{m-1}{k}=0$. Esto significa que cuando $m \leq d+1$, todas las $2^m$ dicotomías posibles son linealmente separables.

- Paso 5: La probabilidad de que una dicotomía aleatoria uniforme de $m$ puntos en posición general en $\mathbb{R}^d$ sea linealmente separable es:

$$P_{\text{sep}}(m,d) = \frac{\Gamma(m,d)}{2^m} = \frac{1}{2^{m-1}} \sum_{k=0}^{\text{mín}(d,m-1)} \binom{m-1}{k}.$$

  Para $m$ fijo y $d \to \infty$, si $d \geq m-1$, entonces $P_{\text{sep}}(m,d) = 1$. Si $d$ crece, pero $m$ crece con $d$ de manera que $m/d \to \alpha$, entonces $P_{\text{sep}}(m,d)$ tiende a una función escalón en $\alpha = 2$.

□

Aunque se trasladen los datos a una nueva dimensión en la que son separables y se pueda definir un margen, no hay pérdida de información, como se demostrará mediante los teoremas siguientes. De hecho, para llevar un conjunto de datos que no es separable a un espacio de mayor dimensión donde sí lo es, se utiliza un kernel gaussiano, el cual se define como

$$K(x_i, x_j) = e^{-(x_i - x_j)^2/2\sigma^2}.$$

Para introducir el teorema que garantiza la separabilidad, se requieren algunos conceptos y resultados clave para entenderlo y dar la prueba, por lo que se exponen a continuación.

**Definición 2.4.4.** Sea $X \subseteq \mathbb{R}^n$. Una función kernel $K : X \times X \to \mathbb{R}$ es simétrica si

$$K(x,y) = K(y,x).$$

La función kernel es definida positiva si

$$\sum_{i=1}^{n} \sum_{j=1}^{n} K(x_i, x_j) c_i c_j \geq 0,$$

para todas las sucesiones finitas de puntos $x_1, x_2, \ldots x_n$ de $X$ y todas las elecciones de números reales $c_1, c_2, \ldots, c_n$.

**Definición 2.4.5.** Sean $X \subset \mathbb{R}^n$ y $K : X \times X \to \mathbb{R}$ un kernel. El kernel $K$ es continuo si, para cada $(x, z) \in X \times X$, se cumple

$$\lim_{(x',z')\to(x,z)} K(x', z') = K(x, z).$$

**Definición 2.4.6.** Sean $X$ y $Y$ espacios normados. Un operador $T : X \to Y$ lineal y continuo es compacto si para toda sucesión acotada $\{x_n\} \subset X$, la sucesión $\{T(x_n)\} \subset Y$ tiene una subsucesión convergente en $Y$.

Recuerde que un conjunto $K$ es compacto en un espacio normado si toda sucesión en $K$ admite una subsucesión convergente con límite en $K$.

**Definición 2.4.7** (Operador relativamente compacto)**.** Sea $X$ un espacio normado y sea $T : X \to X$ un operador lineal. Se dice que $T$ es *relativamente compacto* si la imagen de la bola unitaria cerrada

$$B_X = \{x \in X : \|x\| \leq 1\}$$

es un conjunto relativamente compacto en $X$, es decir, si su clausura es compacta.

En espacios normados (en particular, en espacios de Banach o de Hilbert), la noción de operador relativamente compacto coincide con la de operador compacto, en esta ocasión, se omitirá la prueba de este hecho.

Recuerde que una sucesión $\{x_n\} \subset X$ de un espacio normado se llama *sucesión de Cauchy* si para todo $\varepsilon > 0$ existe $N \in \mathbb{N}$ tal que $\|x_n - x_m\| < \varepsilon,$ para todo $n,\ m \geq N$. Y que un espacio de Hilbert $H$ es un espacio prehilbertiano en el cual toda sucesión de Cauchy en $H$ converge (con respecto a la norma inducida por el producto interno) a un elemento de $H$. En lo siguiente se usará la letra $H$ para denotar un espacio de Hilbert.

**Definición 2.4.8.** Sea $(X, \mathcal{A}, \mu)$ un espacio de medida. Se define el espacio $L^2(X)$ como el conjunto de clases de equivalencia de funciones medibles $f : X \to \mathbb{R}$ (o $\mathbb{C}$) tales que

$$\int_X |f(x)|^2 \, d\mu(x) < \infty.$$

Dos funciones $f$ y $g$ se identifican si coinciden $\mu$-casi en todas partes, es decir,

$$f(x) = g(x) \quad \text{para casi todo } x \in X.$$

El espacio $L^2(X)$ se dota del producto interno:

$$\langle f, g \rangle_{L^2} = \int_X f(x)\, g(x)\, d\mu(x) \quad \left( \text{o} \int_X f(x)\, \overline{g(x)}\, d\mu(x) \text{ en el caso complejo} \right),$$

el cual induce la norma

$$\|f\|_2 = \left( \int_X |f(x)|^2 \, d\mu(x) \right)^{1/2}.$$

Con esta estructura, $L^2(X)$ es un espacio de Hilbert.

**Definición 2.4.9** (Operador autoadjunto)**.** Un operador lineal $T : H \to H$ se llama autoadjunto si

$$\langle Tx, y \rangle = \langle x, Ty \rangle, \quad \text{para todo } x, y \in H.$$

**Definición 2.4.10** (Valor propio y vector propio)**.** Sea $T : H \to H$ un operador lineal. Un escalar $\lambda \in \mathbb{F}$ se llama valor propio de $T$ si existe un vector no nulo $x \in H$ tal que

$$Tx = \lambda x.$$

En tal caso, $x$ se denomina vector propio asociado a $\lambda$.

**Definición 2.4.11** (Multiplicidad de un valor propio)**.** La multiplicidad (geométrica) de un valor propio $\lambda$ de $T$ se define como la dimensión del subespacio

$$\ker(T - \lambda I) = \{x \in H : (T - \lambda I)x = 0\}.$$

**Definición 2.4.12** (Conjunto ortonormal)**.** Un conjunto $\{e_n\}_{n\geq 1} \subset H$ se llama ortonormal si

$$\langle e_n, e_m \rangle = \delta_{nm}, \quad \text{para todo } n, m,$$

donde $\delta_{nm}$ es el delta de Kronecker.

**Definición 2.4.13** (Span y clausura)**.** Dado un conjunto $A \subset H$, el subespacio

$$\operatorname{span}(A)$$

es el conjunto de todas las combinaciones lineales finitas de elementos de $A$. Su clausura en la norma de $H$ se denota por

$$\overline{\operatorname{span}}(A).$$

**Definición 2.4.14** (Núcleo de un operador)**.** El núcleo de un operador lineal $T : H \to H$ se define como

$$\ker(T) = \{x \in H : Tx = 0\}.$$

**Definición 2.4.15** (Suma directa ortogonal)**.** Sean $M, N \subset H$ subespacios cerrados. Se dice que

$$H = M \oplus N$$

si todo $x \in H$ puede escribirse de manera única como $x = m + n$ con $m \in M$, $n \in N$, y además $M \perp N$.

**Definición 2.4.16** (Base ortonormal completa)**.** Un conjunto ortonormal $\{e_n\} \subset H$ se llama base ortonormal completa si

$$\overline{\operatorname{span}}\{e_n\} = H.$$

**Teorema 2.4.17** (Teorema espectral para operadores compactos y autoadjuntos)**.** Sea $H$ un espacio de Hilbert y sea $T : H \to H$ un operador lineal. Si $T$ es compacto y autoadjunto, entonces se cumple lo siguiente:

1. Todos los valores propios de $T$ son reales.
2. Existe una sucesión (finita o infinita) de valores propios no negativos $\{\lambda_n\}_{n\geq 1}$ tal que $\lambda_n \to 0$ cuando $n \to \infty$, y cada valor propio distinto de cero tiene multiplicidad finita.
3. Existen vectores propios $\{e_n\}_{n\geq 1}$ asociados a $\{\lambda_n\}$ que forman un conjunto ortonormal en $H$.
4. Para todo $x \in H$ se tiene la expansión

$$Tx = \sum_{n=1}^{\infty} \lambda_n \langle x, e_n \rangle e_n,$$

donde la serie converge en la norma de $H$.

5. Además, si $\ker(T)$ denota el núcleo de $T$, entonces

$$H = \ker(T) \oplus \overline{\text{span}}\{e_n : n \geq 1\},$$

y los $e_n$ junto con una base ortonormal del núcleo de $T$ forman una base ortonormal completa de $H$.

*Demostración.* La demostración se divide en varios pasos.

**Paso 1: Los valores propios son reales.**

Sea $\lambda$ un valor propio de $T$ con vector propio asociado $x \neq 0$, es decir, $Tx = \lambda x$. Entonces,

$$\langle Tx, x \rangle = \lambda \langle x, x \rangle.$$

Como $T$ es autoadjunto,

$$\langle Tx, x\rangle = \langle x, Tx\rangle = \overline{\lambda}\langle x, x\rangle.$$

Dado que $\langle x, x\rangle > 0$, se concluye que $\lambda = \overline{\lambda}$, y por tanto $\lambda$ es real.

**Paso 2: Existencia de un valor propio no nulo.**

Supóngase que $T \neq 0$. Considérese el funcional

$$\varphi(x) = \langle Tx, x\rangle, \quad \|x\| = 1.$$

Como $T$ es compacto y autoadjunto sobre un espacio de Hilbert, el conjunto $\{Tx : \|x\| = 1\}$ es relativamente compacto, y por tanto $\varphi$ alcanza su máximo en la esfera unitaria. Sea $x_1$ tal que

$$\langle Tx_1, x_1\rangle = \max_{\|x\|=1} \langle Tx, x\rangle = \lambda_1.$$

Usando el método de variaciones, se demuestra que $x_1$ satisface

$$Tx_1 = \lambda_1 x_1,$$

por lo que $\lambda_1$ es un valor propio real. Además, $\lambda_1 \geq 0$.

**Paso 3: Construcción inductiva de valores propios.**

Sea $H_1 = \{x_1\}^{\perp}$. El subespacio $H_1$ es cerrado y $T$ deja invariante a $H_1$. La restricción $T|_{H_1}$ sigue siendo compacta y autoadjunta. Si $T|_{H_1} \neq 0$, el argumento anterior garantiza la existencia de un nuevo valor propio $\lambda_2$ con vector propio $x_2 \in H_1$.

Repitiendo este procedimiento se obtiene una sucesión (finita o infinita) de valores propios no negativos

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq 0,$$

con vectores propios ortonormales $\{e_n\}_{n\geq 1}$.

**Paso 4: Acumulación únicamente en cero y multiplicidad finita.**

Como $T$ es compacto, todo conjunto infinito de valores propios distintos de cero debe acu-

mularse en 0. Además, cada valor propio no nulo tiene multiplicidad finita; de lo contrario, se obtendría una sucesión ortonormal $\{x_n\}$ tal que $\|Tx_n\|$ no admite subsucesión convergente, contradiciendo la compacidad de $T$.

**Paso 5: Descomposición espectral.**

Sea

$$M = \overline{\text{span}}\{e_n : n \geq 1\}.$$

Se demuestra que $M^{\perp} = \ker(T)$ y que

$$H = \ker(T) \oplus M.$$

Para todo $x \in H$ se tiene

$$Tx = \sum_{n=1}^{\infty} \lambda_n \langle x, e_n \rangle e_n,$$

donde la convergencia es en la norma de $H$, como consecuencia de la ortonormalidad de $\{e_n\}$ y la compacidad de $T$.

**Paso 6: Base ortonormal completa.**

Finalmente, los vectores propios $\{e_n\}$ junto con una base ortonormal del núcleo $\ker(T)$ forman una base ortonormal completa de $H$.

Esto concluye la demostración.

□

**Teorema 2.4.18** (Teorema de Mercer)**.** Sea $K : X \times X \to \mathbb{R}$ un kernel simétrico y continuo en un espacio compacto $X \subset \mathbb{R}^n$. Entonces, $K$ es un kernel definido positivo si y sólo si existe una expansión espectral de la forma

$$K(x, z) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(x)\varphi_i(z),$$

donde:

1. $\{\varphi_i\}$ es un sistema ortonormal completo en $L^2(X)$,
2. $\lambda_i \geq 0$ para todo $i$,
3. la serie converge absolutamente y uniformemente en $X \times X$.

*Demostración.* La demostración hace uso de diversos conceptos en teoría de la medida; en particular, las medidas de Lebesgue, de Borel y de Radon. Se recomienda al lector consultar estas definiciones en [7, Capítulos 1 y 7].

Sea $X$ un espacio compacto de $\mathbb{R}^n$, es decir, cerrado y acotado, por el Teorema de Heine-Borel, y sea $\mu$ una medida de Borel finita con soporte $X$ (por ejemplo, la medida de Lebesgue restringida a $X$ o cualquier medida de Borel con $\mu(X) > 0$). Sea

$$K : X \times X \to \mathbb{R}$$

una función continua, simétrica ($K(x, y) = K(y, x)$) y definida positiva, es decir, para todo $m \in \mathbb{N}$, $x_1, \ldots, x_m \in X$ y $c_1, \ldots, c_m \in \mathbb{R}$,

$$\sum_{i,j} c_i c_j K(x_i, x_j) \geq 0.$$

Se desea probar que existe una expansión

$$K(x, z) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(x)\varphi_i(z)$$

que cumple con los incisos 1, 2 y 3 del enunciado.

Defínase el operador integral $T_K : L^2(X) \to L^2(X)$ por

$$(T_K f)(x) = \int_X K(x, y) f(y)\, d\mu(y).$$

Como $K$ es continua en el compacto $X \times X$, por el teorema del máximo existe

$$M := \sup_{(x,y)\in X\times X} |K(x, y)| < \infty.$$

Entonces, para todo $(x, y) \in X \times X$,

$$|K(x, y)|^2 \leq M^2.$$

Si $\mu$ es una medida de Borel (por ejemplo, una medida de Radon) sobre el compacto $X$, entonces $\mu(X) < \infty$. El producto de medidas satisface

$$(\mu \times \mu)(X \times X) = \mu(X)\, \mu(X) = \mu(X)^2 < \infty.$$

Por tanto,

$$\iint_{X\times X} |K(x, y)|^2\, d\mu(x)\, d\mu(y) \leq M^2\, (\mu \times \mu)(X \times X) = M^2\, \mu(X)^2 < \infty. \tag{2.4.1}$$

En consecuencia, $K \in L^2(X \times X)$.

Ahora, se prueba que $T_K$ es un operador compacto: Si $K \in L^2(X \times X)$ se define la norma de Hilbert-Schmidt como

$$\|T_K\|_{HS} := \left( \iint_{X\times X} |K(x, y)|^2\, d\mu(x)\, d\mu(y) \right)^{1/2},$$

la cual es finita por la desigualdad (2.4.1).

Además, por la simetría y continuidad del kernel, $T_K$ es autoadjunto:

$$\langle T_K f, g\rangle = \iint K(x,y) f(y) g(x)\, d\mu(y) d\mu(x) = \langle f, T_K g\rangle.$$

La positividad del kernel se traduce en positividad del operador: para todo $f \in L^2(X)$,

$$\langle T_K f, f\rangle = \iint K(x,y) f(y) \overline{f(x)}\, d\mu(y) d\mu(x) \geq 0.$$

Por el Teorema Espectral para operadores compactos autoadjuntos 2.4.17, existen una sucesión (finita o infinita) de autovalores reales $\{\lambda_n\}$ (contados con multiplicidad) y una base ortonormal de $L^2(X)$ formada por vectores propios del operador en el cierre del rango, tal que:

$$T_K e_n = \lambda_n e_n, \qquad n = 1, 2, \ldots,$$

con $\lambda_n \to 0$. Como $T_K$ es positivo, todos los $\lambda_n \geq 0$. Se puede tomar $\{e_n\}$ ortonormal en $L^2(X)$. Además, siendo $T_K$ compacto, la suma $\sum_n \lambda_n^2$ es finita, y en particular $\sum_n \lambda_n < \infty$ si $T_K$ es de traza (demostrado más adelante).

Dado que el operador es compacto, por el Teorema espectral 2.4.17, la integral kernel $K$ coincide (como elemento de $L^2(X \times X)$ con la medida producto $\mu \times \mu$) con la serie de núcleos

$$\sum_{n=1}^{\infty} \lambda_n e_n(x) e_n(y),$$

es decir, la serie converge en norma $L^2(X \times X)$ y, por tanto, se tiene la igualdad

$$K(x,y) = \sum_{n=1}^{\infty} \lambda_n e_n(x) e_n(y), \quad \text{para casi todo } (x,y) \in X \times X.$$

Ahora, se desea mejorar la igualdad casi en todas partes por igualdad puntual y, además, establecer la convergencia absoluta y uniforme en $X \times X$.

Observe que para $f \in L^2(X)$ la función $T_K f$ es continua en $X$. De hecho, por continuidad y

acotación de $K$ y compacidad de $X$, y con la desigualdad de Cauchy–Schwarz, se cumple:

$$|(T_K f)(x_1) - (T_K f)(x_2)| \leq \int_X |K(x_1, y) - K(x_2, y)|\,|f(y)|\,d\mu(y) \leq \|K(x_1, \cdot) - K(x_2, \cdot)\|_2 \|f\|_2,$$

y la continuidad de $x \mapsto K(x, \cdot)$ en $L^2(X)$ (por continuidad puntual y compacidad) da la continuidad de $T_K f$. En particular, si $e_n$ es autofunción (es decir, $T_K e_n = \lambda_n e_n$), entonces $e_n = \lambda_n^{-1} T_K e_n$ es continua siempre que $\lambda_n \neq 0$. Por tanto, todas las autofunciones asociadas a autovalores no nulos son continuas en $X$.

A continuación, defínase la aproximación parcial:

$$K_N(x, y) := \sum_{n=1}^{N} \lambda_n e_n(x) e_n(y).$$

Se quiere mostrar que $K_N \to K$ uniformemente en $X \times X$.

Obsérvese que para cualquier $N$, la diferencia $R_N := T_K - \sum_{n=1}^{N} \lambda_n \langle \cdot, e_n \rangle e_n$ es un operador positivo (ya que se obtiene con la diferencia de proyectos asociados a los primeros $N$ autovalores) y compacto. Su kernel es

$$r_N(x, y) := K(x, y) - K_N(x, y).$$

Para todo $f \in L^2(X)$,

$$\langle R_N f, f \rangle = \iint r_N(x, y) f(y) f(x)\,d\mu(y) d\mu(x) \geq 0.$$

En particular, se toma $f = \chi_A$ (función indicadora de un conjunto medible $A \subset X$) y dado que $r_N$ es simétrica, se obtiene

$$\int_A r_N(x, x)\,d\mu(x) = \langle R_N \chi_A, \chi_A \rangle \geq 0.$$

Así $r_N(x, x) \geq 0$ para casi todo $x$. Además, la *traza* del operador $T_K$ puede calcularse como

$$\operatorname{tr}(T_K) = \int_X K(x, x)\,d\mu(x) = \sum_{n=1}^{\infty} \lambda_n,$$

lo que implica (al restar las primeras $N$ componentes) que

$$\int_X r_N(x,x)\,d\mu(x) = \sum_{n=N+1}^{\infty} \lambda_n \xrightarrow[N\to\infty]{} 0.$$

Dado que $r_N(x,x) \geq 0$ casi en todos lados y su integral tiende a cero, se sigue que $r_N(x,x) \to 0$ en $L^1(X)$. Pero como cada $r_N(x,x)$ es continua en $x$ (pues $K$ y $K_N$ son continuas), la convergencia en $L^1$ junto con la positividad implica convergencia uniforme a 0 de $r_N(x,x)$ en $X$; en efecto, si una sucesión con elementos positivos tiene integrales que tienden a cero en un compacto, entonces el supremo de las integrales (llamada suprema) debe tender a cero; de lo contrario, existiría una subsucesión con suprema mayor o igual a un $\varepsilon > 0$ y por compacidad se podría extraer una subsucesión que contradice la convergencia de integrales.

Luego, se utiliza la desigualdad de Cauchy–Schwarz para controlar la diferencia fuera de la diagonal, esto es, para todo $x, y \in X$,

$$|r_N(x,y)|^2 \leq r_N(x,x)\, r_N(y,y).$$

Esta desigualdad sigue de la positividad del kernel $r_N$ aplicada a vectores de la forma $c_1\delta_x + c_2\delta_y$ o bien de la representación de $r_N$ como kernel de un operador positivo. Puesto que $r_N(x,x) \to 0$ uniformemente en $X$, se deduce que $r_N(x,y) \to 0$ uniformemente en $X \times X$. Por tanto

$$\sup_{x,y\in X} |K(x,y) - K_N(x,y)| \xrightarrow[N\to\infty]{} 0,$$

es decir, la convergencia de $K_N$ a $K$ es uniforme en $X \times X$. Además, la desigualdad

$$|K_N(x,y)| \leq \sum_{n=1}^{N} \lambda_n |e_n(x)||e_n(y)|$$

y la uniformidad anterior implican la convergencia absoluta de la serie; esto por la consideración del término sobrante y la convergencia de la diagonal.

Resta probar la ortonormalidad completa en $L^2(X)$ y determinar los signos de los autovalores. Por construcción, las funciones $e_n$ son ortonormales en $L^2(X)$ y generan la clausura del

rango de $T_K$. La positividad de $T_K$ garantiza $\lambda_n \geq 0$ para todo $n$. Si además $K$ es definida positiva en el sentido estricto (no degenerada), la nulidad del operador ocurre solo en el subespacio trivial; en cualquier caso, el núcleo se complementa ortogonalmente como en el enunciado.

Con todo lo anterior, se ha mostrado que existe una base ortonormal $\{e_n\}$ (las autofunciones correspondientes a los autovalores no nulos), con $\lambda_n \geq 0$, tal que

$$K(x, y) = \sum_{n=1}^{\infty} \lambda_n e_n(x) e_n(y),$$

donde la serie converge absolutamente y uniformemente en $X \times X$. Esto demuestra el Teorema de Mercer en la forma solicitada con $\varphi_i = e_i$.

□

**Definición 2.4.19** (Espacio de Hilbert de Reproducción)**.** Sea $X$ un conjunto no vacío y $\mathcal{H}$ un espacio de Hilbert de funciones $f : X \to \mathbb{R}$. El conjunto $\mathcal{H}$ es un *Espacio de Hilbert de Reproducción* (RKHS) si cumple:

1. Para todo $x \in X$, el funcional de evaluación

$$L_x : \mathcal{H} \to \mathbb{R}, \quad L_x(f) = f(x)$$

es lineal y continuo.

2. Existe una función $K : X \times X \to \mathbb{R}$, llamada *núcleo de reproducción*, tal que

$$f(x) = \langle f, K(\cdot, x) \rangle_{\mathcal{H}}, \quad \text{para toda } f \in \mathcal{H}, \text{ para todo } x \in X.$$

Ahora, surge la pregunta ¿qué funciones componen un RKHS? Sea $K : X \times X \to \mathbb{R}$ un núcleo (kernel) positivo definido. El espacio de Hilbert de reproducción asociado a $K$, denotado por $\mathcal{H}_K$, está formado por

$$\mathcal{H}_K = \overline{\left\{ f(\cdot) = \sum_{i=1}^{n} \alpha_i K(\cdot, x_i) \;\middle|\; n \in \mathbb{N},\ \alpha_i \in \mathbb{R},\ x_i \in X \right\}},$$

donde la clausura se toma respecto a la norma de Hilbert inducida por $K$. Es decir,

$$f(x) = \sum_{i=1}^{\infty} \alpha_i \, K(x_i, x),$$

para alguna sucesión $\{\alpha_i\}$ y puntos $\{x_i\}$, con condiciones de convergencia impuestas por la norma de $\mathcal{H}_K$. En resumen, las funciones de un RKHS son combinaciones finitas (y límites de combinaciones finitas) de traslaciones del kernel $K$. Equivalentemente, toda función $f \in \mathcal{H}_K$ puede escribirse como

$$f(x) = \langle w, \Phi(x) \rangle_{\mathcal{H}},$$

para algún $w$ en un espacio de Hilbert adecuado, donde $\Phi$ es la función inducida por el kernel. El siguiente teorema muestra que todo núcleo definido positivo induce un espacio de Hilbert de reproducción. El ejemplo que lo acompaña muestra los kernels con sus correspondientes espacios de Hilbert generados.

**Teorema 2.4.20** (Moore–Aronszajn, [4])**.** Sea $X$ un conjunto y $K : X \times X \to \mathbb{R}$ una función simétrica y definida positiva, es decir, $K$ es una función kernel. Entonces existe un *único* (salvo isometrías) espacio de Hilbert $\mathcal{H}$ formado por funciones $f : X \to \mathbb{R}$ cuyo núcleo reproductor es $K$, es decir, se cumple

1. para todo $x \in X$, la función $K(\cdot, x)$ pertenece a $\mathcal{H}$;
2. (propiedad de reproducción) para todo $f \in \mathcal{H}$ y todo $x \in X$,

$$f(x) = \langle f, \, K(\cdot, x) \rangle_{\mathcal{H}}.$$

Tal espacio $\mathcal{H}$ se llama el *espacio de Hilbert de reproducción* asociado a $K$ y se abrevia como RKHS.

*Demostración.* La demostración se divide en los siguientes pasos:

1. Construcción del espacio prehilbertiano: Se define el espacio vectorial de combinaciones

finitas del conjunto, llamado el generado del conjunto $\{ K(\cdot, x) : x \in X \}$ y se define como

$$\mathcal{H}_0 := \text{gen}\{ K(\cdot, x) : x \in X \}.$$

Un elemento arbitrario de $\mathcal{H}_0$ tiene la forma

$$f = \sum_{i=1}^{n} \alpha_i K(\cdot, x_i), \qquad \alpha_i \in \mathbb{R},\ x_i \in X.$$

Así también, se define un producto sesquilineal en $\mathcal{H}_0$ por

$$\left\langle \sum_{i=1}^{n} \alpha_i K(\cdot, x_i),\ \sum_{j=1}^{m} \beta_j K(\cdot, y_j) \right\rangle_{\mathcal{H}_0} := \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j\, K(x_i, y_j). \tag{2.4.2}$$

La buena definición de la construcción 2.4.2 está dada en forma de lema:

**Lema 2.4.21** (Buena definición y positividad)**.** La forma (2.4.2) está bien definida (independiente de la representación), es simétrica, y verifica

$$\langle f, f \rangle_{\mathcal{H}_0} \geq 0,$$

para todo $f \in \mathcal{H}_0$. Además, si $\langle f, f \rangle_{\mathcal{H}_0} = 0$, implica que la función $f$ es la función cero sobre $X$.

*Demostración.* La sesquilinealidad y la simetría siguen por la definición y por la propiedad $K(x, y) = K(y, x)$. Para la positividad, considere $f = \sum_{i=1}^{n} \alpha_i K(\cdot, x_i)$; entonces

$$\langle f, f \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \overline{\alpha_j}\, K(x_i, x_j) \geq 0$$

por la definición de positividad de $K$.

Si $\langle f, f \rangle_{\mathcal{H}_0} = 0$, entonces la forma cuadrática anterior se anula. Por la positividad de la

matriz de Gram $[K(x_i, x_j)]$, esto implica que para todo $y \in X$,

$$0 = \langle f, K(\cdot, y)\rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \alpha_i \, K(x_i, y).$$

Pero el lado derecho es exactamente $f(y)$; por tanto $f(y) = 0$ para todo $y \in X$, es decir, $f \equiv 0$ como función en $X$. Esto garantiza que la norma asociada separa puntos y la definición del producto interior no depende de la representación: si dos combinaciones finitas representan la misma función, su diferencia tiene norma cero y por tanto la forma da el mismo valor. □

En consecuencia del Lema 2.4.21, $(\mathcal{H}_0, \langle\cdot,\cdot\rangle_{\mathcal{H}_0})$ es un prehilbertiano.

2. Propiedad de reproducción en $\mathcal{H}_0$: Sea $f = \sum_{i=1}^{n} \alpha_i K(\cdot, x_i) \in \mathcal{H}_0$ y sea $x \in X$. Entonces

$$\langle f, K(\cdot, x)\rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \alpha_i \, K(x_i, x),$$

y por definición de $f$ también

$$f(x) = \sum_{i=1}^{n} \alpha_i \, K(x_i, x).$$

Por tanto, para todo $f \in \mathcal{H}_0$ y todo $x \in X$ se cumple la identidad de reproducción

$$f(x) = \langle f, K(\cdot, x)\rangle_{\mathcal{H}_0}.$$

En particular, $K(\cdot, x) \in \mathcal{H}_0$ para todo $x \in X$.

3. Obtención del RKHS: Sea $\mathcal{H}$ la completación de $\mathcal{H}_0$ respecto a la norma inducida por $\langle\cdot,\cdot\rangle_{\mathcal{H}_0}$. Dado que la identidad de reproducción en $\mathcal{H}_0$ expresa cada evaluación $f \mapsto f(x)$ como el producto interior $\langle f, K(\cdot, x)\rangle$, y puesto que $K(\cdot, x) \in \mathcal{H}_0$ (por construcción), la evaluación es continua en la norma de $\mathcal{H}_0$ y, por continuidad, la identidad se extiende a

toda $\mathcal{H}$. Más concretamente, si $f_n \in \mathcal{H}_0$ converge en norma a $f \in \mathcal{H}$, entonces

$$f_n(x) = \langle f_n, K(\cdot, x)\rangle_{\mathcal{H}_0} \to \langle f, K(\cdot, x)\rangle_{\mathcal{H}},$$

y como $f_n(x) \to f(x)$ puntualmente, se obtiene

$$f(x) = \langle f, K(\cdot, x)\rangle_{\mathcal{H}}, \quad \text{para todo } x \in X.$$

Así $\mathcal{H}$ es un espacio de Hilbert de funciones en $X$ con $K$ como núcleo de reproducción.

4. Unicidad (salvo isometrías con los kernels $K(\cdot, x)$ fijos): Supóngase que $\mathcal{G}$ es otro espacio de Hilbert de funciones en $X$ con la propiedad de que $K(\cdot, x) \in \mathcal{G}$ y que

$$g(x) = \langle g, K(\cdot, x)\rangle_{\mathcal{G}} \quad \text{para todo } g \in \mathcal{G},\ x \in X.$$

Considérese los subespacios generados por los núcleos:

$$\mathcal{H}_0 = \operatorname{span}\{K(\cdot, x) : x \in X\} \subset \mathcal{H}, \qquad \mathcal{G}_0 = \operatorname{span}\{K(\cdot, x) : x \in X\} \subset \mathcal{G}.$$

Se define la aplicación lineal $T : \mathcal{H}_0 \to \mathcal{G}_0$ como

$$T\Big(\sum_{i=1}^{n} \alpha_i K(\cdot, x_i)\Big) := \sum_{i=1}^{n} \alpha_i K(\cdot, x_i),$$

la suma del lado izquierdo es elemento de $\mathcal{H}_0$ y la del derecho es elemento de $\mathcal{G}_0$. Se usa la propiedad de reproducción en ambos espacios y se verifica que $T$ es isometría en $\mathcal{H}_0$:

$$\Big\| \sum_{i=1}^{n} \alpha_i K(\cdot, x_i)\Big\|_{\mathcal{H}}^2 = \sum_{i,j} \alpha_i \alpha_j\, K(x_i, x_j) = \Big\| \sum_{i=1}^{n} \alpha_i K(\cdot, x_i)\Big\|_{\mathcal{G}}^2.$$

Por tanto, $T$ preserva normas y se extiende de manera única (por completación) a una isometría lineal entre $\mathcal{H}$ y la clausura de $\mathcal{G}_0$ (que es todo $\mathcal{G}$ si $\mathcal{G}$ está generado por los núcleos). Esta isometría fija cada $K(\cdot, x)$ y, por tanto, identifica $\mathcal{H}$ y $\mathcal{G}$ como RKHS idénti-

cos. Esto demuestra la unicidad.

□

**Ejemplo 2.4.22.**

1. **Kernel lineal:** $K(x,z) = \langle x, z\rangle$. El RKHS asociado es $\mathbb{R}^n$ con el producto interno usual. En efecto, sea $X = \mathbb{R}^n$ (o un subconjunto que genere $\mathbb{R}^n$) y considérese el kernel

$$K(x,z) = \langle x, z\rangle_{\mathbb{R}^n}.$$

A continuación se demuestra que el espacio de Hilbert con kernel reproductor asociado a $K$ es isométricamente $\mathbb{R}^n$ con su producto interno usual.

Para cada $a \in \mathbb{R}^n$, la sección del kernel es

$$K(\cdot, a)(x) = K(x,a) = \langle x, a\rangle.$$

Toda combinación lineal finita de tales secciones tiene la forma

$$f(x) = \sum_{i=1}^{m} \alpha_i K(x, a_i) = \sum_{i=1}^{m} \alpha_i \langle x, a_i\rangle = \Big\langle x, \sum_{i=1}^{m} \alpha_i a_i \Big\rangle.$$

Por tanto, cualquier función en el subespacio generado por las secciones es lineal en $x$.

Sea $\mathcal{S} := \operatorname{span}\{K(\cdot, a) : a \in \mathbb{R}^n\}$. Se define su producto interno imponiendo la propiedad reproductora:

$$\langle K(\cdot, a), K(\cdot, b)\rangle_{\mathcal{H}} := K(a,b) = \langle a, b\rangle_{\mathbb{R}^n}.$$

Por bilinealidad y positividad, esto define un prehilbertiano en $\mathcal{S}$. Se define la aplicación lineal

$$\Phi : \mathbb{R}^n \to \mathcal{S}, \qquad \Phi(w) = K(\cdot, w) = \langle \cdot, w\rangle.$$

Para cualesquiera $u, v \in \mathbb{R}^n$,

$$\langle \Phi(u), \Phi(v) \rangle_{\mathcal{H}} = \langle K(\cdot, u), K(\cdot, v) \rangle_{\mathcal{H}} = K(u, v) = \langle u, v \rangle_{\mathbb{R}^n}.$$

Así, $\Phi$ es una isometría lineal. También es inyectiva; si $\Phi(w) = 0$, entonces $\langle x, w \rangle = 0$ para todo $x$, lo que implica $w = 0$. La dimensión de $\mathcal{S}$ es $n$, por lo que $\Phi$ es sobre.
El subespacio $\mathcal{S}$ es de dimensión finita, por lo que es completo. La RKHS $\mathcal{H}_K$ se define como la clausura de $\mathcal{S}$, pero al ser éste cerrado se obtiene:

$$\mathcal{H}_K = \mathcal{S}.$$

Cada $f \in \mathcal{H}_K$ tiene la forma

$$f(x) = \langle x, w \rangle, \qquad \text{para un único } w \in \mathbb{R}^n.$$

La norma inducida coincide con la euclidiana:

$$\|f\|_{\mathcal{H}_K} = \|w\|_{\mathbb{R}^n}.$$

La propiedad reproductora se verifica directamente:

$$f(a) = \langle a, w \rangle = \langle f, K(\cdot, a) \rangle_{\mathcal{H}_K}.$$

Por lo tanto, se concluye que el RKHS asociado al núcleo $K(x, z) = \langle x, z \rangle$ es exactamente $\mathbb{R}^n$ con su producto interno usual, identificado mediante la correspondencia

$$w \longmapsto f_w(x) = \langle x, w \rangle.$$

2 **Kernel polinomial:** $K(x,z) = (\langle x,z\rangle + c)^d$. El RKHS asociado corresponde al espacio de todos los polinomios de grado $\leq d$ en las variables de $x$. Por el teorema de Moore-Aronszajn, el RKHS asociado a $K$ está dado por la clausura del espacio generado por las funciones $K(\cdot,z)$, $z \in \mathbb{R}^n$.

Observe primero que

$$\langle x,z\rangle = \sum_{i=1}^{n} x_i z_i.$$

Por el teorema multinomial, se tiene

$$(\langle x,z\rangle + c)^d = \sum_{k=0}^{d} \binom{d}{k} c^{d-k} \left(\sum_{i=1}^{n} x_i z_i\right)^k = \sum_{k=0}^{d} \binom{d}{k} c^{d-k} \sum_{|\alpha|=k} \frac{k!}{\alpha!} x^\alpha z^\alpha,$$

donde $\alpha = (\alpha_1,\ldots,\alpha_n) \in \mathbb{N}^n$ es un multiíndice, $|\alpha| = \alpha_1 + \cdots + \alpha_n$, $\alpha! = \alpha_1!\cdots\alpha_n!$ y

$$x^\alpha = x_1^{\alpha_1}\cdots x_n^{\alpha_n}.$$

Reordenando términos, puede escribirse

$$K(x,z) = \sum_{|\alpha|\leq d} a_\alpha\, x^\alpha z^\alpha,$$

con coeficientes $a_\alpha > 0$ que dependen únicamente de $d$ y $c$.

Defina entonces el mapeo de características

$$\phi : \mathbb{R}^n \to \mathbb{R}^N, \qquad \phi(x) = \left(\sqrt{a_\alpha}\, x^\alpha\right)_{|\alpha|\leq d},$$

donde $N = \binom{n+d}{d}$ es el número de monomios de grado menor o igual que $d$. Con esta definición se verifica que

$$K(x,z) = \langle \phi(x), \phi(z)\rangle_{\mathbb{R}^N}.$$

Por la caracterización del RKHS asociada a un kernel de producto interno, toda función $f$ en el RKHS tiene la forma

$$f(x) = \langle w, \phi(x) \rangle = \sum_{|\alpha| \leq d} w_\alpha x^\alpha,$$

para algún $w \in \mathbb{R}^N$. En consecuencia, $f$ es un polinomio de grado menor o igual que $d$. Recíprocamente, todo polinomio de grado menor o igual que $d$ puede escribirse como una combinación lineal finita de los monomios $x^\alpha$, $|\alpha| \leq d$, y por tanto pertenece al espacio generado por $\phi$.

Se concluye que el RKHS asociado al kernel polinomial $(\langle x, z \rangle + c)^d$ coincide con el espacio de todos los polinomios en $n$ variables de grado a lo más $d$, equipado con el producto interno inducido por $K$.

Así, se establece que la elección de algún kernel mapea los datos a su correspondiente espacio de Hilbert de reproducción. Sólo resta demostrar que, sobre esos espacios, las dos clases de datos se pueden separar mediante un hiperplano en el espacio de características, que se corresponde con el límite no lineal en el espacio de entrada; esto se establece mediante el teorema:

**Teorema 2.4.23** (Separabilidad mediante aumento de dimensionalidad)**.** Sea $\{(x_i, y_i)\}_{i=1}^m$ un conjunto de datos con $x_i \in \mathbb{R}^n$ y etiquetas $y_i \in \{-1, +1\}$. Si los datos no son linealmente separables en el espacio de entrada $\mathbb{R}^n$, entonces, mediante un mapeo no lineal

$$\psi : \mathbb{R}^n \to \mathbb{R}^N, \quad N \gg n,$$

existe con alta probabilidad un hiperplano en el espacio transformado tal que

$$y_i \left( \langle w, \psi(x_i) \rangle + b \right) > 0, \quad \text{para todo } i = 1, \cdots, m.$$

En particular, si $\psi$ corresponde a un *kernel definido positivo*, los datos se vuelven siempre separables en el espacio de Hilbert de reproducción asociado (RKHS).

*Demostración.* Sea $\psi : \mathbb{R}^n \to \mathbb{R}^N$ un mapeo, y considérese la matriz de Gram $K \in \mathbb{R}^{m\times m}$ asociada a los puntos $\{x_i\}_{i=1}^m$:

$$K_{ij} \;=\; \langle \psi(x_i), \psi(x_j)\rangle, \qquad i, j = 1, \ldots, m.$$

La invertibilidad de $K$ implica que los vectores $\{\psi(x_i)\}_{i=1}^m$ son linealmente independientes. En efecto, si

$$\sum_{i=1}^{m} \alpha_i \, \psi(x_i) \;=\; 0,$$

entonces, al tomar producto interno con cada $\psi(x_j)$, se obtiene

$$\sum_{i=1}^{m} \alpha_i \, K_{ij} \;=\; 0, \qquad j = 1, \ldots, m.$$

En notación matricial, esto se escribe como $K\alpha = 0$. Si $K$ es invertible, la única solución es $\alpha = 0$, por lo que las imágenes $\psi(x_i)$ son linealmente independientes.

Sea el vector de etiquetas $y = (y_1, \ldots, y_m)^t \in \mathbb{R}^m$. Considérese el sistema lineal

$$Kc = y,$$

que tiene única solución $c = K^{-1}y \in \mathbb{R}^m$. Ahora, se define el vector

$$w \;:=\; \sum_{j=1}^{m} c_j \, \psi(x_j) \;\in\; \mathbb{R}^N.$$

Entonces, para cada $i = 1, \ldots, m$,

$$\langle w, \psi(x_i)\rangle = \Big\langle \sum_{j=1}^{m} c_j \psi(x_j),\; \psi(x_i) \Big\rangle = \sum_{j=1}^{m} c_j \, \langle \psi(x_j), \psi(x_i)\rangle = (Kc)_i = y_i.$$

Al tomar $b = 0$ se obtiene

$$y_i(\langle w, \psi(x_i)\rangle + b) = y_i y_i = y_i^2 = 1 > 0, \qquad \text{para toda } i = 1, \ldots, m, \tag{2.4.3}$$

esto es dado que las etiquetas $y_i \in \{-1, 1\}$. Es decir, en la desigualdad (2.4.3) se demuestra que existe un hiperplano que separa las clases. Por tanto, cuando $K$ es invertible, se ha construido explícitamente un hiperplano (definido por $w, b$) que separa los datos, es decir, los datos son linealmente separables.

Ahora, resta probar por qué $K$ suele ser invertible bajo un mapeo no lineal $\psi$. En este sentido, se analizan dos casos:

1. Si $\psi$ es un mapeo *aleatorio* a una dimensión muy alta (por ejemplo, $\psi$ construida con características aleatorias), entonces los vectores $\psi(x_i)$ estarán en una posición general con una alta probabilidad, es decir, no hay alineaciones, coplanariedades o coincidencias exactas entre los puntos, y la matriz de Gram será invertible. Salvo que el mapeo sea degenerado, es decir, envíe a todos los puntos en uno solo. Lo cual en la práctica no podría ser elegido. Es por esto que enunciados informales dicen "con alta probabilidad": un mapeo suficientemente variado a alta dimensión evita relaciones lineales entre las imágenes de los puntos de entrada.

2. Si $\psi$ proviene de un kernel definido positivo $K(\cdot, \cdot)$ y dicho kernel es estrictamente definido positivo sobre el conjunto de puntos $\{x_i\}$, entonces la matriz de Gram $K_{ij} = K(x_i, x_j)$ es *definida positiva*, en particular, invertible. En ese caso, el argumento del caso 1 se aplica y garantiza separación perfecta en el RKHS asociado.

Combinando ambos pasos, se concluye que, tras un mapeo no lineal suficientemente variado (o al usar un kernel estrictamente definido positivo), existe un vector $w$ (con $b = 0$ en la construcción anterior) que separa todas las etiquetas:

$$y_i(\langle w, \psi(x_i)\rangle + b) > 0, \qquad i = 1, \ldots, m.$$

Esto prueba el teorema. □

La introducción de un kernel con su correspondiente espacio de características (espacio de Hilbert de reproducción) se conoce como "sustitución del kernel". A la luz del teorema de separabilidad mediante aumento de dimensionalidad (Teorema 2.4.23), se introduce un kernel $K(x, y)$; de esta manera, la formulación de clasificación binaria en la ecuación (2.2.13) se transforma en maximizar el problema dual:

$$\text{Maximizar} \quad \mathcal{W}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \tag{2.4.4}$$

$$\text{sujeto a} \quad \alpha_i \geq 0 \tag{2.4.5}$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0. \tag{2.4.6}$$

La función objetivo es cóncava por la Proposición 2.2.7, por lo que se puede utilizar el método del gradiente ascendente para hallar el vector $\alpha$. Nótese que es casi el mismo problema que en el caso linealmente separable, la diferencia es que en la función (2.4.4) se utiliza el kernel.

Ahora bien, el sesgo $b$ se calcula como lo muestra la proposición siguiente.

**Proposición 2.4.24.** Sean los valores $\alpha_i \geq 0$ obtenidos al resolver el problema (2.4.4). El sesgo $b$ se calcula como

$$b = -\frac{1}{2}\left[ \max_{\{i|y_i=-1\}} \left( \sum_{j=1}^{m} \alpha_j y_j K(x_i, x_j) \right) + \min_{\{i|y_i=+1\}} \left( \sum_{j=1}^{m} \alpha_j y_j K(x_i, x_j) \right) \right]. \tag{2.4.7}$$

*Demostración.* Nótese que para un dato con $y_i = 1$ se tiene que:

$$\min_{\{i|y_i=+1\}} [\langle w, x_i \rangle + b] = \min_{\{i|y_i=+1\}} \left[ \sum_{j=1}^{m} \alpha_j y_j K(x_i, x_j) \right] + b = 1.$$

Se toma otro punto para el cual $y_i = -1$. Así, se tiene que:

$$\max_{\{i|y_i=-1\}} [\langle w, x_i \rangle + b] = \max_{\{i|y_i=-1\}} \left[ \sum_{j=1}^{m} \alpha_j y_j K(x_i, x_j) \right] + b = -1.$$

Se suman estas dos expresiones se obtiene que:

$$\max_{\{i|y_i=-1\}}\left[\sum_{j=1}^{m}\alpha_j y_j K(x_i,x_j)\right]+b+\min_{\{i|y_i=+1\}}\left[\sum_{j=1}^{m}\alpha_j y_j K(x_i,x_j)\right]+b=-1+1.$$

Se despeja $b$ de esta ecuación:

$$\max_{\{i|y_i=-1\}}\left[\sum_{j=1}^{m}\alpha_j y_j K(x_i,x_j)\right]+\min_{\{i|y_i=+1\}}\left[\sum_{j=1}^{m}\alpha_j y_j K(x_i,x_j)\right]=-2b.$$

En consecuencia,

$$b=-\frac{1}{2}\left[\max_{\{i|y_i=-1\}}\left(\sum_{j=1}^{m}\alpha_j y_j K(x_i,x_j)\right)+\min_{\{i|y_i=+1\}}\left(\sum_{j=1}^{m}\alpha_j y_j K(x_i,x_j)\right)\right]. \tag{2.4.8}$$

□

Para un nuevo vector de entrada $z$, su clase predicha está dada por la función:

$$\phi(z)=\text{signo}\left(\sum_{i=1}^{m}\alpha_i^* y_i K(x_i,z)+b^*\right), \tag{2.4.9}$$

donde $b^*$ denota el sesgo óptimo y $\alpha^*$ denota los máximos multiplicadores de Lagrange.

El espacio de características es un espacio vectorial, (el cual es un espacio de Hilbert) al que los datos originales se proyectan mediante una transformación $\psi$, con el propósito de que en ese espacio los datos sean separables mediante un hiperplano. Cuando el hiperplano de margen máximo es encontrado, solo los puntos más cercanos al hiperplano tienen valores $\alpha_i^* > 0$ y estos puntos son llamados vectores soporte. Todos los demás puntos tienen valores $\alpha_i^* = 0$, y la función de decisión es independiente de estas muestras, como se establece en la siguiente proposición.

**Proposición 2.4.25.** Si $x_i$ es un vector soporte, entonces $\alpha_i > 0$. En caso contrario, $\alpha_i = 0$. En consecuencia, si se elimina uno de los vectores que no son vectores soporte, el hiperplano de separación no se ve afectado, es decir, la formulación del hiperplano de separación sigue siendo la misma.

*Demostración.* La solución del problema de optimización con restricciones satisface las condiciones de Karush-Kuhn-Tucker (KKT) por el Teorema 1.2.38. Una de las condiciones KKT es la complementariedad, es decir, para el punto óptimo del problema de optimización de SVM, y el vector de parámetros óptimo $(w^\star, b^\star, \alpha^\star)$, se cumple que

$$\alpha_i^\star g(x^\star) = 0,$$
$$\alpha_i^\star (y_i(\langle w^\star, x_i\rangle + b^\star) - 1) = 0.$$

Donde cada $\alpha_i^\star \geq 0$. Nótese que para cada vector se cumple que $y_i(\langle w, x_i\rangle + b) \geq 1$. Si $x_i$ es vector no soporte, se vuelve una desigualdad estricta, $y_i(\langle w, x_i\rangle + b) > 1$, es decir, $y_i(\langle w, x_i\rangle + b) - 1 > 0$, por lo tanto, $\alpha_i^\star = 0$.

En dado caso de que $x_i$ sea un vector soporte, se tiene $y_i(\langle w, x_i\rangle + b) = 1$, lo que implica que $\alpha_i^\star > 0$. □

Los datos con valores $\alpha_i^*$ relativamente grandes tienen una influencia significativa en la orientación del hiperplano y en la función de decisión; podrían ser datos que resulten correctamente clasificados o no al final del entrenamiento. Por lo que se deben examinar para ver si deben descartarse, a este proceso se le denomina *limpieza de datos*. Un valor atípico correctamente etiquetado puede tener una influencia indebida, por lo que lo mejor es reducir su impacto mediante el uso de un método de margen suave como el mostrado en la sección 2.3.

A continuación se muestra un ejemplo de la implementación de SVM para clasificación utilizando kernel y aumento de dimensionalidad. Considérese el siguiente conjunto de datos en $\mathbb{R}^2$:

| $i$ | $x_i$ | $y_i$ |
|---|---|---|
| 1 | $(1,1)$ | $+1$ |
| 2 | $(-1,-1)$ | $+1$ |
| 3 | $(1,-1)$ | $-1$ |
| 4 | $(-1,1)$ | $-1$ |

Graficamente se pueden ver estos datos en la Figura 2.9.

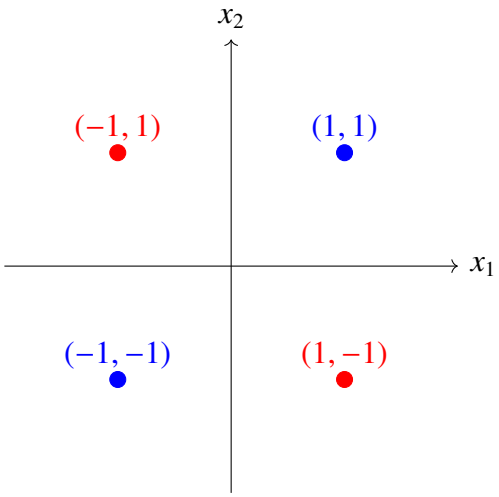


**Figura 2.9:** Puntos de entrenamiento.

Este conjunto de datos no es linealmente separable en el espacio original. Ahora, se utiliza el *kernel polinomial de grado dos* definido por:

$$K(x,z) = (\langle x,z\rangle)^2.$$

Luego, se desea hallar la transformación no lineal $\psi : \mathbb{R}^2 \to \mathbb{R}^3$ correspondiente al kernel $K$: sea $x = (x_1,x_2)$ y $z = (z_1,z_2) \in \mathbb{R}^2$, entonces

$$\begin{aligned} K(x,z) &= (\langle x,z\rangle)^2 \\ &= (x_1z_1 + x_2z_2)^2 \\ &= x_1^2z_1^2 + 2x_1z_1x_2z_2 + x_2^2z_2^2. \end{aligned}$$

Esta expresión puede escribirse como un producto interno en $\mathbb{R}^3$:

$$K(x, z) = \langle \psi(x), \psi(z) \rangle = x_1^2 z_1^2 + 2x_1 z_1 x_2 z_2 + x_2^2 z_2^2,$$

donde la transformación de características es:

$$\psi(x) = (x_1^2,\ \sqrt{2}x_1 x_2, x_2^2),$$

que es la transformación deseada. Así, el problema dual de la SVM con kernel se escribe como:

$$\begin{aligned} \text{Maximizar} \quad & \mathcal{W}(\alpha) = \sum_{i=1}^{4} \alpha_i - \frac{1}{2} \sum_{i=1}^{4} \sum_{j=1}^{4} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{sujeto a} \quad & \sum_{i=1}^{4} \alpha_i y_i = 0, \\ & 0 \leq \alpha_i. \end{aligned}$$

Se calculan los valores del kernel:

$$K = \begin{pmatrix} 4 & 4 & 0 & 0 \\ 4 & 4 & 0 & 0 \\ 0 & 0 & 4 & 4 \\ 0 & 0 & 4 & 4 \end{pmatrix}.$$

Y se define la matriz

$$Q_{ij} = y_i y_j K(x_i, x_j),$$

con

$$y = (1, 1, -1, -1),$$

con lo que se obtiene:

$$Q = \begin{pmatrix} 4 & 4 & 0 & 0 \\ 4 & 4 & 0 & 0 \\ 0 & 0 & 4 & 4 \\ 0 & 0 & 4 & 4 \end{pmatrix},$$

de modo que la función dual $\mathcal{W}$ se escribe como

$$\mathcal{W}(\alpha) = \mathbf{1}^t\alpha - \frac{1}{2}\alpha^t Q\alpha,$$

Y su gradiente respecto a $\alpha$ es

$$\nabla\mathcal{W} = \mathbf{1} - Q\alpha.$$

El método de gradiente ascendente se define por

$$\alpha^{(k+1)} = \alpha^{(k)} + \eta(\mathbf{1} - Q\alpha^{(k)}),$$

donde $\eta > 0$ es el tamaño de paso.

Como valores iniciales para comenzar a iterar el método de gradiente ascendente se toman los siguientes valores:

$$\alpha^{(0)} = (0, 0, 0, 0).$$

Tras varias iteraciones, el método converge a

$$\alpha^\star = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix},$$

Para hallar $b$, se utiliza lo establecido en la Proposición 2.4.24 con el conjunto de índices $I = \{1, 2, 3, 4\}$. Sean

$$\text{máx}\{i \in I : y_i = -1\} = 4 \qquad \text{y} \qquad \text{mín}\{i \in I : y_i = 1\} = 1. \tag{2.4.10}$$

luego

$$\sum_{j=1}^{4} \alpha_j y_j K(x_4, x_j) = \frac{1}{4}(0 + 0 - 4 - 4) = -2 \tag{2.4.11}$$

y

$$\sum_{j=1}^{4} \alpha_j y_j K(x_1, x_j) = \frac{1}{4}(4 + 4 + 0 + 0) = 2. \tag{2.4.12}$$

Finalmente se calcula el promedio de los resultados (2.4.11) y (2.4.12):

$$b = -\frac{1}{2}(2 - 2) = 0. \tag{2.4.13}$$

Sólo resta sustituir los valores óptimos en la función de decisión, con lo que se tiene:

$$\begin{aligned} \phi(x) &= \text{signo}\left(\sum_{i=1}^{4} \frac{1}{4} y_i (\langle x_i, x \rangle)^2\right) \\ &= 2uv. \end{aligned}$$

Para un nuevo vector de entrada $x = (u, v) \in \mathbb{R}^2$

Como se observa, en el espacio original, los datos no son separables linealmente. Sin embargo, al aplicar el kernel polinomial, la SVM encuentra un hiperplano en $\mathbb{R}^3$ que induce un hiperplano de separación no lineal en $\mathbb{R}^2$.

Este ejemplo ilustra el llamado *kernel trick*: no es necesario conocer explícitamente la transformación $\psi$, basta con evaluar productos internos mediante el kernel.

CAPÍTULO 3

# Máquina de soporte vectorial para regresión

Las máquinas de soporte vectorial también pueden ser utilizadas en problemas de regresión mediante la introducción de una función de pérdida alternativa, la cual se modifica para incluir una medida de distancia; esta función de pérdida se denomina $\varepsilon$-*insensitiva* y será introducida en esta sección. Estos modelos son llamados **SVMR (Support Vector Machine for Regression)** y son una herramienta útil para el análisis predictivo de datos, donde se tiene el problema de aproximar a un conjunto de entrenamiento. Si la regresión es lineal, se denominan máquinas de soporte vectorial para regresión lineal (Linear-SVMR); este tipo de modelos ha sido utilizado para múltiples problemas de distintas áreas, tales como la estimación de parámetros petrofísicos, diseño de antenas, predicción de hipertensión arterial, entre muchos otros (véase [3, 5, 16, 26]). Cabe resaltar que en la parte final del capítulo se encontrará el modelo SVM para regresión por mínimos cuadrados, que se sabe que es la base para el modelo SVM en su versión cuántica.

## 3.1. Formulación lineal y no lineal.

Para formular el modelo Linear-SVMR, considérese el conjunto de entrenamiento

$$\{(x_1, y_1), ..., (x_m, y_m)\} \subset X \times \mathbb{R},$$

donde $X$ denota el espacio de entrada de la muestra, para este caso de estudio se toma $X = \mathbb{R}^n$. El objetivo es encontrar una función $f$ que tenga a lo más un valor $\varepsilon$ de desviación de los valores objetivos reales $y_i$ para todos los datos de entrenamiento, es decir, que los errores sean menores que una tolerancia $\varepsilon$. En primera instancia, la función planteada es de tipo lineal, $\overline{f} : X \to \mathbb{R}$, la cual se formula de la siguiente manera:

$$\overline{f}(x) = \langle w, x \rangle + b. \tag{3.1.1}$$

El objetivo es minimizar la norma del vector para que la función de regresión sea lo menos inclinada posible, es decir, minimizar $\|w\|^2 = \langle w, w \rangle$. A esto se le conoce como "planitud de la recta".

El problema escrito como un problema de optimización convexa queda como sigue

$$\text{Minimizar} \quad \frac{1}{2}\|w\|^2, \tag{3.1.2}$$

$$\text{sujeto a} \quad y_i - \langle w, x_i \rangle - b \leq \varepsilon; \tag{3.1.3}$$

$$\langle w, x_i \rangle + b - y_i \leq \varepsilon. \tag{3.1.4}$$

El supuesto es que existe una función $\overline{f}$ que se aproxima a todos los pares $(x_i, y_i)$ con precisión $\varepsilon$, es decir $|\overline{f}(x_i) - y_i| \leq \varepsilon$. En otras palabras, que el problema de optimización convexa es factible; sin embargo, este puede no ser el caso en general. En esta situación se suelen permitir algunos errores mediante la introducción de variables de holgura $\xi_i$, $\xi_i^*$, de forma análoga a la función de "margen suave", las cuales fueron usadas por Vapnik y Cortes en su artículo de Máquinas de Soporte Vectorial de 1995 (véase [13]). Así es como se obtiene la formulación de

Vapnik:

$$\text{Minimizar} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*), \tag{3.1.5}$$

$$\begin{aligned} \text{sujeto a} \quad & y_i - \langle w, x_i\rangle - b \leq \varepsilon + \xi_i; \\ & \langle w, x_i\rangle + b - y_i \leq \varepsilon + \xi_i^*; \\ & \xi_i,\ \xi_i^* \geq 0. \end{aligned} \tag{3.1.6}$$

La constante $C > 0$ compensa la planitud de $f$, es decir, el grado en que la función se curva en ciertas direcciones, y el porcentaje de error permitido. Luego, de acuerdo a la Definición 1.2.36, se construye la función de Lagrange $\mathcal{L} : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$, la cual se formula de la siguiente manera:

$$\begin{aligned} \mathcal{L}(w, b, \xi, \xi^*, \eta, \eta^*, \alpha, \alpha^*) =& \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) - \sum_{i=1}^{m}(\eta_i\xi_i + \eta_i^*\xi^*) \\ & - \sum_{i=1}^{m}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) \\ & - \sum_{i=1}^{m}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b). \end{aligned} \tag{3.1.7}$$

La función $\mathcal{L}$ es el lagrangiano y los valores $\eta_i$, $\eta_i^*$, $\alpha_i$, $\alpha_i^*$ son los multiplicadores de Lagrange. Por lo tanto, las variables duales en la ecuación anterior deben satisfacer las restricciones de positividad, es decir,

$$\alpha_i^*, \alpha, \eta_i^*, \eta \geq 0. \tag{3.1.8}$$

La siguiente proposición muestra el valor del parámetro buscado $w$ que proporciona la optimalidad de $\mathcal{L}$.

**Teorema 3.1.1.** El valor $w$ de la función $\mathcal{L}$ (3.1.7), el cual es parámetro para la función de regresión $\overline{f} : \mathbb{R}^n \to \mathbb{R}$ dada por

$$\overline{f}(x) = \langle w, x\rangle + b, \qquad \text{para cada } x \in \mathbb{R}^n,$$

es de la forma

$$w = \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)x_i, \tag{3.1.9}$$

con $\alpha_i,\ \alpha_i^*$ los parámetros óptimos del problema dual.

*Demostración.* Se calcula la derivada de $\mathcal{L}$ con respecto a $w$, esto con el fin de encontrar un punto que satisfaga la condición KKT de complementariedad mostrada en el Teorema 1.2.38.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w} &= \frac{\partial}{\partial w}\Bigg(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) - \sum_{i=1}^{m}(\eta_i\xi_i + \eta_i^*\xi^*) \\
&\quad - \sum_{i=1}^{m}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) - \sum_{i=1}^{m}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b)\Bigg) \\
&= \frac{\partial}{\partial w}\Bigg(\frac{1}{2}\|w\|^2 - \sum_{i=1}^{m}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) - \sum_{i=1}^{m}\alpha_i^*(\varepsilon + \xi_i^* + y_i \\
&\quad - \langle w, x_i\rangle - b)\Bigg) \\
&= w - \sum_{i=1}^{m}\alpha_i x_i + \sum_{i=1}^{m}\alpha_i^* x_i \\
&= w + \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)x_i \\
&= w - \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)x_i.
\end{aligned}$$

Se iguala la derivada parcial a cero y se obtiene el valor del parámetro $w$:

$$w = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)x_i. \tag{3.1.10}$$

□

Sólo resta hallar los valores de los $\alpha_i$ y $\alpha_i^*$. Para hacerlo, se calculan las derivadas parciales de $\mathcal{L}$ con respecto a cada una de las variables, se igualan a cero y se obtiene:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial b} &= \frac{\partial}{\partial b}\Bigg(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) - \sum_{i=1}^{m}(\eta_i\xi_i + \eta_i^*\xi^*) \\
&\quad - \sum_{i=1}^{m}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) - \sum_{i=1}^{m}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b)\Bigg) \\
&= \frac{\partial}{\partial b}\Bigg(-\sum_{i=1}^{m}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) - \sum_{i=1}^{m}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b)\Bigg) \\
&= -\sum_{i=1}^{m}\alpha_i b + \sum_{i=1}^{m}\alpha_i^* b \\
&= -b\sum_{i=1}^{m}\alpha_i + b\sum_{i=1}^{m}\alpha_i^* \\
&= b\sum_{i=1}^{m}(\alpha_i^* - \alpha_i).
\end{aligned}$$

Se iguala a cero y se obtiene lo siguiente:

$$\begin{aligned}
b\sum_{i=1}^{m}(\alpha_i^* - \alpha_i) = 0, \\
\sum_{i=1}^{m}(\alpha_i^* - \alpha_i) = 0.
\end{aligned} \tag{3.1.11}$$

Ahora, se calcula la derivada respecto a la variable $\xi$:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \xi} &= \frac{\partial}{\partial \xi}\Bigg(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) - \sum_{i=1}^{m}(\eta_i\xi_i + \eta_i^*\xi_i^*) \\
&\quad - \sum_{i=1}^{m}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) - \sum_{i=1}^{m}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b)\Bigg) \\
&= \frac{\partial}{\partial \xi}\Bigg(C\sum_{i=1}^{m}(\xi_i + \xi_i^*) - \sum_{i=1}^{m}(\eta_i\xi_i + \eta_i^*\xi^*) - \sum_{i=1}^{m}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b)\Bigg) \\
&= C - \eta_i - \alpha_i.
\end{aligned}$$

Se iguala a cero y se obtiene

$$C - \eta_i - \alpha_i = 0. \tag{3.1.12}$$

Finalmente, se calcula la derivada parcial respecto a la variable $\xi^*$:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \xi^*} &= \frac{\partial}{\partial \xi^*}\Bigg(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) - \sum_{i=1}^{m}(\eta_i\xi_i + \eta_i^*\xi_i^*) \\
&\quad - \sum_{i=1}^{m}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) - \sum_{i=1}^{m}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b)\Bigg) \\
&= \frac{\partial}{\partial \xi}\left(C\sum_{i=1}^{m}(\xi_i + \xi_i^*) - \sum_{i=1}^{m}(\eta_i\xi_i + \eta_i^*\xi^*) - \sum_{i=1}^{m}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b)\right) \\
&= C - \eta_i^* - \alpha_i^*.
\end{aligned}$$

Se iguala la derivada a cero:

$$C - \eta_i^* - \alpha_i^* = 0. \tag{3.1.13}$$

Se despejan los parámetros de Lagrange $\eta_i$ y $\eta_i^*$ de las ecuaciones (3.1.12) y (3.1.13) y se obtiene que

$$\eta_i = C - \alpha_i, \qquad \eta_i^* = C - \alpha_i^*. \tag{3.1.14}$$

Sustituyendo el valor de $w$ dado en la igualdad (3.1.9) y por las igualdades (3.1.11) y (3.1.14) en la función (3.1.7) se tiene el problema de optimización dual:

$$\begin{aligned}
\mathcal{W}(\alpha, \alpha^*, \eta, \eta^*) &= \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) - \sum_{i=1}^{m}(\eta_i\xi_i + \eta_i^*\xi_i^*) \\
&\quad - \sum_{i=1}^{m}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) - \sum_{i=1}^{m}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b) \\
&= \frac{1}{2}\left\langle \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)x_i, \sum_{j=1}^{m}(\alpha_j - \alpha_j^*)x_j \right\rangle + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) - \sum_{i=1}^{m}(\eta_i\xi_i + \eta_i^*\xi_i^*)
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{i=1}^{m} \alpha_i \left( \varepsilon + \xi_i - y_i + \left\langle \sum_{j=1}^{m} (\alpha_j - \alpha_j^*) x_j, x_i \right\rangle + b \right) \\
&- \sum_{i=1}^{m} \alpha_i^* \left( \varepsilon + \xi_i^* + y_i - \left\langle \sum_{j=1}^{m} (\alpha_j - \alpha_j^*) x_j, x_i \right\rangle - b \right) \\
&= \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) - \sum_{i=1}^{m} (\eta_i \xi_i + \eta_i^* \xi_i^*) \\
&- \sum_{i=1}^{m} \alpha_i \left( \varepsilon + \xi_i - y_i + \sum_{j=1}^{m} (\alpha_j - \alpha_j^*) \langle x_j, x_i \rangle + b \right) \\
&- \sum_{i=1}^{m} \alpha_i^* \left( \varepsilon + \xi_i^* + y_i - \sum_{j=1}^{m} (\alpha_j - \alpha_j^*) \langle x_j, x_i \rangle - b \right) \\
&= \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) - \sum_{i=1}^{m} (\eta_i \xi_i + \eta_i^* \xi_i^*) \\
&- \sum_{i=1}^{m} \alpha_i \varepsilon - \sum_{i=1}^{m} \alpha_i \xi_i + \sum_{i=1}^{m} \alpha_i y_i - \sum_{i=1}^{m} \alpha_i \sum_{j=1}^{m} (\alpha_j - \alpha_j^*) \langle x_j, x_i \rangle - b \sum_{i=1}^{m} \alpha_i \\
&- \sum_{i=1}^{m} \alpha_i^* \varepsilon - \sum_{i=1}^{m} \alpha_i^* \xi_i^* - \sum_{i=1}^{m} \alpha_i^* y_i + \sum_{i=1}^{m} \alpha_i^* \sum_{j=1}^{m} (\alpha_j - \alpha_j^*) \langle x_j, x_i \rangle + b \sum_{i=1}^{m} \alpha_i^* \\
&= \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) - \sum_{i=1}^{m} (\eta_i \xi_i + \eta_i^* \xi_i^*) \\
&- \varepsilon \sum_{i=1}^{m} \alpha_i + \alpha_i^* - \sum_{i=1}^{m} \alpha_i \xi_i + \sum_{i=1}^{m} \alpha_i y_i - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i (\alpha_j - \alpha_j^*) \langle x_j, x_i \rangle \\
&- \sum_{i=1}^{m} \alpha_i^* \xi_i^* - \sum_{i=1}^{m} \alpha_i^* y_i + \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i^* (\alpha_j - \alpha_j^*) \langle x_j, x_i \rangle + b \sum_{i=1}^{m} \alpha_i^* - \alpha_i \\
&= \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) \\
&- \varepsilon \sum_{i=1}^{m} \alpha_i + \alpha_i^* - \sum_{i=1}^{m} \alpha_i \xi_i + \sum_{i=1}^{m} \sum_{j=1}^{m} (\alpha_i^* - \alpha_i)(\alpha_j - \alpha_j^*) \langle x_j, x_i \rangle \\
&- \sum_{i=1}^{m} \alpha_i^* \xi_i^* + \sum_{i=1}^{m} y_i (\alpha_i - \alpha_i^*) - \sum_{i=1}^{m} ((C - \alpha_i) \xi_i + (C - \alpha_i^*) \xi_i^*)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\langle x_i,x_j\rangle + C\sum_{i=1}^{m}(\xi_i+\xi_i^*)\\
&\quad -\varepsilon\sum_{i=1}^{m}\alpha_i+\alpha_i^* - \sum_{i=1}^{m}\alpha_i\xi_i - \sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\langle x_j,x_i\rangle\\
&\quad -\sum_{i=1}^{m}\alpha_i^*\xi_i^* + \sum_{i=1}^{m}y_i(\alpha_i-\alpha_i^*) - \sum_{i=1}^{m}((C-\alpha_i)\xi_i+(C-\alpha_i^*)\xi_i^*)\\
&= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)\langle x_i,x_j\rangle + C\sum_{i=1}^{m}(\xi_i+\xi_i^*)\\
&\quad -\varepsilon\sum_{i=1}^{m}\alpha_i+\alpha_i^* - \sum_{i=1}^{m}\alpha_i\xi_i - \sum_{i=1}^{m}\alpha_i^*\xi_i^* + \sum_{i=1}^{m}y_i(\alpha_i-\alpha_i^*)\\
&\quad -C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i\xi_i - C\sum_{i=1}^{m}\xi_i^* + \sum_{i=1}^{m}\alpha_i^*\xi_i^*\\
&= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)\langle x_i,x_j\rangle + C\sum_{i=1}^{m}(\xi_i+\xi_i^*)\\
&\quad -\varepsilon\sum_{i=1}^{m}\alpha_i+\alpha_i^* - \sum_{i=1}^{m}\alpha_i\xi_i - \sum_{i=1}^{m}\alpha_i^*\xi_i^* + \sum_{i=1}^{m}y_i(\alpha_i-\alpha_i^*)\\
&\quad -C\sum_{i=1}^{m}(\xi_i+\xi_i^*) + \sum_{i=1}^{m}\alpha_i\xi_i + \sum_{i=1}^{m}\alpha_i^*\xi\\
&= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)\langle x_i,x_j\rangle - \varepsilon\sum_{i=1}^{m}(\alpha_i+\alpha_i^*) + \sum_{i=1}^{m}y_i(\alpha_i-\alpha_i^*).
\end{aligned}
$$

Por lo tanto, la función dual queda como

$$
\mathcal{W}(\alpha,\alpha^*) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)\langle x_i,x_j\rangle - \varepsilon\sum_{i=1}^{m}(\alpha_i+\alpha_i^*) + \sum_{i=1}^{m}y_i(\alpha_i-\alpha_i^*). \quad (3.1.15)
$$

La cual se desea maximizar, por ello se obtiene el problema de optimización dual:

$$\text{Maximizar } -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\langle x_i, x_j\rangle - \varepsilon\sum_{i=1}^{m}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{m} y_i(\alpha_i - \alpha_i^*) \quad (3.1.16)$$

$$\text{sujeto a} \quad \sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0 \quad (3.1.17)$$

$$\alpha_i, \alpha_i^* \in [0, C].$$

De manera similar a como se hizo en la sección 2.2, los valores de los parámetros $\alpha_i$ y $\alpha_i^*$ se estiman mediante un método del gradiente ascendente, similar a cuando se resolvió el problema Dual (véase (2.2.13)).

Después de haber encontrado el vector $\alpha$ óptimo se tiene que encontrar el valor del parámetro $b$ de la función de regresión, este valor se describe en la siguiente proposición.

**Proposición 3.1.2.** Sea $\{(x_i, y_i)\}_{i=1}^n \subset X\times\mathbb{R}$ el conjunto de entrenamiento y sean $\alpha, \alpha^* \in \mathbb{R}^n$ las soluciones del problema dual (3.1.15). Defínase los escalares $\beta_i$ y $f : X \to \mathbb{R}$ como

$$\beta_i := \alpha_i - \alpha_i^*, \qquad f(x) := \sum_{j=1}^{n}\beta_j \langle x_j, x\rangle.$$

Sean además

$$\mathcal{I} := \{\, k : \ 0 < \alpha_k < C \,\}, \qquad \mathcal{I}^* := \{\, l : \ 0 < \alpha_l^* < C \,\},$$

los cuales son subconjuntos de $\{1, ..., n\}$. Si $|\mathcal{I}| + |\mathcal{I}^*| > 0$, donde $|\mathcal{I}|$ y $|\mathcal{I}^*|$ denota la cardinalidad de los conjuntos $\mathcal{I}$ y $\mathcal{I}^*$ respectivamente, entonces $b$ puede tomarse como el promedio de las igualdades KKT correspondientes:

$$b \ = \ \frac{1}{|\mathcal{I}| + |\mathcal{I}^*|}\left(\sum_{i\in\mathcal{I}}\big(y_i - f(x_i) - \varepsilon\big) \ + \ \sum_{i\in\mathcal{I}^*}\big(y_i - f(x_i) + \varepsilon\big)\right). \quad (3.1.18)$$

En caso de que no existan índices “libres” (es decir, $|\mathcal{I}| + |\mathcal{I}^*| = 0$), se utiliza el conjunto de vectores de soporte

$$\mathcal{S} := \{\, i : \ \alpha_i > 0 \text{ o } \alpha_i^* > 0 \,\}$$

y se define

$$b \ = \ \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} (y_i - f(x_i)). \qquad \text{si } \ |\mathcal{S}| > 0. \tag{3.1.19}$$

donde $|\mathcal{S}|$ denota la cardinalidad del conjunto $\mathcal{S}$.

Si, por último, $\mathcal{S}$ también es vacío (caso degenerado), se toma

$$b = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i)).$$

*Demostración.* Sean $\beta_i := \alpha_i - \alpha_i^* \in \mathbb{R}$ y la función $f : X \to \mathbb{R}$ dada por $f(x) := \sum_{j=1}^{n} \beta_j \langle x_j, x \rangle$. Por el Teorema 3.1.1 se tiene

$$\langle w, x_i \rangle = \sum_{j=1}^{n} (\alpha_j^* - \alpha_j) \langle x_j, x_i \rangle = - \sum_{j} \beta_j \langle x_j, x_i \rangle = -f(x_i).$$

La condición de complementariedad (KKT) mostrada en el Teorema 1.2.38, aplicada al problema de optimización de las SVM para regresión (3.1.5-3.1.6), para cada $i = 1, ..., m$, es:

$$\alpha_i (y_i - \langle w, x_i \rangle - b - \varepsilon - \xi_i) = 0, \tag{3.1.20}$$

$$\alpha_i^* (\langle w, x_i \rangle + b - y_i - \varepsilon - \xi_i^*) = 0, \tag{3.1.21}$$

$$\eta_i \xi_i = 0, \qquad \eta_i^* \xi_i^* = 0. \tag{3.1.22}$$

Ahora, se procede a derivar las expresiones para $b$. Se analizan los casos por separado en los que $0 < \alpha_i < C$ o $0 < \alpha_i^* < C$.

Si $0 < \alpha_i < C$, entonces por (3.1.14) se tiene $\eta_i = C - \alpha_i > 0$, por lo que $\eta_i > 0$, la condición (3.1.22) fuerza a que $\xi_i = 0$. Además, por (3.1.20) con $\alpha_i > 0$ debe cumplirse

$$y_i - \langle w, x_i \rangle - b - \varepsilon - \xi_i = 0.$$

Sustituyendo $\xi_i = 0$ y $\langle w, x_i \rangle = f(x_i)$ se obtiene

$$y_i - f(x_i) - b - \varepsilon = 0 \quad \Longrightarrow \quad b = y_i - f(x_i) - \varepsilon.$$

Análogamente, si $0 < \alpha_i^* < C$, entonces $\xi_i^* = 0$ y por (3.1.21) se cumple

$$\langle w, x_i \rangle + b - y_i - \varepsilon = 0 \quad \Longrightarrow \quad b = y_i - f(x_i) + \varepsilon.$$

Por tanto, para cada índice $i$ en $\mathcal{I} = \{i : 0 < \alpha_i < C\}$ ó $\mathcal{I}^* = \{i : 0 < \alpha_i^* < C\}$ se tiene un valor

$$b_i = y_i - f(x_i) - \varepsilon, \qquad \text{o} \qquad b_i = y_i - f(x_i) + \varepsilon,$$

respectivamente. Como todas estas igualdades se deben satisfacer simultáneamente en el óptimo, se toma el promedio de estos valores. De ahí la fórmula (3.1.18):

$$b \;=\; \frac{1}{|\mathcal{I}| + |\mathcal{I}^*|}\Bigg( \sum_{i \in \mathcal{I}} \big(y_i - f(x_i) - \varepsilon\big) \;+\; \sum_{i \in \mathcal{I}^*} \big(y_i - f(x_i) + \varepsilon\big)\Bigg).$$

Ahora, supóngase que $|\mathcal{I}| + |\mathcal{I}^*| = 0$, es decir, todos los $\alpha_i$ y $\alpha_i^*$ están en los extremos $\alpha_i = 0$ o $\alpha_i = C$, esto se debe a la segunda condición del problema de optimización (3.1.16-3.1.17) . En ese caso no se cuenta con una expresión explícita para $b$. Para hallar el valor del parámetro $b$ se realiza lo siguiente:

Sea $\mathcal{S} := \{\, i : \;\; \alpha_i > 0 \text{ o } \alpha_i^* > 0 \,\} \neq \emptyset$ (es decir $|\mathcal{S}| > 0$, donde $|\mathcal{S}|$ denota la cardinalidad del conjunto $\mathcal{S}$), el conjunto de índices de los vectores de entrenamiento con $\alpha_i = C$ o $\alpha_i^* = C$. Como en los vectores de entrenamiento asociados a los índices $i \in \mathcal{S}$ la predicción $f(x_i) + b$ debe aproximar $y_i$ dentro de la $\varepsilon$-banda, es decir $|f(x_i) + b - y_i| \leq \varepsilon$, se toma $b$ que minimice la diferencia $f(x_i) + b) - y_i$ al cuadrado en $\mathcal{S}$, esto es,

$$b^\star = \underset{b \in \mathbb{R}}{\operatorname{argmin}} \sum_{i \in \mathcal{S}} \big((f(x_i) + b) - y_i\big)^2.$$

Se deriva $\sum_{i\in\mathcal{S}}\left((f(x_i)+b)-y_i\right)^2$ respecto de $b$ y se tiene

$$\frac{d}{db}\sum_{i\in\mathcal{S}}\left((f(x_i)+b)-y_i\right)^2 = 2\sum_{i\in\mathcal{S}}(f(x_i)+b^\star-y_i). \tag{3.1.23}$$

Se iguala (3.1.23) a cero:

$$\begin{aligned} 2\sum_{i\in\mathcal{S}}(f(x_i)+b^\star-y_i) &= 0,\\ \sum_{i\in\mathcal{S}}(f(x_i)-y_i)+|\mathcal{S}|b^\star &= 0,\\ b^\star &= \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}(y_i-f(x_i)), \end{aligned} \tag{3.1.24}$$

que coincide con la fórmula (3.1.19).

Supóngase que $\mathcal{S}$ fuese vacío, es decir, $\alpha_i = 0$ y $\alpha_i^* = 0$ para cada $i = 1, ..., n$. Entonces el parámetro $b$ es el promedio sobre todo el conjunto de entrenamiento:

$$b = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i)).$$

□

Finalmente, con la igualdad (3.1.9) sustituida en (3.1.1) se tiene la *función de regresión del problema Linear-SVMR* $\overline{f}:\mathbb{R}^n\to\mathbb{R}$, dada por

$$\overline{f}(z) = \sum_{i=1}^{m}(\alpha_i-\alpha_i^*)\langle x_i, z\rangle + b^\star, \tag{3.1.25}$$

esta es la llamada *expansión de vectores de soporte*, esta función determina el valor para una nueva entrada $z\in\mathbb{R}^n$.

En la mayoría de los casos, los datos de entrenamiento no pueden ser modelados con una SVM para regresión lineal; por ello, se implementa el uso de funciones kernel, ya que a través de estas funciones se pueden establecer relaciones no lineales entre los datos, lo que provoca una mejor predicción de los resultados. Para la formulación de la SVMR, se reemplaza el producto interior por una función kernel $K(x, y)$; esta sustitución se realiza desde el planteamiento del

modelo, de esta forma, el problema a solucionar queda definido como

$$\text{Minimizar} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*), \tag{3.1.26}$$

$$\text{sujeto a} \quad y_i - K(w, x_i) - b \leq \varepsilon + \xi_i; \tag{3.1.27}$$

$$K(w, x_i) + b - y_i \leq \varepsilon + \xi_i^*;$$

$$\xi_i,\ \xi_i^* \geq 0.$$

El problema dual del problema de optimización es de la siguiente forma:

$$\text{Maximizar} \ -\frac{1}{2}\sum_{i,j=1}^{m}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) - \varepsilon\sum_{i=1}^{m}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{m} y_i(\alpha_i - \alpha_i^*) \tag{3.1.28}$$

$$\text{sujeto a} \quad \sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0, \tag{3.1.29}$$

$$\alpha_i, \alpha_i^* \in [0, C]. \tag{3.1.30}$$

Por lo que la solución está dada por

$$f(x) = \sum_{i=1}^{m}(\alpha_i^* - \alpha)K(x_i, x) + b. \tag{3.1.31}$$

Los vectores de soporte son aquellos puntos $x_i$ en los cuales el error de interpolación es mayor o igual a $\xi$, a diferencia de los puntos en los que el error de interpolación es menor que $\xi$; estos nunca son vectores de soporte y no forman parte de la solución.

Una vez que los puntos que no son vectores soporte han sido encontrados, pueden ser eliminados del conjunto de datos, y si se resuelve el problema de programación nuevamente sobre el conjunto reducido, se encuentra la misma solución.

A continuación se presenta un ejemplo de regresión SVM utilizando el kernel gaussiano. Considere el siguiente conjunto formado por parejas de datos unidimensionales y su respectivo

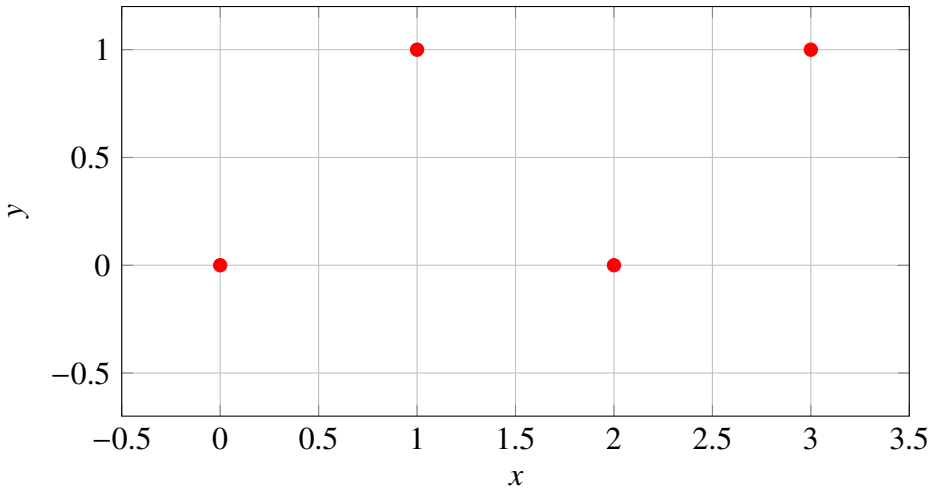


**Figura 3.1:** Datos de entrenamiento.

valor de la variable dependiente:

| $i$ | $x_i$ | $y_i$ |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 1 | 1 |
| 3 | 2 | 0 |
| 4 | 3 | 1 |

Estos datos no pueden ajustarse adecuadamente mediante un modelo lineal como se ve en la Figura 3.1, por lo que se utilizará el modelo SVMR con kernel gaussiano. El kernel gaussiano se define como:

$$K(x, z) = \exp\left(-\gamma \|x - z\|^2\right), \quad \gamma > 0.$$

Este kernel induce un espacio de Hilbert de dimensión infinita y permite modelar relaciones no lineales. Para este ejemplo se fijan los valores:

$$\varepsilon = 0.1, \quad C = 10, \quad \gamma = 1.$$

Al resolver el problema dual por medio del método del gradiente ascendente, todos los

puntos resultan ser vectores soporte y se obtienen los siguientes valores:

| $x_i$ | $\alpha_i - \alpha_i^*$ |
|---|---|
| 0 | 0.8 |
| 1 | −0.9 |
| 2 | 0.9 |
| 3 | −0.8 |

La función de regresión para este caso es de la forma:

$$f(x) = -0.8\, e^{-(x-0)^2} + 0.9\, e^{-(x-1)^2} - 0.9\, e^{-(x-2)^2} + 0.8\, e^{-(x-3)^2} + b.$$

Para hallar el valor del sesgo $b$ se ocupa la Proposición 3.1.2. Para ello, note que los conjuntos $I$ y $I^*$ de la Proposición 3.1.2 cumplen $|I| + |I^*| = 4 > 0$. Se realizan los cálculos y se obtiene $b = 0.5$. La función resultante se puede observar en la Figura 3.2.

La función $f(x)$ es una combinación lineal de funciones gaussianas centradas en los vectores soporte. Cada término contribuye localmente al ajuste, lo que permite capturar la estructura no lineal de los datos.

Este ejemplo ilustra cómo el kernel gaussiano convierte el SVR en un aproximador no lineal flexible sin necesidad de construir explícitamente el mapeo $\phi(x)$.

## 3.2. Máquina de soporte vectorial de mínimos cuadrados

En la sección (3.1) se mencionó que las máquinas de soporte vectorial se resuelven como un problema de programación cuadrática, obteniendo la solución (3.1.31). No obstante, existe la posibilidad de simplificar aspectos de la formulación de las **SVMR**, sin perder sus ventajas; para eso se utilizan las máquinas de soporte vectorial de mínimos cuadrados (LS-SVM).

El modelo LS-SVM es una reformulación de las máquinas de soporte vectorial de Vapnik, en las que la optimización lleva a resolver un sistema de ecuaciones lineales, lo cual es más simple de utilizar que las soluciones de programación cuadrática anteriormente descritas en el

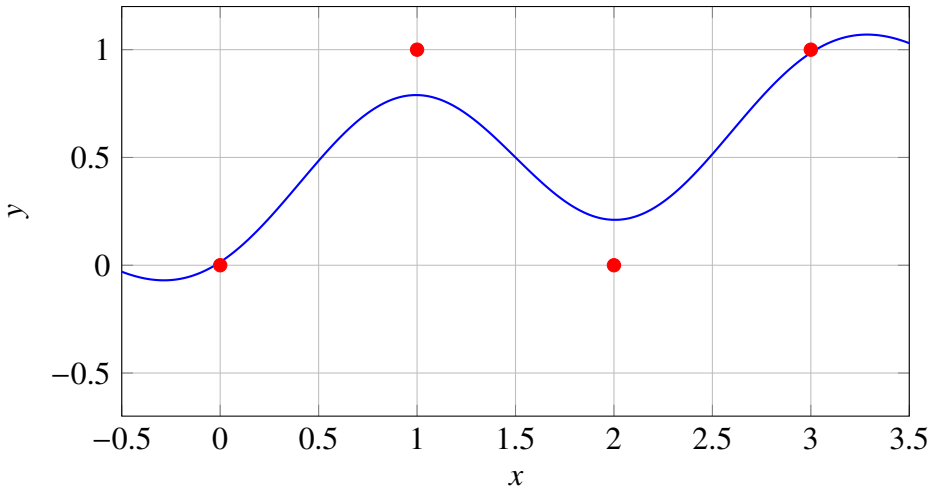


**Figura 3.2:** Función de regresión.

problema de optimización conformado por (3.1.26)-(3.1.27).

La formulación del problema de optimización (3.1.26)-(3.1.27) descrita en la subsección 3.1 se modifica en dos puntos. El primer punto se basa en el uso de ecuaciones de igualdad en lugar de las inecuaciones (3.1.6). En dichas ecuaciones, el valor de la derecha se toma como valor objetivo, más que como un umbral. Sobre este valor se permite un valor de estimación variable, el cual juega un papel similar a las variables de holgura en **SVMR**, es por ello que esta variable se denomina $\xi$. Como siguiente punto, la función de pérdida de Vapnik se sustituye por una función de pérdida cuadrática $\xi^2$. Estas modificaciones simplifican de manera significativa el problema. En específico, se considera la siguiente ecuación:

$$\overline{f}(x) = \langle w, k(x) \rangle + b, \tag{3.2.1}$$

donde $x \in X,\ y \in \mathbb{R},\ k : \mathbb{R}^n \to \mathbb{R}^p$ con $p \gg n$ es un mapeo hacia un espacio de características de dimensión mayor. Dado un conjunto de entrenamiento $\{(x_1, y_1), ..., (x_m, y_m)\} \subset X \times \mathbb{R}$, con

$X \subseteq \mathbb{R}^n$. El problema de optimización se expresa como

$$\text{Minimizar} \qquad \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m}(\xi_i^2), \tag{3.2.2}$$

$$\text{sujeto a} \qquad y_i = \langle w, k(x_i)\rangle + b + \xi_i. \tag{3.2.3}$$

Se formula su función de Lagrange por medio de la Definición 1.2.36 como la función $\mathcal{L} : \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ dada por

$$\mathcal{L}(w, b, \alpha, \xi) = \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2 - \sum_{i=1}^{m}\alpha_i(\langle w, k(x_i)\rangle + b + \xi_i - y_i), \tag{3.2.4}$$

y se encuentran las derivadas parciales de la función $\mathcal{L}$:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial b} &= \frac{\partial}{\partial b}\left(\frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2 - \sum_{i=1}^{m}\alpha_i(\langle w, k(x_i)\rangle + b + \xi_i - y_i)\right) \\
&= \frac{\partial}{\partial b}\left(-\sum_{i=1}^{m}\alpha_i(\langle w, k(x_i)\rangle + b + \xi_i - y_i)\right) \\
&= \frac{\partial}{\partial b}\left(-\sum_{i=1}^{m}\alpha_i\langle w, k(x_i)\rangle - b\sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i\xi_i + \sum_{i=1}^{m}\alpha_i y_i\right) \\
&= -\frac{\partial}{\partial b}\sum_{i=1}^{m}\alpha_i\langle w, k(x_i)\rangle - \frac{\partial}{\partial b}b\sum_{i=1}^{m}\alpha_i - \frac{\partial}{\partial b}\sum_{i=1}^{m}\alpha_i\xi_i + \frac{\partial}{\partial b}\sum_{i=1}^{m}\alpha_i y_i. \\
&= \sum_{i=1}^{m}\alpha_i.
\end{aligned}$$

Al igualar la derivada anterior a cero se tiene

$$\sum_{i=1}^{m}\alpha_i = 0, \tag{3.2.5}$$

Para hallar el valor del parámetro $w$ de la función de regresión (3.2.1) se tiene la siguiente

proposición.

**Proposición 3.2.1.** El valor de $w$ de la función $\mathcal{L}$ en (3.2.4) y además parámetro $w$ de la función de regresión (3.2.1) está dado por

$$w = \sum_{i=1}^{m} \alpha_i k(x_i).$$

*Demostración.* Se calcula la derivada parcial de $\mathcal{L}$ con respecto a $w$:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w} &= \frac{\partial}{\partial w}\left(\frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2 - \sum_{i=1}^{m}\alpha_i(\langle w, k(x_i)\rangle + b + \xi_i - y_i)\right) \\
&= \frac{\partial}{\partial w}\left(\frac{1}{2}\|w\|^2 - \sum_{i=1}^{m}\alpha_i(\langle w, k(x_i)\rangle + b + \xi_i - y_i)\right) \\
&= \frac{\partial}{\partial w}\left(\frac{1}{2}\|w\|^2\right) - \frac{\partial}{\partial w}\left(\sum_{i=1}^{m}\alpha_i\langle w, k(x_i)\rangle\right) \\
&= w - \sum_{i=1}^{m}\alpha_i k(x_i).
\end{aligned}$$

Se iguala a cero y se despeja a $w$, con lo que se tiene la siguiente igualdad:

$$w = \sum_{i=1}^{m} \alpha_i k(x_i). \tag{3.2.6}$$

□

La siguiente proposición proporciona una metodología para hallar los valores de los parámetros de la función de regresión sin recurrir a métodos de gradiente ascendente, sólo basta resolver un problema de álgebra lineal.

**Proposición 3.2.2.** Para el problema de optimización (3.2.1) con función de Lagrange $\mathcal{L}$ dada en (3.2.4), los valores de $\alpha_i$ y $b$ están dados por la solución del problema

$$\begin{bmatrix} 0 & \bar{1}^T \\ \bar{1} & K + \frac{1}{C}I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \tag{3.2.7}$$

donde $\bar{1}$ denota el vector de unos de tamaño $m$, $I$ denota la matriz identidad de tamaño $m \times m$, $\alpha$ denota el vector $\alpha = [\alpha_1\ \alpha_2\ ...\ \alpha_m]^T$ y

$$K = \begin{pmatrix} \langle k(x_1), k(x_1)\rangle & \langle k(x_2), k(x_1)\rangle & \cdots & \langle k(x_m), k(x_1)\rangle \\ \langle k(x_1), k(x_2)\rangle & \langle k(x_2), k(x_2)\rangle & \cdots & \langle k(x_m), k(x_2)\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle k(x_1), k(x_m)\rangle & \langle k(x_2), k(x_m)\rangle & \cdots & \langle k(x_m), k(x_m)\rangle \end{pmatrix}.$$

Por lo tanto, la función de regresión $\overline{f} : \mathbb{R}^n \to \mathbb{R}$ es de la forma

$$\bar{f}(x) = \sum_{i=1}^{N} \alpha_i \langle k(x_i), k(x)\rangle + b.$$

*Demostración.* Se calcula la derivada parcial de $\mathcal{L}$ con respecto a $\xi$ y se tiene lo siguiente.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \xi} &= \frac{\partial}{\partial \xi}\left(\frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2 - \sum_{i=1}^{m}\alpha_i(\langle w, k(x_i)\rangle + b + \xi_i - y_i)\right) \\ &= \frac{\partial}{\partial \xi}\left(\frac{C}{2}\sum_{i=1}^{m}\xi_i^2 - \sum_{i=1}^{m}\alpha_i(\langle w, k(x_i)\rangle + b + \xi_i - y_i)\right) \\ &= \frac{\partial}{\partial \xi}\left(\frac{C}{2}\sum_{i=1}^{m}\xi_i^2\right) - \frac{\partial}{\partial \xi}\left(\sum_{i=1}^{m}\alpha_i(\langle w, k(x_i)\rangle + b + \xi_i - y_i)\right) \\ &= \frac{\partial}{\partial \xi}\left(\frac{C}{2}\sum_{i=1}^{m}\xi_i^2\right) - \frac{\partial}{\partial \xi}\left(\sum_{i=1}^{m}\alpha_i\xi_i\right) \\ &= C\xi - \alpha. \end{aligned}$$

Se iguala a cero y se despeja a $\xi$, se tiene que

$$\xi = \frac{\alpha}{C}. \tag{3.2.8}$$

Finalmente, se calcula la derivada parcial con respecto a $\alpha$.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \alpha} &= \frac{\partial}{\partial \alpha}\left(\frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2 - \sum_{i=1}^{m}\alpha_i(\langle w, k(x_i)\rangle + b + \xi_i - y_i)\right) \\ &= -\frac{\partial}{\partial \alpha}\left(\sum_{i=1}^{m}\alpha_i(\langle w, k(x_i)\rangle + b + \xi_i - y_i)\right)\end{aligned}$$

$$= -(\langle w, k(x_i)\rangle + b + \xi_i - y_i).$$

Se iguala a cero y se tiene lo siguiente:

$$\langle w, k(x_i)\rangle + b + \xi_i - y_i = 0, \qquad \text{para cada } i = 1, 2, ..., m. \tag{3.2.9}$$

Sustituyendo el valor de $w$ dado en (3.2.6) y la igualdad (3.2.8) dentro de (3.2.9), se obtiene la ecuación:

$$y_i = \sum_{j=1}^{m}\alpha_j\langle k(x_j), k(x_i)\rangle + \frac{\alpha_i}{C} + b, \qquad \text{para cada } i = 1, 2, ..., m. \tag{3.2.10}$$

Por las ecuaciones (3.2.5) y (3.2.10), se obtiene el siguiente sistema de ecuaciones:

$$\begin{cases} \sum_{i=1}^{m}\alpha_i = 0; \\ \\ \sum_{j=1}^{m}(\alpha_j\langle k(x_j), k(x_i)\rangle) + \frac{\alpha_i}{C} + b = y_i, \qquad \text{para cada } i = 1, 2, ..., m. \end{cases} \tag{3.2.11}$$

Reescribiendo las ecuaciones en forma de sistema matricial, se tiene:

$$\begin{bmatrix} 0 & \bar{1}^T \\ \\ \bar{1} & K + \frac{1}{C}I \end{bmatrix}\begin{bmatrix} b \\ \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ \\ y \end{bmatrix}, \tag{3.2.12}$$

donde $\bar{1}$ denota el vector de unos de tamaño $m$, $I$ denota la matriz identidad de tamaño $m \times m$, $\alpha$ denota el vector $\alpha = [\alpha_1\ \alpha_2\ ...\ \alpha_m]^T$, $y = (y_1, \ldots, y_m)^t \in \mathbb{R}^m$ denota los valores asociados a cada

valor de entrenamiento y

$$K = \begin{pmatrix} \langle k(x_1), k(x_1)\rangle & \langle k(x_2), k(x_1)\rangle & \cdots & \langle k(x_m), k(x_1)\rangle \\ \langle k(x_1), k(x_2)\rangle & \langle k(x_2), k(x_2)\rangle & \cdots & \langle k(x_m), k(x_2)\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle k(x_1), k(x_m)\rangle & \langle k(x_2), k(x_m)\rangle & \cdots & \langle k(x_m), k(x_m)\rangle \end{pmatrix}.$$

Al resolver el sistema de ecuaciones (3.2.7) se obtienen los valores de $b$ y $\alpha$, por lo que la función de regresión se expresa de la siguiente forma:

$$\bar{f}(x) = \sum_{i=1}^{m} \alpha_i \langle k(x_i), k(x)\rangle + b. \tag{3.2.13}$$

□

Si se denota al vector cuyas entradas son los productos internos $\langle k(x_i), k(x)\rangle$ por $k(x_i, x)$, entonces la función de regresión (3.2.13) se escribe como

$$\bar{f}(x) = \langle k(x_i, x), \alpha\rangle + b. \tag{3.2.14}$$

Solo resta hallar las condiciones bajo las cuales el problema de álgebra lineal tiene solución, y con ello se garantizará que el método funciona para estimar los parámetros de la función de regresión buscada, la siguiente definición es la norma utilizada por el teorema que proporciona una de estas condiciones.

**Definición 3.2.3.** Sea una matriz $A \in \mathbb{R}^{n\times m}$. La norma $||A||_2$ de la matriz $A$ está dada por:

$$||A||_2 = \sqrt{\sum_{i,j} A_{ij}^2}\ , \tag{3.2.15}$$

donde $A_{ij}$ son las entradas de la matriz.

**Teorema 3.2.4.** Si las columnas de la matriz

$$Q = \begin{bmatrix} 0 & \bar{1}^T \\ \bar{1} & K + \frac{1}{C} I \end{bmatrix}$$

son linealmente independientes, entonces el problema (3.2.7) tiene solución y es de la forma:

$$\begin{bmatrix} b^* \\ \alpha^* \end{bmatrix} = (Q^t Q)^{-1} Q^t \begin{bmatrix} 0 \\ y \end{bmatrix}. \tag{3.2.16}$$

Donde $\alpha^* = (\alpha_1^*, ..., \alpha_m^*)^T \in \mathbb{R}^n$ y $b \in \mathbb{R}$.

*Demostración.* Se puede reescribir el sistema (3.2.7) de la siguiente manera:

$$\begin{bmatrix} 0 \\ y \end{bmatrix} - \begin{bmatrix} 0 & \bar{1}^T \\ \bar{1} & K + \frac{1}{C} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = 0. \tag{3.2.17}$$

Se calcula la norma $\| \cdot \|_2$ de la matriz

$$\left\| \begin{bmatrix} 0 \\ y \end{bmatrix} - \begin{bmatrix} 0 & \bar{1}^T \\ \bar{1} & K + \frac{1}{C} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} \right\|_2 = 0. \tag{3.2.18}$$

Por el Teorema 1.2.10 existe un vector

$$\begin{bmatrix} b^* \\ \alpha^* \end{bmatrix} \tag{3.2.19}$$

en $\mathbb{R}^{1+m}$ el cual minimiza la función (3.2.18) y es de la forma:

$$\begin{bmatrix} b^* \\ \alpha^* \end{bmatrix} = (Q^t Q)^{-1} Q^t \begin{bmatrix} 0 \\ y \end{bmatrix}. \tag{3.2.20}$$

□

De esta manera, se establece una metodología para estimar los parámetros de la función de regresión mediante la resolución de un problema de álgebra lineal. A continuación se muestra un ejemplo del funcionamiento de este modelo.

Consideremos el siguiente conjunto de entrenamiento en $\mathbb{R}$:

| $x_i$ | $y_i$ |
|---|---|
| 1 | 1 |
| 2 | 2 |
| 3 | 2 |

El objetivo es aproximar la relación entre $x$ e $y$ mediante una función de regresión. En LS-SVM para regresión se busca una función de la forma

$$\overline{f}(x) = wx + b,$$

donde $w \in \mathbb{R}$ y $b \in \mathbb{R}$. En este ejemplo se toma $C = 1$.

En forma matricial, el problema se muestra de la siguiente forma:

$$\begin{bmatrix} 0 & \bar{1}^T \\ \bar{1} & K + I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \tag{3.2.21}$$

donde $K_{ij} = x_i x_j$. Para los datos dados:

$$K = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{pmatrix}, \qquad K + I = \begin{pmatrix} 2 & 2 & 3 \\ 2 & 5 & 6 \\ 3 & 6 & 10 \end{pmatrix}.$$

Por lo tanto,

$$\begin{bmatrix} 0 & \bar{1}^T \\ \bar{1} & K+I \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 2 & 2 & 3 \\ 1 & 2 & 5 & 6 \\ 1 & 3 & 6 & 10 \end{bmatrix} \tag{3.2.22}$$

Nótese que la matriz tiene columnas linealmente independientes, en efecto, se comprueba que para la combinación lineal

$$x_1 \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} + x_2 \begin{pmatrix} 1 \\ 2 \\ 2 \\ 3 \end{pmatrix} + x_3 \begin{pmatrix} 1 \\ 2 \\ 5 \\ 6 \end{pmatrix} + x_4 \begin{pmatrix} 1 \\ 3 \\ 6 \\ 10 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{3.2.23}$$

las variables $x_i = 0$ con $i = 1, 2, 3, 4$. Esto se comprueba resolviendo el sistema

$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 2 & 2 & 3 \\ 1 & 2 & 5 & 6 \\ 1 & 3 & 6 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{3.2.24}$$

para el cual efectivamente las soluciones son $x_1 = x_2 = x_3 = x_4 = 0$. Ahora, se aplica el Teorema 3.2.4, por lo que se halla la inversa y la transpuesta de la matriz:

$$Q = \begin{bmatrix} 0 & \bar{1}^T \\ \bar{1} & K + \frac{1}{C} I \end{bmatrix} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 2 & 2 & 3 \\ 1 & 2 & 5 & 6 \\ 1 & 3 & 6 & 10 \end{pmatrix},$$

Como $Q$ es simétrica, se tiene que $Q = Q^t$, por lo que la matriz de coeficientes es

$$(Q^tQ)^{-1}Q^t = \left(\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 2 & 2 & 3 \\ 1 & 2 & 5 & 6 \\ 1 & 3 & 6 & 10 \end{pmatrix} * \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 2 & 2 & 3 \\ 1 & 2 & 5 & 6 \\ 1 & 3 & 6 & 10 \end{pmatrix}\right)^{-1} * \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 2 & 2 & 3 \\ 1 & 2 & 5 & 6 \\ 1 & 3 & 6 & 10 \end{pmatrix}$$

$$= \left(\begin{pmatrix} 3 & 7 & 13 & 19 \\ 7 & 18 & 33 & 49 \\ 13 & 33 & 66 & 97 \\ 19 & 49 & 97 & 146 \end{pmatrix}\right)^{-1} * \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 2 & 2 & 3 \\ 1 & 2 & 5 & 6 \\ 1 & 3 & 6 & 10 \end{pmatrix}$$

$$= \begin{pmatrix} 4 & -\frac{13}{9} & -\frac{5}{9} & \frac{1}{3} \\ -\frac{13}{9} & \frac{11}{9} & 0 & -\frac{2}{9} \\ -\frac{5}{9} & 0 & \frac{7}{9} & -\frac{4}{9} \\ \frac{1}{3} & -\frac{2}{9} & -\frac{4}{9} & \frac{1}{3} \end{pmatrix} * \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 2 & 2 & 3 \\ 1 & 2 & 5 & 6 \\ 1 & 3 & 6 & 10 \end{pmatrix}$$

$$= \begin{pmatrix} -\frac{5}{3} & 1 & \frac{1}{3} & -\frac{1}{3} \\ 1 & \frac{1}{3} & -\frac{1}{3} & 0 \\ \frac{1}{3} & -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & 0 & -\frac{1}{3} & \frac{1}{3} \end{pmatrix}.$$

Se utiliza la Proposición 3.2.2 por lo que se obtiene el sistema:

$$\begin{pmatrix} -\frac{5}{3} & 1 & \frac{1}{3} & -\frac{1}{3} \\ 1 & \frac{1}{3} & -\frac{1}{3} & 0 \\ \frac{1}{3} & -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & 0 & -\frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 2 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ -\frac{1}{3} \\ \frac{1}{3} \\ 0 \end{pmatrix}. \tag{3.2.25}$$

Y resulta

$$\alpha^* = (-\frac{1}{3}, \frac{1}{3}, 0) \qquad \text{y} \qquad b^* = 1.$$

Con estos valores la función de regresión LS-SVM es:

$$\begin{aligned} f(x) &= \sum_{i=1}^{3} \alpha_i x_i x + b \\ &= (-\frac{1}{3} * 1 + \frac{1}{3} * 2 + 0 * 3)x + 1 \\ &= \frac{1}{3}x + 1. \end{aligned}$$

La recta de ajuste y los datos de entrenamiento se visualizan en la Figura 3.3. En este caso se usó el kernel lineal, sin embargo, se puede utilizar otro tipo de kernel para obtener una función no lineal que ajuste de mejor manera a los datos.

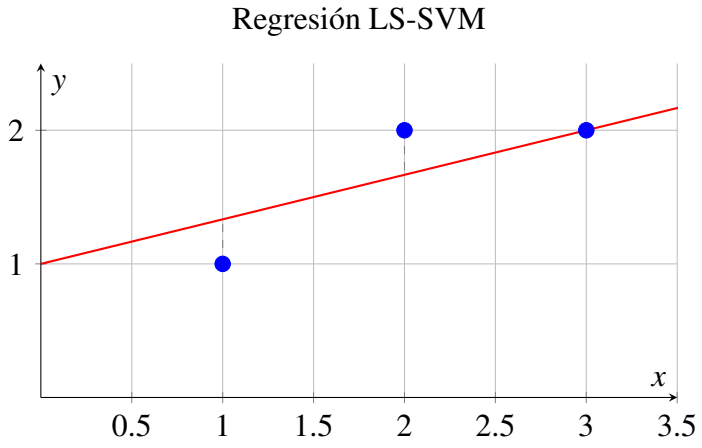


**Figura 3.3:** Datos de entrenamiento y la recta de ajuste.

# Conclusiones

El aprendizaje automático es un área que se ha desarrollado desde finales del siglo pasado; sin embargo, ha tomado relevancia desde la década pasada. Algunos algoritmos de aprendizaje se diseñaron para mejorar los procesos que requieren de manipulación de datos, principalmente en las tareas de clasificación y regresión. Uno de los principales procedimientos para realizar estas tareas es el modelo de máquina de soporte vectorial (SVM), el cual fue presentado por Vladimir Vapnik en colaboración con Corina Cortes en su artículo denominado *Support-Vector Networks* en el año de 1995 [13]. A lo largo del tiempo se ha ido analizando de manera más exhaustiva la formulación matemática de las SVM's, sin embargo, en la revisión bibliográfica [1, 2, 6, 8, 10, 11, 12, 15, 17, 22, 24, 25] se detectó que hay una deficiencia en los documentos que muestren de manera completa todo el desarrollo y la teoría que se ocupa para el correcto funcionamiento del modelo. Por esta razón, el objetivo del presente trabajo de tesis fue realizar una memoria autocontenida, en la medidad de lo posible, que contenga una formalización matemática rigurosa para el modelo de máquina de soporte vectorial (SVM), que permita comprender sus propiedades limitaciones y comportamiento en aplicaciones de clasificación y regresión complejas, para lograrlo, en un primer paso se recopiló la teoría matemática que sustenta el modelo de SVM, esta se encuentra en la subsección 2.2.

Así también, se definió un marco matemático de las propiedades fundamentales de la SVM: en la sección 2.3 se incluyen las formulaciones para el caso de margen duro para datos linealmente separables, margen suave para datos no separables y el uso de una función kernel. Un resultado destacable es el Teorema 2.4.23, el cual es poco conocido dentro de la literatura y que establece que es posible construir un espacio de dimensión alta en el que un conjunto no linealmente separable se puede separar. Luego, el caso de regresión se presenta en la sección 3, en esta se incluye la formulación clásica y se resuelve mediante operadores de Lagrange.

Por otra parte, se agrega el problema de regresión resuelto mediante la formulación de mínimos cuadrados.

Los resultados principales de esta investigación son la Proposición 2.2.6 y la Proposición 2.2.8 las cuales describen la forma de los valores de los parámetros cuando los datos de entrada son linealmente separables. La Proposición 2.3.3 y Proposición 2.3.4, muestran la forma de los parámetros del hiperplano cuando los datos no son linealmente separables, es decir, el caso de margen suave. La Proposición 2.4.24 muestra el valor del parámetro $b$ del hiperplano de separación utilizando una función kernel. El Teorema 3.1.1 indica el valor del parámetro $w$ y la Proposición 3.1.2 describe el valor de $b$ en la función de regresión. Finalmente, la Proposición 3.2.2 formula el problema de regresión utilizando mínimos cuadrados y el Teorema 3.2.4 muestra la forma de la solución. En conjunto, estos resultados muestran el valor de los parámetros necesarios para la formulación de las SVM en sus diferentes presentaciones, por lo que se puede observar a detalle cómo es el funcionamiento de una SVM.

Cabe mencionar que la complejidad de una SVM depende de dos aspectos, el primero es la cantidad de datos de muestra que se tenga en el problema, en el caso de tener un volumen de datos demasiado grande, los cálculos resultan demasiado costosos en términos de tiempo y recursos computacionales. El segundo punto es el uso de las funciones kernel que trasladan los datos a un espacio de dimensión más alta donde estos sean linealmente separables.

Con este trabajo de tesis se despejan algunas interrogantes como ¿Cuales son las bases que dan el sustento teórico para la formulación de cada una de las versiones del modelo de SVM?, ¿Cuál es la formulación detallada del modelo de SVM?, ¿Cómo se obtiene el valor de los parámetros para la función de clasificación y regresión? La primera pregunta se despeja en su mayoría con la sección de preliminares de este trabajo de tesis, la teoría restante se encuentra en las secciones de formulación del modelo. La siguiente interrogante se responde con cada una de las secciones dentro del capítulo de formulación del modelo SVM, en cada una se indican detalladamente los pasos a seguir para poder formular el modelo en cada una de sus versiones. La tercera interrogante se responde con los diferentes teoremas y proposiciones desarrollados en el trabajo de tesis donde se indican explícitamente los valores de los parámetros para las funciones de clasificación y regresión respectivas. En la práctica las condiciones que requieren

cada uno de los teoremas y proposiciones por lo general son omitidas en la bibliografía revisada. Este trabajo de tesis, abre paso a futuras investigaciones como la formalización matemática de las máquinas de soporte vectorial en su versión cuántica [21].

# Bibliografía


[1] Aggarwal, C. *Data Classification: Algorithms and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Taylor & Francis, 2014.

[2] Aggarwal, C. *Data Mining: The Textbook*. Springer International Publishing, 2015.

[3] Akande, K. O., Owolabi, T. O., Olatunji, S. O., and AbdulRaheem, A. A. A hybrid particle swarm optimization and support vector regression model for modelling permeability prediction of hydrocarbon reservoir. *Journal of Petroleum Science and Engineering 150* (2017), 43–53. Published 2017 — PSO-SVR for permeability prediction.

[4] Aronszajn, N. Theory of reproducing kernels. *Transactions of the American Mathematical Society 68*, 3 (1950).

[5] Ayestaran, R., and Las-Heras, F. Support vector regression for the design of array antennas. *IEEE Antennas and Wireless Propagation Letters 4* (2005), 414–416.

[6] Bi, J., Bennett, K., Embrechts, M., Breneman, C., and Song, M. Dimensionality reduction via sparse support vector machines. *J. Mach. Learn. Res. 3* (March 2003), 1229–1243.

[7] Bogachev, V. I. *Measure Theory*, vol. I. Springer, Berlin, 2007.

[8] Boswell, D. *Introduction to Support Vector Machines*. 2002.

[9] Boyd, S., and Vandenberghe, L. *Convex Optimization*. Berichte über verteilte messysteme. Cambridge University Press, 2004.

[10] Burges, C. J. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery 2* (1998), 121–67.

[11] Calafiore, G., and Ghaoui, L. *Optimization Models*. Cambridge University Press, 2014.

[12] Campbell, C., and Ying, Y. *Learning with Support Vector Machines*. Morgan & Claypool Publishers, 2011.

[13] Cortes, C., and Vapnik, V. Support-vector networks. *Machine Learning 20* (1995), 273–297.

[14] Cover, T. M. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *TRANSACTIONS ON ELECTRONIC COMPUTERS* (1965).

[15] Cristianini, N., and Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[16] Dai, X., Mi, D., Wu, H., and Zhang, Y. Design of compact patch antenna based on support vector regression. *Radioengineering 31* (09 2022), 339–345.

[17] Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2009.

[18] Holmes, D. Convex hulls solve svms. MIT, 2013.

[19] Luenberger, D. *Optimization by Vector Space Methods*. Professional Series. Wiley, 1997.

[20] Ovchinnikov, S. *Functional Analysis: An Introductory Course*. Universitext. Springer International Publishing, 2018.

[21] Quesada Perez, C. *Máquinas de soporte vectorial cuánticas*. Universidad D'Alacant, 2023.

[22] Rebentrost, P., Mohseni, M., and Lloyd, S. Quantum support vector machine for big data classification. *Physical Review Letters 113*, 13 (2014).

[23] Rockafellar, R. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1997.

[24] Scholkopf, B., and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[25] Tianzhu, Y. Quantum support vector machines: theory and applications. *Quantum Machine Learning: Bridging Quantum Physics and Computational Simulations* (2024).

[26] Zhang, B., Ren, H., Huang, G., Cheng, Y., and Hu, C. Predicting blood pressure from physiological index data using the svr algorithm. *BMC Bioinformatics* (2019 February).