Universidad de Ingeniería y Tecnología - UTEC
Departamento de Ingeniería Mecatrónica
Curso: Tópicos de Ingeniería Mecatrónica (2020 -2)

Tema 5:
Support Vector Machines (SVM) - Parte 1

Prof. Oscar E. Ramos Ponce

El clasificador denominado *máquina de vectores de soporte*, comúnmente llamado SVM (del inglés *Support Vector Machine*), es un clasificador lineal que trata de maximizar la separación entre dos clases a través del uso de los denominados vectores de soporte. Como notación, se asumirá que los valores de entrada son $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$, con d atributos, y que su correspondiente salida binaria está dada por

$$y^{(i)} \in \{-1, +1\}.$$

Nótese que, a diferencia de los métodos vistos anteriormente, en este caso se utiliza como salida las etiquetas -1 (en lugar de 0) y $+1$. Esto resulta de utilidad debido a la forma en la que trabaja el SVM.

1. Márgenes

El clasificador llamado máquina de vectores de soporte (SVM) tiene dicho nombre debido a que el hiperplano de separación se obtiene maximizando el margen de dicho plano hacia los llamados vectores de soporte. En esta sección se presenta algunos conceptos que permiten formular el problema.

1.1. Hiperplano de Separación

En un clasificador binario lineal, el hiperplano de separación, también denominado *hiperplano de decisión* o *frontera de decisión*, es el hiperplano $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ que se utiliza para separar las dos clases de la siguiente manera.

- Si una instancia $\boldsymbol{x}^{(i)}$ queda en un lado de este hiperplano; es decir $\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b > 0$, esta instancia se clasifica como perteneciente a la clase 1.
- Si una instancia queda en el otro lado del hiperplano, $\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b < 0$, la instancia se clasifica como perteneciente a la clase -1.

En caso que la instancia se encuentra exactamente sobre el hiperplano, $\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b = 0$, no es posible determinar a qué clase pertenece, y su clasificación sería arbiraria (en la práctica es muy poco probable que se presente este caso debido a errores numéricos). Nótese que el

hiperplano tiene dos casos particulares: cuando se tiene dos atributos (x_1, x_2), el hiperplano se convierte en una recta, y cuando se tiene tres atributos (x_1, x_2, x_3), el hiperplano se convierte en un plano tridimensional.

Seguridad de la Predicción. Intuitivamente, mientras más se aleje una instancia del hiperplano de separación, más seguridad se tiene sobre la clase a la cual pertenece. Para cuantificar esta idea, se define la *seguridad de la predicción* $s\left(\boldsymbol{x}^{(i)}\right) \in \mathbb{R}$, asociada a la instancia $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$, como la "distancia" de la instancia al hiperplano. Si se tiene una instancia positiva, con $y^{(i)} = 1$, se puede considerar $\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b > 0$ como su medida de distancia. Por el contrario, si se tiene una instancia negativa, con $y^{(i)} = -1$, y dado que en este caso $\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b < 0$, se puede considerar $y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) > 0$ como su medida de distancia. Así, de manera general, la seguridad de predicción se expresa como

$$s\left(\boldsymbol{x}^{(i)}\right) = y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b).$$

Mientras más lejos se encuentre la instancia del hiperplano, más grande será el valor de $y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b)$ y más seguridad se tendrá sobre su predicción. Nótese que este valor es positivo siempre que la instancia correspondiente sea clasificada adecuadamente.

1.2. Vectores de Soporte

Los vectores de soporte son las instancias que se encuentran más cercanas al hiperplano de separación y, por tanto, son aquellas instancias donde se tiene menor seguridad en la predicción. Debido a esta baja seguridad, se convierten en las instancias más difíciles de clasificar. Esta potencial dificultad de clasificación hace que sean las instancias más críticas de todo el conjunto de entrenamiento, llegando a tener una influencia directa en la ubicación del hiperplano de separación más óptimo.

El clasificador SVM busca obtener el hiperplano que se encuentre más alejado de los vectores de soporte de ambas clases. Con este fin, además del hiperplano de separación $\boldsymbol{w}^T\boldsymbol{x} + b = 0$, se definen los hiperplanos paralelos $\boldsymbol{w}^T\boldsymbol{x} + b + \alpha = 0$ y $\boldsymbol{w}^T\boldsymbol{x} + b - \alpha = 0$ que pasan exactamente por los vectores de soporte. Como se puede notar, ambos hiperplanos son equidistantes del hiperplano de separación. Sin pérdida de generalidad, se suele denominar a estos dos hiperplanos de la siguiente manera:

- $\boldsymbol{w}^T\boldsymbol{x} + b = -1$: es el hiperplano que pasa exactamente por los vectores de soporte correspondientes a la clase -1.
- $\boldsymbol{w}^T\boldsymbol{x} + b = 1$: es el hiperplano que pasa exactamente por los vectores de soporte correspondientes a la clase +1.

El hecho de asignar de manera arbitraria un -1 y un +1 a estos hiperplanos no les quita generalidad, y no significa que se encuentren a una distancia de 1 del hiperplano de separación. Esta asignación (+1 y -1) es genérica debido a que al multiplicar por un factor la ecuación de un plano se obtiene otra representación del mismo plano y, de este modo, se podría usar cualquier otro valor en lugar de +1 y -1 con igual generalidad. En resumen, los vectores de soporte siempre se encuentran sobre el hiperplano

$$\boldsymbol{w}^T\boldsymbol{x} + b = \pm 1,$$

donde el signo se escoge según la clase (+1 o -1) a la cual pertenecen los respectivos vectores de soporte. Con este mismo criterio, se puede afirmar que los vectores de soporte satisfacen

$$y(\boldsymbol{w}^T\boldsymbol{x} + b) = 1,$$

lo cual se puede verificar algebraicamente para cada caso, o de manera intuitiva considerando que el producto de la clase clasificada adecuadamente justo en el margen, con su respectivo valor real, siempre será 1.

Ejemplo. Considérese el plano $x_1 - x_2$, con coordenadas $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$. La recta, en este plano, dada por $\boldsymbol{w}_1^T \boldsymbol{x} + b_1 = 1$, donde $\boldsymbol{w}_1 = \begin{bmatrix} -3 & 4 \end{bmatrix}$, y $b_1 = 2$, es equivalente a la recta dada por $\boldsymbol{w}_2^T \boldsymbol{x} + b_2 = 10$, donde $\boldsymbol{w}_2 = \begin{bmatrix} -30 & 40 \end{bmatrix}$, y $b_2 = 20$. Esta equivalencia se puede fácilmente demostrar multiplicando al primer plano por un factor de 10.

1.3. Margen

El margen γ de un hiperplano de separación $\boldsymbol{w}^T \boldsymbol{x} + b = 0$ es la distancia más pequeña que existe entre dicho hiperplano y cualquiera de las instancias del conjunto de entrenamiento que se desea clasificar. A partir de la definición de los vectores de soporte, se puede ver que el margen siempre estará asociado a un vector de soporte, ya que son las instancias más cercanas al hiperplano de separación y, por tanto, las instancias para las cuales existirá la menor distancia con respecto al hiperplano de separación.

Para determinar de manera matemática el valor del margen con respecto al hiperplano de separación, considérese que el vector de soporte denominado $\boldsymbol{z} \in \mathbb{R}^d$ se encuentra sobre el plano $\boldsymbol{w}^T \boldsymbol{x} + b = 1$, cumpliéndose $\boldsymbol{w}^T \boldsymbol{z} + b = 1$. El vector $\boldsymbol{z}$ proyectado en el hiperplano de separación será $\boldsymbol{z} - \gamma \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$, ya que la normal al hiperplano está dada por el vector unitario $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$, y la distancia de este vector de soporte $\boldsymbol{z}$ al hiperplano de separación es el valor del margen γ. Para esta proyección se cumplirá

$$\boldsymbol{w}^T \left(\boldsymbol{z} - \gamma \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \right) + b = 0.$$

Desarrollando el producto y agrupando adecuadamente se tiene $(\boldsymbol{w}^T \boldsymbol{z} + b) - \gamma \boldsymbol{w}^T \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} = 0$, donde el primer término en paréntesis es igual a 1 por definición. Con esta simplificación se llega a la expresión $1 - \gamma \boldsymbol{w}^T \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} = 0$, de donde se despeja el margen γ como

$$\gamma = \frac{\|\boldsymbol{w}\|}{\boldsymbol{w}^T \boldsymbol{w}} = \frac{1}{\|\boldsymbol{w}\|}, \tag{1}$$

debido a que $\boldsymbol{w}^T \boldsymbol{w} = \|\boldsymbol{w}\|^2$. Como se verá más adelante, es este margen el que busca maximizar un SVM.

2. SVM de Margen Duro

Un SVM de margen duro, denominado en inglés *Hard SVM*, es el SVM más simple y se basa en la suposición que ambas clases son linealmente separables. Por tanto, asume que existe un hiperplano capaz de separar a ambas clases de manera "perfecta".

2.1. Función de Hipótesis

Considerando que la seguridad de la predicción depende de la distancia al hiperplano de separación $\boldsymbol{w}^T \boldsymbol{x} + b = 0$, la máquina de vectores de soporte busca establecer un margen a ambos lados de este hiperplano, generando dos nuevos hiperplanos. Estos hiperplanos

se encuentran definidos como $\boldsymbol{w}^T\boldsymbol{x} + b = 1$ y $\boldsymbol{w}^T\boldsymbol{x} + b = -1$. Utilizando estos nuevos hiperplanos, la función de hipótesis $h_w(\boldsymbol{x})$ de un SVM se define como

$$h_w(\boldsymbol{x}) = \begin{cases} +1, & \boldsymbol{w}^T\boldsymbol{x} + b \geq 1 \\ -1, & \boldsymbol{w}^T\boldsymbol{x} + b \leq -1 \end{cases}. \tag{2}$$

Esta expresión indica que si $y = 1$, y la instancia asociada está bien clasificada, entonces $(1)(\boldsymbol{w}^T\boldsymbol{x} + b) \geq 1$. Igualmente, si $y = -1$ y la instancia asociada está bien clasificada, entonces $\boldsymbol{w}^T\boldsymbol{x} + b \leq -1$, o, $(-1)(\boldsymbol{w}^T\boldsymbol{x} + b) \geq 1$, donde se multiplicó por -1 a ambos términos. Considerando que y solo puede tomar el valor de 1 o de -1, ambas desigualdades se pueden escribir de manera compacta como

$$y(\boldsymbol{w}^T\boldsymbol{x} + b) \geq 1. \tag{3}$$

Esta expresión solo es válida cuando hay una correcta predicción en un SVM. En otros términos, la predicción en un SVM se considerará correcta si se cumple (3). Nótese que el valor de "1" es arbitrario y solamente referencial ya que el margen dependerá del valor que tenga el parámetro $\boldsymbol{w}$, tal como se muestra en (1).

2.2. Maximización del Margen

El objetivo de un SVM consiste en maximizar el margen γ, de tal modo que se tenga un clasificador más confiable, garantizando que siempre se realizará una clasificación correcta. De manera matemática esto se puede escribir como

$$\begin{aligned} &\max_{\boldsymbol{w},\gamma} \gamma \\ &\text{s.a.} \\ &\quad y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) \geq 1, \quad i = 1, \cdots, n \end{aligned}$$

donde s.a. significa "sujeto a" e indica la restricción al problema de optimización y se asume que hay n instancias de entrenamiento en total. Notar que la restricción $y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) \geq 1$ indica que cada instancia debe ser clasificada de manera correcta, lo cual es posible solamente si ambas clases son linealmente separables. Debido a esta restricción, a este problema de optimización se le conoce como *hard* SVM o SVM de margen duro.

De manera equivalente, dado que $\gamma = \frac{1}{\|\boldsymbol{w}\|}$, el problema de maximizar γ es equivalente al problema de minimizar $\|\boldsymbol{w}\|$. Usando esta equivalencia, se puede eliminar la variable γ del problema, reescribiendo este como

$$\begin{aligned} &\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{s.a.} \\ &\quad y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) \geq 1, \quad i = 1, \cdots, n. \end{aligned} \tag{4}$$

En este problema, por conveniencia matemática, se ha utilizado $\|\boldsymbol{w}\|^2$ en lugar de $\|\boldsymbol{w}\|$, ya que minimizar uno de ellos es equivalente a minimizar el otro. Igualmente, se ha añadido la constante $\frac{1}{2}$ por pura conveniencia matemática sin afectar el resultado final. Además, es posible demostrar que el problema dado en (4) es un problema de optimización convexa, dado que la función objetivo es convexa y las restricciones son lineales.

Si bien el problema de minimización dado en (4) es convexo, también es un problema cuadrático, ya que $\|\boldsymbol{w}\|^2 = \boldsymbol{w}^T\boldsymbol{w}$. Por este motivo, para su resolución se puede utilizar de manera directa paquetes computacionales que resuelvan problemas cuadráticos (llamados QPs, del inglés *Quadratic Program*).

2.3. Solución al Problema de Optimización

El problema de maximización de margen dado en (4) es equivalente a la minimización del vector $\boldsymbol{w}$. Este problema contiene una desigualdad, por lo que se puede utilizar las condiciones de Karush-Kuhn-Tucker (KKT), especificadas en el apéndice A. Estas condiciones no brindan una solución directa al problema, pero sí brindan condiciones que debe satisfacer la solución.

En el formato estándar del problema de optimización (apéndice A), la desigualdad se encuentra como menor o igual, pero en el problema de maximización del margen se encuentra como mayor o igual. Por este motivo, la restricción de desigualdad $y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)}+b) \geq 1$ se debe escribir como $1-y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)}+b) \leq 0$. Utilizando este arreglo, el Lagrangiano del sistema queda definido como

$$L(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b)), \tag{5}$$

donde α_i se conocen como los *multiplicadores de Lagrange*. Luego, las condiciones KKT establecen que la solución al problema de optimización debe satisfacer las siguientes igualdades y desigualdades:

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w}^* - \sum_{i=1}^{n} \alpha_i y^{(i)} \boldsymbol{x}^{(i)} = 0 \tag{6}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i y^{(i)} = 0 \tag{7}$$

$$\alpha_i \geq 0 \tag{8}$$

$$\alpha_i (1 - y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b^*)) = 0 \tag{9}$$

$$1 - y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b^*) \leq 0, \tag{10}$$

donde los asteriscos hacen referencia a que son los valores óptimos los que satisfacen dichas condiciones.

Para satisfacer la condición (9), alguno de los dos términos tiene que ser cero. Se sabe a partir de (8) que α_i o es 0 o es positivo. Cuando $\alpha_i = 0$, no se puede hacer ninguna afirmación sobre el valor del término $1 - y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b^*)$ a partir de (9). Sin embargo, cuando $\alpha_i > 0$, entonces sí se debe tener que $1-y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)}+b^*) = 0$ o, equivalentemente, $y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)}+b^*) = 1$. Esto significa que las instancias i asociadas a $\alpha_i > 0$ son los *vectores de soporte*, los cuales serán representados como $\boldsymbol{x}_{sv}^{(i)}$ (del inglés *support vector*). Igualmente, los valores asociados a estos vectores de soporte serán denominados $y_{sv}^{(i)}$. El número total de vectores de soporte que existe será el número de elementos α_i que son mayores que cero, y se denominará N_{sv}.

Por otro lado, a partir de (7) se obtiene la condición

$$\sum_{i=1}^{n} \alpha_i y^{(i)} = 0, \tag{11}$$

la cual establece una restricción que deben satisfacer los multiplicadores de Lagrange α_i en función de sus respectivas salidas deseadas $y^{(i)}$. Esta es una restricción que será de utilidad al calcular la expresión final del Lagrangiano.

Valor óptimo de $\boldsymbol{w}$. A partir de la condición dada en (6) se obtiene que el valor óptimo del vector de parámetros de entrenamiento $\boldsymbol{w}$, denotado como $\boldsymbol{w}^*$, debe satisfacer

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i y^{(i)} \boldsymbol{x}^{(i)}. \tag{12}$$

En esta sumatoria, se considera las n instancias de entrenamiento. Sin embargo, debido a la condición (8), algunos α_i son cero, anulando su respectivo término en la sumatoria, mientras que otros son positivos. Dado que (9) establece que cada $\alpha_i > 0$ está asociado a un vector de soporte, se concluye de (12) que el vector de parámetros $\boldsymbol{w}$ óptimo será una combinación lineal solamente de los vectores de soporte, y no de todas las instancias de entrenamiento. Así, (12) se puede escribir como

$$\boldsymbol{w}^* = \sum_{i=1}^{N_{sv}} \alpha_i y_{sv}^{(i)} \boldsymbol{x}_{sv}^{(i)}, \tag{13}$$

donde $\boldsymbol{x}_{sv}^{(i)}$ son los vectores de soporte, $y_{sv}^{(i)}$ son sus valores reales asociados, y N_{sv} es el número total de vectores de soporte. Este resultado refuerza el papel crucial que juegan los vectores de soporte en un SVM, ignorando el resto de instancias de entrenamiento.

Valor óptimo de b. Considerando que se tiene calculado el valor $\boldsymbol{w}^*$ óptimo, el valor de b^* óptimo se puede obtener a partir de (9) para $\alpha_i > 0$; es decir, para los vectores de soporte. Despejando b^* de $y_{sv}^{(i)}(\boldsymbol{w}^T \boldsymbol{x}_{sv}^{(i)} + b^*) = 1$, y utilizando cualquier vector de soporte, se obtiene

$$b^* = y_{sv}^{(i)} - \boldsymbol{w}^T \boldsymbol{x}_{sv}^{(i)},$$

donde la simplificación ha considerado que $\frac{1}{y} = y$, debido a que solamente interesa el signo de $y = \{+1, -1\}$. Sin embargo, una manera más robusta de calcular b^* en la práctica resulta de promediar todos los valores de b obtenidos a partir de los diferentes vectores de soporte. Esto es:

$$b^* = \frac{1}{N_{sv}} \sum_{i=1}^{N_{sv}} y_{sv}^{(i)} - \boldsymbol{w}^T \boldsymbol{x}_{sv}^{(i)}, \tag{14}$$

donde N_{sv} es el número total de vectores de soporte.

Desarrollo del Lagrangiano. El Lagrangiano (5) del problema de optimización del SVM se puede desarrollar término a término llegando a

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y^{(i)} \boldsymbol{w}^T \boldsymbol{x}^{(i)} - b \sum_{i=1}^{n} \alpha_i y^{(i)},$$

donde se ha reemplazado $\|\boldsymbol{w}\|^2$ por su equivalente $\boldsymbol{w}^T\boldsymbol{w}$, y se ha definido el vector $\boldsymbol{\alpha}$ como $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{bmatrix}^T$. En esta expresión del Lagrangiano se elimina el último término al utilizar la condición dada en (11). Luego, al reemplazar el valor de $\boldsymbol{w}$ dado en (12), se llega a

$$L(\boldsymbol{\alpha}) = \frac{1}{2} \sum_{i=1}^{n} \alpha_i y^{(i)} \boldsymbol{x}^{(i)T} \sum_{j=1}^{n} \alpha_j y^{(j)} \boldsymbol{x}^{(j)} + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y^{(i)} \left(\sum_{j=1}^{n} \alpha_j y^{(j)} \boldsymbol{x}^{(j)T} \right) \boldsymbol{x}^{(i)},$$

donde se observa que el primero y el último término son semejantes. Notar que se ha reemplazado la expresión (12) para $\boldsymbol{w}$, usando todos los valores de α_i, y no se ha utilizado (13), que es la expresión simplificada, dado que en este punto aún no se conoce los valores de α_i (y, por tanto, se desconoce cuáles α_i son mayores que cero). Obsérvese además que ahora el Lagrangiano solamente depende de α, ya que todos los demás términos son obtenidos del conjunto de entrenamiento. Sumando estos términos, y ordenando tanto las sumatorias como los términos dentro de las mismas, se obtiene

$$L(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}\boldsymbol{x}^{(i)T}\boldsymbol{x}^{(j)} + \sum_{i=1}^{n}\alpha_i. \tag{15}$$

Dado que ahora este Lagrangiano solamente depende de los multiplicadores de Lagrange α_i, también llamados variables duales, puede ser utilizado para formular el problema dual que permita calcular los valores de α_i.

Problema Dual. El problema dual consiste en maximizar el Lagrangiano dado en (15) sujeto a las restricciones (8) y (9) de las condiciones KKT, que es donde aparece el valor de α_i. De manera concreta, este problema se define como

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}\boldsymbol{x}^{(i)T}\boldsymbol{x}^{(j)} \right\} \tag{16}$$

s.a.

$$\sum_{i=1}^{n}\alpha_i y^{(i)} = 0$$

$$\alpha_i \geq 0, \quad i = 1, \cdots, n.$$

Este es nuevamente un problema cuadrático que puede ser resuelto con algún paquete que solucione QPs. La ventaja de este problema dual es que solamente el valor de $\boldsymbol{\alpha}$ está involucrado. Además, se tiene el término $\boldsymbol{x}^{(i)T}\boldsymbol{x}^{(j)} = \boldsymbol{x}^{(i)} \cdot \boldsymbol{x}^{(j)}$ en la función objetivo, el cual indica que para nuevas instancias solamente es de importancia el producto punto entre estas. Lo importante de esto es que luego se puede definir funciones sobre las instancias, llamadas *kernels*, y solamente será importante la definición del producto punto de estos kernels.

2.4. Implementación de la Solución

A manera de resumen de la sección anterior, para implementar la solución al problema del SVM de margen duro se debe seguir los siguientes pasos:

1. Se calcula los valores óptimos de α_i usando un programa de optimización cuadrática para resolver el problema de optimización dado en (16).

2. Se calcula el vector óptimo de pesos $\boldsymbol{w}^*$, utilizando (13) y los valores óptimos obtenidos para cada α_i.

3. Se calcula el sesgo óptimo b^* usando (14) y el valor de $\boldsymbol{w}^*$ obtenido.

Una vez que se realiza estos cálculos, ya se puede aplicar el clasificador y se dice que el SVM se encuentra entrenado.

Como se observa, el problema principal radica en el cálculo de α_i, ya que una vez calculados, los valores de $\boldsymbol{w}$ y b son fácilmente obtenidos. Esto significa, que de manera práctica, se debe resolver el problema dado en (16). La resolución implica expresar la función objetivo explícitamente como una función cuadrática, para poder luego aplicar paquetes que resuelven QPs. Dado que se definió $\boldsymbol{\alpha} = \begin{bmatrix}\alpha_1 & \alpha_2 & \cdots & \alpha_n\end{bmatrix}^T$, se puede escribir la sumatoria de los términos α_i como

$$\sum_{i=1}^{n} \alpha_i = \alpha_1 + \alpha_2 + \cdots + \alpha_n = \boldsymbol{v}_1^T \boldsymbol{\alpha}, \tag{17}$$

donde $\boldsymbol{v}_1^T = \begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}$ es un vector que contiene n unos. Por otro lado, dado que la doble sumatoria de la función objetivo de (16) es, en realidad, una forma cuadrática, se puede escribir como

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \boldsymbol{x}^{(i)T} \boldsymbol{x}^{(j)} = \boldsymbol{\alpha}^T M \boldsymbol{\alpha},$$

donde

$$M = \begin{bmatrix} y^{(1)}y^{(1)}\boldsymbol{x}^{(1)T}\boldsymbol{x}^{(1)} & y^{(1)}y^{(2)}\boldsymbol{x}^{(1)T}\boldsymbol{x}^{(2)} & \cdots & y^{(1)}y^{(n)}\boldsymbol{x}^{(1)T}\boldsymbol{x}^{(n)} \\ y^{(2)}y^{(1)}\boldsymbol{x}^{(2)T}\boldsymbol{x}^{(1)} & y^{(2)}y^{(2)}\boldsymbol{x}^{(2)T}\boldsymbol{x}^{(2)} & \cdots & y^{(2)}y^{(n)}\boldsymbol{x}^{(2)T}\boldsymbol{x}^{(n)} \\ \vdots & \vdots & \ddots & \vdots \\ y^{(n)}y^{(1)}\boldsymbol{x}^{(n)T}\boldsymbol{x}^{(1)} & y^{(n)}y^{(2)}\boldsymbol{x}^{(n)T}\boldsymbol{x}^{(2)} & \cdots & y^{(n)}y^{(n)}\boldsymbol{x}^{(n)T}\boldsymbol{x}^{(n)} \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

Debido a la organización de los términos de M, y a su simetría, la matriz M se puede, a su vez, representar como el producto exterior $M = X_y X_y^T$, donde

$$X_y = \begin{bmatrix} y^{(1)}\boldsymbol{x}^{(1)T} \\ y^{(2)}\boldsymbol{x}^{(2)T} \\ \vdots \\ y^{(n)}\boldsymbol{x}^{(n)T} \end{bmatrix} \in \mathbb{R}^{n\times d},$$

recordando que se tiene n instancias de $\boldsymbol{x}$, y que cada $\boldsymbol{x}$ tiene d atributos. Notar que en X_y, cada elemento $y^{(i)}\boldsymbol{x}^{(i)T}$ constituye una fila.

Finalmente, las restricciones que aparecen en (16) también deben ser reescritas en forma matricial. La sumatoria que aparece en la restricción se puede escribir como

$$\sum_{i=1}^{n} \alpha_i y^{(i)} = \boldsymbol{y}^T \boldsymbol{\alpha},$$

donde $\boldsymbol{y}^T = \begin{bmatrix}y^{(1)} & y^{(2)} & \cdots & y^{(n)}\end{bmatrix}$. Todas las restricciones $\alpha_i \geq 0$ pueden ser escritas de manera compacta como $I\boldsymbol{\alpha} \succeq \mathbf{0}$, donde I es la matriz identidad de $n \times n$, $\mathbf{0} \in \mathbb{R}^n$ es un vector columna con n ceros, y $\succeq$ indica la desigualdad término a término.

Utilizando estas expresiones matriciales y vectoriales, el problema dual de maximización mostrado en (16), que se usa para encontrar el valor de $\boldsymbol{\alpha}$ se convierte en

$$\max_{\boldsymbol{\alpha}} \left\{ \boldsymbol{v}_1^T \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}^T X_y X_y^T \boldsymbol{\alpha} \right\} \tag{18}$$

s.a.

$$\boldsymbol{y}^T \boldsymbol{\alpha} = 0$$
$$I\boldsymbol{\alpha} \succeq \mathbf{0}.$$

La expresión (18) se encuentra escrita de manera explícita como un problema cuadrático de optimización (QP), donde la función objetivo es cuadrática, y las restricciones son lineales. Es esta expresión la que se utilizará, junto con algún paquete computacional que resuelva QPs, para el cálculo de los valores de cada α_i.

Anexo A: Condiciones de Karush-Kuhn-Tucker (KKT)

Dado un problema de minimización

$$\begin{aligned} &\min_{\boldsymbol{x}} f(\boldsymbol{x}) \\ \text{s.a.} \quad & \\ & h_i(\boldsymbol{x}) \leq 0, \quad i = 1, \cdots, m \\ & g_j(\boldsymbol{x}) = 0, \quad i = 1, \cdots, r, \end{aligned}$$

donde $x \in \mathbb{R}^n$, se define el *Lagrangiano* como

$$L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = f(x) + \sum_{i=1}^{m} u_i h_i(\boldsymbol{x}) + \sum_{j=1}^{r} v_j g_j(\boldsymbol{x})$$

donde $\boldsymbol{u} = (u_1, u_2, \cdots, u_m)$ y $\boldsymbol{v} = (v_1, v_2, \cdots, v_r)$.

Las soluciones $\boldsymbol{x}^*$ a este problema de minimización deben satisfacer las llamadas condiciones de Karush-Kuhn-Tucker (usualmente llamadas condiciones KKT), las cuales quedan definidas como

$$\begin{aligned} \frac{\partial L(\boldsymbol{x}^*, \boldsymbol{u}^*, \boldsymbol{v}^*)}{\partial x_i} &= 0 \\ u_i^* &\geq 0 \\ u_i^* h_i(\boldsymbol{x}^*) &= 0 \\ h_i(\boldsymbol{x}^*) &\leq 0 \\ g_j(\boldsymbol{x}^*) &= 0, \end{aligned}$$

donde $i = 1, \cdots, m$, y $j = 1, \cdots, r$, y donde u^*,$\boldsymbol{v}^*$ representan los valores óptimos de los coeficientes de Lagrange.

Anexo B: Uso de un paquete Computacional para QPs

La mayoría de paquetes de software disponibles para la resolución de QPs formulan el problema cuadrático de optimización genérico de la siguiente manera:

$$\begin{aligned} &\min_{\boldsymbol{x}} \left\{ \frac{1}{2} \boldsymbol{x}^T P \boldsymbol{x} + \boldsymbol{q}^T \boldsymbol{x} \right\} \\ \text{s.a.} \quad & \\ & A\boldsymbol{x} = \boldsymbol{b} \\ & G\boldsymbol{x} \preceq \boldsymbol{h}, \end{aligned}$$

de tal modo que es necesario adaptar el problema que se desea resolver a esta forma; es decir, se tendrá que encontrar los valores P, $\boldsymbol{q}$, A, $\boldsymbol{b}$, G y $\boldsymbol{h}$.

Para el SVM de margen duro, se vio que el problema principal se reducía a la solución del problema dual de maximización, a partir del cual se calcula los valores de los coeficientes de Lagrange α_i. Es sabido que el maximizar una función f es equivalente a minimizar una función $-f$. por lo que el problema de maximización de (18) se puede fácilmente convertir en uno de minimización invirtiendo el signo de la función objetivo. Luego de hacer esta modificación, se puede comparar (18) con la expresión general del problema de minimización, encontrando: $P = X_y X_y^T$, $\boldsymbol{q} = -\boldsymbol{v}_1^T$, $A = \boldsymbol{y}^T$, $\boldsymbol{b} = 0$, $G = -I$, $\boldsymbol{h} = \mathbf{0}$.

Es importante notar que existen optimizadores especialmente diseñados para trabajar con los problemas cuadráticos que aparecen en la obtención de los parámetros de un SVM. El más usado es el llamado SMO (*Sequential Minimal Optimization*), el cual divide el problema de optimización cuadrática en subproblemas pequeños, los cuales se resuelven analíticamente evitando iteraciones internas de los algoritmos de optimización. Este algoritmo es bastante eficiente, especialmente si los datos son dispersos.