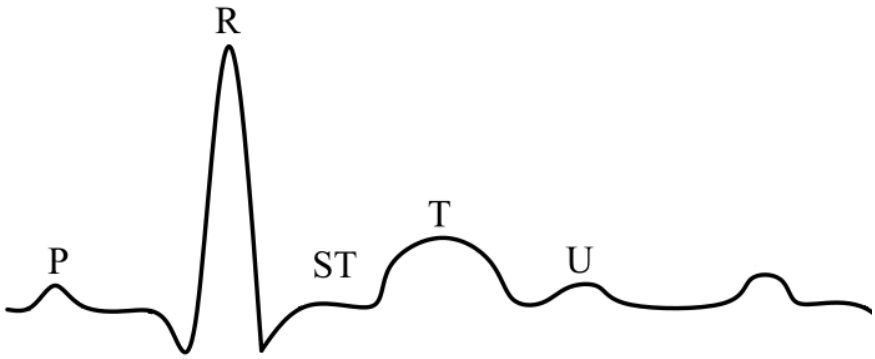
Figura 6

En el trazado tras la onda T puede observarse la onda U



Nota: tomado de Palma (1998)

CAPÍTULO 3. MÁQUINAS DE VECTORES DE SOPORTE

La clasificación de patrones se define como la tarea de categorizar algún objeto dentro de una de las categorías dadas llamadas clases, a partir de un conjunto de patrones asociados a cada objeto. Usamos el término "patrón" para denotar un vector de datos x de dimensión p, donde $x = (x_1, x_p)T$ cuyos componentes xi son las medidas de las características de un objeto. Estas características son las variables especificadas por los investigadores, debido a que por lo regular tienen un peso importante en los resultados de la clasificación.

En general, existen dos enfoques principales de clasificación: clasificación supervisada y clasificación no supervisada. La clasificación no supervisada también es referida frecuentemente como agrupamiento. En este tipo de clasificación, los datos no son etiquetados y se desean encontrar grupos en los datos que se distingan unos de otros a partir de las características. En la clasificación supervisada tenemos un conjunto de datos de prueba, cada uno de estos consiste en medidas sobre un conjunto de variables y asociado a cada dato una etiqueta que define la clase del objeto.

Las redes neuronales, árboles de decisión y SVM son clasificadores de aprendizaje

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

supervisado. Los métodos de aprendizaje supervisado emplean un conjunto de pares entrada-salida, estos clasificadores adquieren una función de decisión que asocia a un nuevo dato una etiqueta de clase dentro de las clases dadas.

En este Capítulo son definidas las características teóricas de las SVM para problemas de clasificación con dos clases. Primero, se definen las funciones decisión y su importancia al generalizar, después se explican las Máquinas de Vectores Soporte con margen duro, para conjuntos de datos de entrenamiento linealmente separables en el espacio de entrada. Una vez concluido esto, se extiende a el caso linealmente no separable y es necesario trasladar el espacio de datos de entrada a un espacio de características altamente dimensional con el propósito de separar linealmente el espacio de características.

3.1 Funciones de decisión

Se considera que el problema de clasificación de un punto cuyas características están dadas por el vector x tal que x = (x_1, ... , x_p)T y este pertenece a una de dos clases posibles. Suponga que se tiene las funciones f_1(x) y f_2(x) que definen las clases 1 y 2 y se clasifica al punto x dentro de la clase 1 si:

$$f_1(x) > 0, f_2(x) < 0 \quad (3)$$

o se clasifica al punto x dentro de la clase 2 si:

$$f_1(x) < 0, f_2(x) > 0, \quad (4)$$

A estas funciones se las llama funciones de decisión. Al proceso de encontrar las funciones de decisión a partir de pares de entrada-salida es llamado entrenamiento. Los métodos convencionales de entrenamiento determinan las funciones de decisión de tal forma que cada par entrada-salida sea correctamente clasificado dentro de la clase a la que pertenece. Por ejemplo, asumiendo que los cuadros pertenecen a la clase 1 y los círculos pertenecen a la clase 2, resulta claro que los datos de entrenamiento no se interceptan en ningún momento y es posible trazar una línea separando los datos de manera perfecta. Sin embargo, ya sea que la función de decisión $f_1(x)$ o la función $f_2(x)$ se muevan hacia la línea punteada de su propio lado, el conjunto de datos de entrenamiento aún sigue siendo

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

correctamente clasificado, dando la certeza de que es posible encontrar un conjunto infinito de hiperplanos que correctamente clasifiquen los datos de entrenamiento. Sin embargo, es claro que la precisión de clasificación al generalizar será directamente afectada por la posición de las funciones de decisión.

Las SVM a diferencia de otros métodos de clasificación consideran esta desventaja y encuentra la función de decisión de tal forma que la distancia entre los datos de entrenamiento es maximizada. Esta función de decisión es llamada función de decisión óptima o hiperplano de decisión óptima.

3.2 Funciones de decisión en las SVM

Recientemente ha habido un incremento impresionante en el número de artículos de investigación sobre SVM. Las SVM han sido aplicadas exitosamente a un gran número de procesos yendo desde identificación de partículas, identificación de rostros y categorización de texto, hasta bioinformática y medicina. El enfoque es motivado por la teoría de aprendizaje estadístico, pues las SVM producen modelos matemáticos elegantes que son geométricamente intuitivos y teóricamente bien fundamentados.

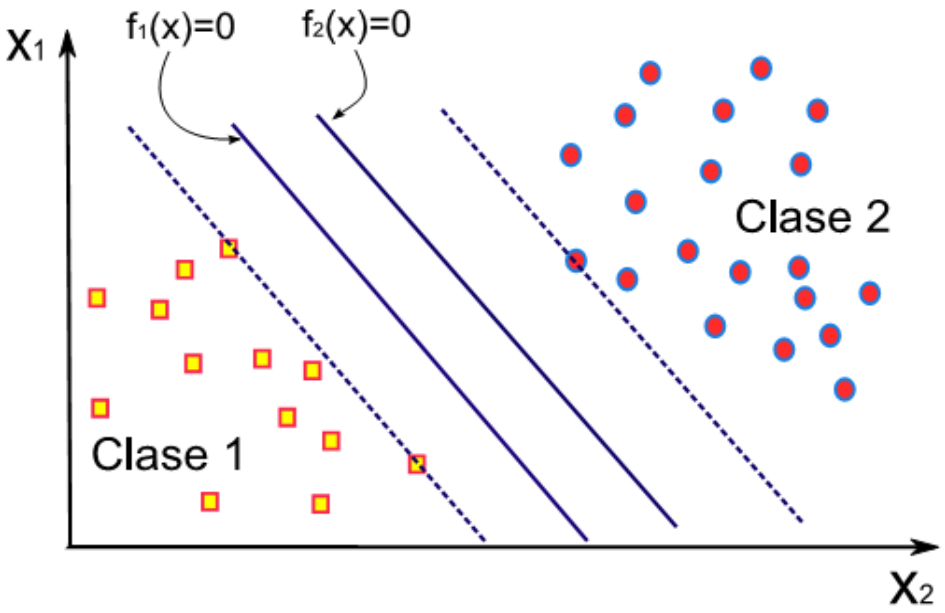
La principal motivación de las SVM es separar varias clases en el conjunto de entrenamiento con una superficie que maximice el margen entre estas. Esta es una implementación del principio de minimización estructural que permite minimizar una cota sobre el error de generalización de un modelo, en lugar de minimizar el error medio cuadrático sobre el conjunto de datos de entrenamiento, que es la filosofía que usan a menudo los métodos de minimización de riesgo empírico.

Entrenar una SVM requiere un conjunto de n ejemplos. Cada ejemplo consiste en un vector de entrada Xi y una etiqueta Yi asociada al vector de entrada. La función de la SVM que tiene que ser entrenada con los ejemplos contiene n parámetros libres, los llamados multiplicadores de Lagrange positivos αi, i = 1, ..., n. Cada αi es una medida de cuanto, el correspondiente ejemplo de entrenamiento influye en la función. La mayoría de los ejemplos no afectan la función y consecuentemente la mayoría de los αi son cero. Ver figura 7:

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

Figura 7

Funciones de las SVM



Nota: elaboración propia.

3.3 Caso linealmente separable

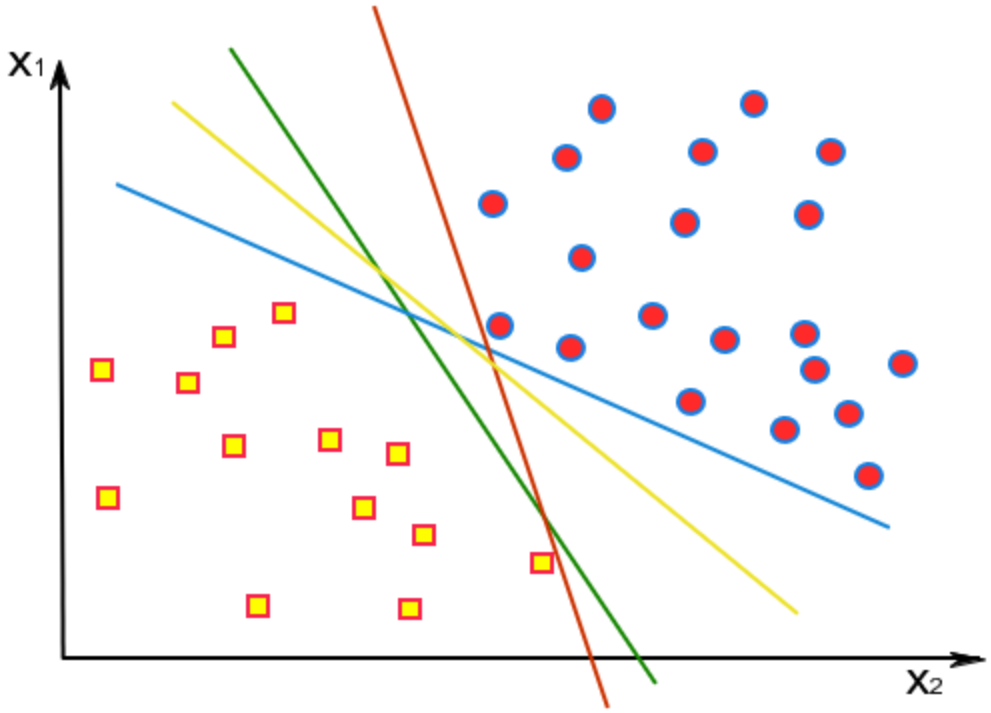
Se considera el problema de clasificación binaria en donde los datos (ver figura 8) de entrenamiento son dados como:

$$(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l), x \in R^n, y \in \{+1, -1\} \quad (5)$$

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

Figura 8

Clasificador convencional



Nota: elaboración propia.

Por razones de visualización, se considera el caso de un espacio de entrada bidimensional- al, i.e., $x \in R2$. Los datos son linealmente separables y existen diferentes hiperplanos que pueden realizar la separación. La Figura 7, muestra varios hiperplanos de decisión que separan perfectamente el conjunto de datos de entrada. Es claro que existe un número infinito de hiperplanos que podrían realizar este trabajo. Sin embargo, la habilidad de generalización depende de la localización del hiperplano de separación y el hiperplano con máximo margen es llamado hiperplano de separación óptima. La cota de decisión, esto es la línea que separa el espacio de entrada es definida por la ecuación $wT\, x_i + b = 0$. Sin embargo, el problema radica en encontrar la mejor cota de decisión, esto es, la función de separación óptima.

El caso más simple de SVM es el caso linealmente separable en el espacio de características. Se optimiza el margen geométrico fijando para ello el margen funcional $\kappa_i =$

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

1 (también llamado Hiperplano Canónico [20]), por lo tanto, el clasificador de línea

$$\begin{aligned} y_i &= \pm 1, \\ \langle w.x^+ \rangle + b &= 1 \\ \langle w.x^- \rangle + b &= -1 \end{aligned} \tag{6}$$

Estos pueden ser combinados dentro de un conjunto de desigualdades:

$$y_1(\langle w.x^+ \rangle + b \geq 1 \forall i \tag{7}$$

El margen geometrico de x^+y x^- es:

$$\begin{aligned} \gamma_i &= \frac{1}{2}\left(\langle \frac{w}{\|w\|}.x^+ \rangle - \langle \frac{w}{\|w\|}.x^- \rangle\right) \\ &= \frac{1}{2\|w\|}(\langle w.x^+ \rangle - \langle w.x^- \rangle) \\ &= \frac{1}{\|w\|} \end{aligned} \tag{8}$$

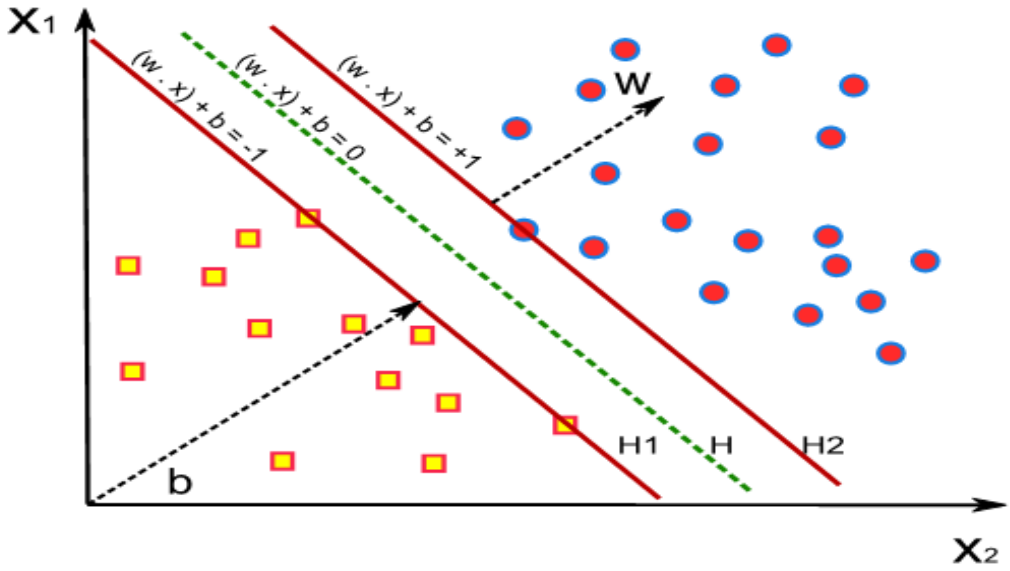
donde w define el hiperplano de separación óptima y b es el sesgo. La distancia entre el hiperplano de separación y el dato de entrenamiento más cercano al hiperplano es llamada margen. La habilidad de generalización depende de la localización del hiperplano de separación y el hiperplano con máximo margen es llamado hiperplano de separación óptima. Es intuitivamente claro que la habilidad de generalización es maximizada si el hiperplano de separación óptima es seleccionado como el hiperplano de separación. Optimizar el margen geométrico significa minimizar la norma del vector de pesos. Al resolver el problema de programación cuadrática se trata de encontrar el hiperplano óptimo y dos hiperplanos (H1 y H2) paralelos. Las distancias entre H1 y H2 es maximizada y no existe ningún dato entre los dos hiperplanos. Cuando la distancia entre H1 y H2 es maximizada, algunos puntos de datos pueden estar sobre H1 y algunos puntos de datos pueden estar sobre H2. Estos puntos de datos son llamados vectores soporte, ya que participan de forma directa en definir el hiperplano de separación, los otros puntos pueden ser removidos o cambiados sin cruzar los planos H1 y H2 y no modificarán de alguna forma la habilidad de generalización del clasificador esto es, la solución de una SVM está dada únicamente por este pequeño conjunto

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

de vectores soporte. Ver figura 9:

Figura 9

Representación de las SVM mediante W, X y B



Nota: elaboración propia.

Cualquier hiperplano puede ser representado mediante w, x y b, donde w es un vector perpendicular al hiperplano. La Figura 9 muestra la representación geométrica del problema de programación cuadrática mostrando H (separador óptimo) y los hiperplanos H1 y H2. De esta forma, el problema original de optimización queda de la siguiente manera.

Proposición 1: Para el caso linealmente separable S = $[(x_1, y_1) \cdots (x_l , y_l)]$, si el hiperplano (w, b) es la solución de:

$$\min_{w,b} \langle w \cdot w \rangle = \|w\|^2$$

$$\text{Sujeto } a\text{: } y_i (\langle w \cdot x_i \rangle + b) \geq 1 \tag{9}$$

entonces el hiperplano tiene un margen máximo (geométrico)

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

$$\gamma = \frac{1}{\|w\|} \tag{10}$$

Ahora se cambia al problema dual utilizando la formulación de Lagrange. Existen dos razones para hacer esto. La primera radica en el hecho de que las condiciones dadas serán reemplazadas por multiplicadores de Lagrange, que son mucho más fáciles de manejar. La segunda proviene de que, en la reformulación del problema, los datos de entrenamiento únicamente aparecerán en la forma de producto punto entre vectores. Esta es una propiedad fundamental que permitirá generalizar el procedimiento en el caso no lineal. De esta manera, el Lagrangiano está dado por:

$$L(w, b, \alpha) \equiv \frac{1}{2}\langle w \cdot w\rangle - \sum_{i=1}^{l} \alpha_i [y_i (\langle w \cdot x_i\rangle + b) - 1] \tag{11}$$

donde α_i son multiplicadores de Lagrange.

El dual es encontrado en dos pasos: primero, diferenciando con respecto a w y b

$$\begin{aligned} \frac{\partial L(w, b, \alpha)}{\partial w} &= w - \sum_{i=1}^{l} \alpha_i y_i x_i = 0 \rightarrow w = \sum_{i=1}^{l} \alpha_i y_i x_i \\ \frac{\partial L(w, b, \alpha)}{\partial b} &= -\sum_{i=1}^{l} \alpha_i y_i = 0 \rightarrow w = \sum_{i=1}^{l} \alpha_i y_i = 0 \end{aligned} \tag{12}$$

y segundo, restituyendo las relaciones obtenidas en el Lagrangiano original

$$\begin{aligned} L(w, b, \alpha) &\equiv \frac{1}{2}\langle w \cdot w\rangle - \sum_{i=1}^{l} \alpha_i [y_i (\langle w \cdot x_i\rangle + b) - 1] \\ &= \frac{1}{2}\langle \sum_{i=1}^{l} \alpha_i y_i x_i * \sum_{i=1}^{l} \alpha_i y_i x_i\rangle - \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \left(\langle x_j \cdot x_i\rangle + b\right) - \sum_{i=1}^{l} \alpha_i \\ &= \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i y_i \alpha_j y_j \langle x_i \cdot x_j\rangle - \sum_{i,j=1}^{l} \alpha_i y_i \alpha_j y_j \langle x_j \cdot x_i\rangle - \sum_{i=1}^{l} \alpha_i y_i b + \sum_{i=1}^{l} \alpha_i \\ &= -\frac{1}{2}\sum_{i,j=1}^{l} \alpha_i y_i \alpha_j y_j \langle x_i \cdot x_j\rangle + \sum_{i=1}^{l} \alpha_i \end{aligned} \tag{13}$$

Aquellos puntos para los cuales $\alpha i > 0$ son llamados "vectores soporte" y quedan en uno de los hiperplanos H1, H2. En todos los otros puntos de entrenamiento $\alpha i = 0$ y soportan o quedan sobre H1 o H2 de tal forma que las condiciones de la ecuación 11 se cumplen. Los vectores soporte son los elementos críticos del conjunto de entrenamiento y estos son los más cercanos a la cota de decisión.

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

Comentario: Al entrenar el conjunto inicial de datos se obtiene un hiperplano, que separa perfectamente estos datos y es definido por un pequeño conjunto de vectores soporte. Si todos los demás puntos fueran eliminados (o desplazados alrededor sin cruzar H1 o H2) y el entrenamiento fuera repetido, se encontraría el mismo hiperplano de separación definido por el mismo conjunto de vectores soporte.

Por lo tanto, el problema original de optimización queda de la siguiente manera.

Proposición 2: Para el caso linealmente separable S = $[(x_1, y_1) \ (x_l, y_l)]$, si α∗

la solución del problema de optimización cuadrático

$$\max_{\alpha_i} -\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i y_i \alpha_j y_j \langle x_i \cdot x_j\rangle + \sum_{i=1}^{l}\alpha_i \text{ sujeto a: } \sum_{i=1}^{l}\alpha_i y_i = 0 \quad (14)$$

entonces $\|w\|^2$ comprende el mínimo $w^* = \sum_{i=1}^{l}\alpha_i^* y_i x_i$ y el margen geométrico $\gamma^* = \frac{1}{\|w^*\|}$ es maximizado.

Condiciones de Karush- Kuhn – Tucker. Las condiciones de Karush-Kuhn-Tucker (KKT) juegan un rol muy importante en la teoría de optimización, ya que dan las condiciones para obtener una solución óptima a un problema de optimización general.

Teorema 1 Dado un problema de optimización con dominio convexo Ω ⊆ Rn

$$minimizar\ f(w), \qquad w \in \Omega$$

$$sujeto\ a \quad \begin{matrix} g_i(w) \le 0, & i=1, \ldots, k, \\ h_i(w) = 0, & i=1, \ldots, m, \end{matrix}$$

con f ∈ w convexa, las condiciones necesarias y suficientes para que un punto normal w^* sea un óptimo son la existencia de α^*, β^* tal que

$$\begin{aligned} \frac{\partial L(w^*, \alpha^*, \beta^*)}{\partial w} &= 0 \\ \frac{\partial L(w^*, \alpha^*, \beta^*)}{\partial \beta} &= 0 \alpha_i^* g_i(w^*) = 0, i = 1, \ldots, k, \\ g_i(w^*) &\le 0, i = 1, \ldots, k, \\ \alpha_i^* &\ge 0, i = 1, \ldots, k. \end{aligned} \quad (15)$$

Por lo tanto, la distancia máxima de un hiperplano es:

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

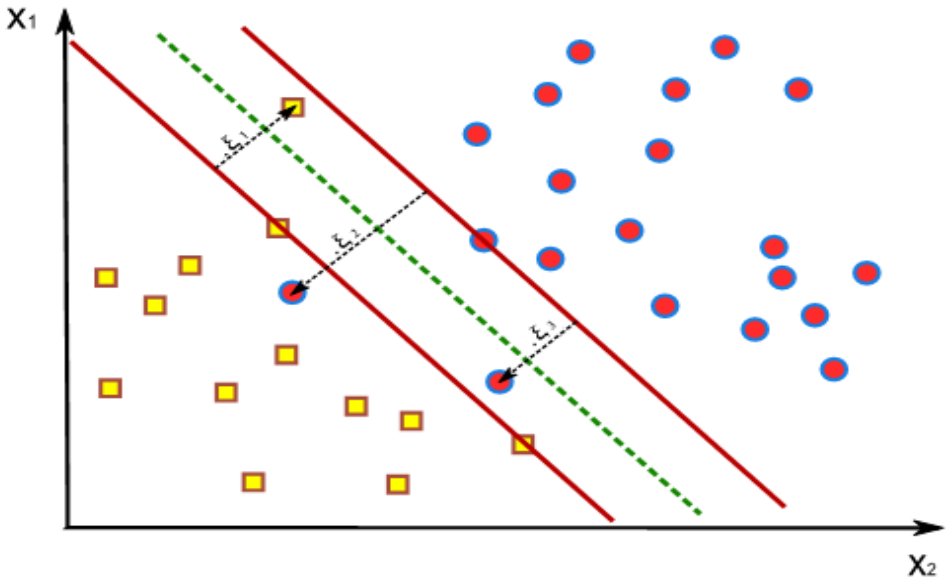
$$\frac{1}{\|w^*\|} = \left(\sum_{i\in\delta v} \alpha_i^*\right)^{-\frac{1}{2}}$$
$$\|w^*\|^2 = \langle w^* \cdot w^* \rangle = \left(\sum_{i\in\delta v} \alpha_i^*\right)^{-\frac{1}{2}} \quad (16)$$

3.4 Hiperplanos con márgenes blandos

El problema de aprendizaje presentado anteriormente es válido para el caso donde los datos son linealmente separables, que significa que el conjunto de datos de entrenamiento no tiene intersecciones. Sin embargo, este tipo de problemas son raros en la práctica. Al mismo tiempo, existen algunos ejemplos en los que el hiperplano de separación lineal puede dar buenos resultados aun cuando los datos se intersecan. Sin embargo, las soluciones de programación cuadrática como están dadas anteriormente no pueden ser usadas en el caso de intersección ya que la condición $y_i(hw \cdot x_{ii} + b) \geq 1 \; \forall i$ no puede ser satisfecha en el caso de intersección (ver figura 10). Los puntos que se encuentran en la intersección no pueden ser correctamente clasificados y para cualquier dato mal clasificado x_i, su correspondiente α_i tenderá a infinito.

Figura 10

Hiperplano con margen blando



Nota: elaboración propia.

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

Para encontrar un clasificador con margen máximo, el algoritmo presentado anteriormente deberá ser cambiado permitiendo un margen blando, por lo tanto es necesario introducir variables flojas no negativas $\xi_i(\geq 0)$.

$$y_i(\langle w^T \cdot x_i \rangle + b) \geq 1 - \xi_i \qquad \forall_i \tag{17}$$

Mediante las variables ξ_i, la solución factible siempre existe. Para los datos de entrenamiento x_i, si $0 < \xi_i < 1$ los datos no poseen el margen máximo, pero pueden ser correctamente clasificados. Por otro lado, el ancho de este margen blando puede ser controlado por el parámetro de penalización C, que determina la relación entre el error de entrenamiento y la dimensión Vapnik-Chervonenkis del módulo.

Definición 1 (Dimensión Vapnik-Chervonenkis -VC-) La dimensión VC describe la capacidad de un conjunto de funciones implementadas en una máquina de aprendizaje. Para clasificación binaria h es el máximo número de puntos en el que pueden ser separadas dos clases en todas la 2h formas posibles usando las funciones de la máquina de aprendizaje.

Un C grande proporciona un pequeño número de errores de clasificación y un gran w^T w. Es claro que tomando $C = \infty$ requiere que el número de datos mal clasificados sea cero. Sin embargo, en este caso no es posible, ya que el problema puede ser factible únicamente para algún valor $C < \infty$. Introduciendo "variables flojas" no negativas $\xi_i(i = 1, l)$ al problema de optimización, ahora en lugar de la condiciones del hiperplano de separación deberá satisfacer

$$\begin{gathered} \min_{w,b,\xi_i} \langle w \cdot w \rangle + C \sum_{i=1}^{l} \xi_i^2 \\ sujeto\ a: y_i(\langle w \cdot x_i \rangle + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{gathered} \tag{18}$$

i.e., sujeto a

$$\begin{gathered} \langle w \cdot x_i \rangle + b \geq +1 - \xi_i, para\ y_i = +1, \xi_i \geq 0 \\ \langle w \cdot x_i \rangle + b \leq +1 - \xi_i, para\ y_i = +1, \xi_i \geq 0 \\ Si\ \xi_i < 0, y_i(\langle w \cdot x_i \rangle + b) \geq 1 - \xi_i \geq 1, \end{gathered} \tag{19}$$

por lo tanto, no consideramos la condición $\xi_i < 0$.

Para el máximo margen blando con Norma-2 $((1/C)\delta_{ij})$ la Lagrangiana original está dado por:

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

$$L(w,b,\xi_i,\alpha) = \frac{1}{2}\langle w \cdot w\rangle - \sum_{i=1}^{l} \alpha_i [y_i(\langle w \cdot x_i\rangle + b) - 1 + \xi_i] + \frac{C}{2}\sum_{i=1}^{l} \xi_i^2 \quad (20)$$

El dual es encontrado en dos pasos: de la misma manera que en el caso linealmente separable primero diferenciando con respecto a w y b, y después re sustituyendo en el Lagrangiano original, de tal forma que el problema dual sería:

$$\begin{gathered} \underset{\alpha_i}{máx} -\frac{1}{2}\sum_{i,j=1}^{l} \alpha_i y_i \alpha_j y_j \left[\langle x_i \cdot x_j\rangle + \frac{1}{C}\delta_{ij}\right] + \sum_{i=1}^{l} \alpha_i \\ sujeto\ a: \sum_{i=1}^{l} \alpha_i y_i = 0 \end{gathered} \quad (21)$$

3.5 Funciones de decisiones en las SVM

La condición de Kuhn-Tucker es:

$$\alpha_i^*[y_i(\langle w^* \cdot x_i\rangle + b^*) - 1 + \xi_i] = 0 \quad (22)$$

Esto es, el problema de optimización cuadrática es prácticamente el mismo que en el caso separable con la única diferencia de las cotas modificadas de los multiplicadores de Lagrange αi. El parámetro C es determinado por el usuario. La selección de una apropiada C es realizada experimentalmente usando alguna técnica de validación cruzada.

3.6 Kernels

En una SVM, el hiperplano óptimo es determinado para maximizar su habilidad de generalización. Pero, si lo datos de entrenamiento no son linealmente separables, el clasificador obtenido puede no tener una alta habilidad de generalización, aun cuando los hiperplanos sean determinados óptimamente, para maximizar el espacio entre clases, el espacio de entrada original es transformado dentro de un espacio altamente dimensional llamado "espacio de características".

La idea básica en diseño de SVM no lineales es transformar los vectores de entrada $x \in R^n$ dentro de vectores Φ(x) de un espacio de características altamente dimensional F (donde Φ representa el mapeo: $R^n \rightarrow R^f$) y resolver el problema de clasificación lineal en este espacio de características:

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

$$x \in R^n \rightarrow \Phi(\mathrm{x}) = [\emptyset_1(x), \emptyset_2(x), \ldots, \emptyset_n(x)]^T \in R^f \quad (23)$$

El conjunto de hipótesis que consideraremos será funciones de tipo:

$$f(x) = \sum_{i=1}^{l} w_i \emptyset_i(x) + b \quad (24)$$

donde $\varphi : X \rightarrow F$ es un mapeo no lineal desde un espacio de entrada a un espacio de características, el procedimiento de aprendizaje consiste en dos pasos: primero, un mapeo no lineal transforma los datos dentro de un espacio de características F y después, una máquina lineal es utilizada para clasificar los datos en un espacio de características.

Como se vio anteriormente, una propiedad de las máquinas de aprendizaje lineal es que éstas pueden ser expresadas en una representación dual, esto significa que puede ser expresada como una combinación lineal de los puntos de entrenamiento.

Por lo tanto, la regla de decisión puede ser evaluada usando productos punto:

$$f(x) = \sum_{i=1}^{l} \alpha_i y_i \langle \emptyset(x_i) \cdot \emptyset(x) \rangle + b \quad (25)$$

Si se tiene una forma de capturar el producto hφ(xi) · φ(x)i en el espacio de características, directamente como una función de los puntos de entrada originales, esto hace posible unir los dos pasos necesarios para construir una máquina de aprendizaje no-lineal. A este método de cómputo directo se le llama función kernel.

Definición 2 Un kernel es una función K, tal que, para todo x, z ∈ X

$$K(x, z) = \langle \emptyset(x) \cdot \emptyset(z) \rangle \quad (26)$$

donde φ es un mapeo de X a un espacio de características F. La clave del enfoque es encontrar una función kernel que pueda ser evaluada eficientemente. Una vez que tenemos tal función de decisión, la regla puede ser evaluada:

$$f(x) = \sum_{i=1}^{l} \alpha_i y_i K\langle x_i \cdot x \rangle + b \quad (27)$$

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

3.7 Condición de Mercer

El teorema de Mercer provee una caracterización de cuando una función K (x, z) es un kernel. Dado un espacio de entrada finito X = {x1, ..., xn} y suponiendo que K (x, z) es una función simétrica de X, entonces la matriz

$$K = (K(x_i \cdot x_j))_{i,j=1}^{n} \tag{28}$$

Ya que K es simétrica existe una matriz ortogonal V tal que K = VΛV0, donde Λ es la matriz diagonal que contiene los eigenvalores λt de K, con sus correspondientes eigenvectores vt = (vti)i=1. Asumiendo que todos los eigenvalores son no-negativos y considerando el mapeo se tiene que

$$\langle \emptyset(x_i) \cdot \emptyset(x_j) \rangle = \sum_{t=1}^{n} \lambda_t v_{ti} v_{tj} = (\mathrm{V}\Lambda\mathrm{V}')_{ij} = K_{ij} = K(x_i \cdot x_j) \tag{29}$$

implica que K (x, z) es una función kernel correspondiente al mapeo φ. El requerimiento de que los eigenvalores de K sean no negativos es necesario, ya que, si se tiene un eigenvalor negativo λs en el eigenvector vs, el punto en el espacio de características podría tener norma cuadrada.

$$\|z\|^2 = \langle z \cdot z \rangle = v_s' V \sqrt{\Lambda}\sqrt{\Lambda} V' v_s = v_s' V \Lambda V' v_s = v_s' K v_s = \lambda_s < 0, \tag{30}$$

Contradiciendo la geometría de este espacio. Esto nos lleva a la siguiente proposición

Proposición 3: Sea X un espacio de entrada finito con una función simétrica sobre X K (x, z). Se decide que K (x, z) es una función kernel si y solamente si la matriz es positiva semi definida (tiene eigenvalores no negativos).

$$K = (K(x_i \cdot x_j))_{i,j}^{n} = 1 \tag{31}$$

Permitiendo una ligera generalización de un producto punto en un espacio de Hilbert, introduciendo un peso λi para cada dimensión

$$\langle \emptyset(x) \cdot \emptyset(z) \rangle = \sum_{i=1}^{\infty} \lambda_i \emptyset_i(x) \cdot \emptyset_i(z) = K(x, z), \tag{32}$$

por lo tanto, el vector de características sería

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

$$\emptyset: x_i \mapsto \left(\sqrt{\lambda_t}\, v_{ti}\right)_{t=1}^{n} \in R^n, i = 1, \dots, n.$$
$$\emptyset(x) = (\emptyset_i(x), \emptyset_i(x), \dots, \emptyset_i(x), \dots). \tag{33}$$

El teorema de Mercer da las condiciones necesarias y suficientes para que una función simétrica continua K (x, z) sea representada:

$$K(x, z) = \sum_{i=1}^{\infty} \lambda_i \emptyset_i(x) \cdot \emptyset_i(z) \tag{34}$$

con λi no negativos, que es equivalente a que K (x, z) sea un producto punto en el espacio de características F ⊇ φ(X), donde F es el espacio l_2 de todas las secuencias

$$z = \sum_{i=1}^{n} v_{si} \emptyset(x_i) = \sqrt{\Lambda}\, V' v_s \tag{35}$$

Para el cual

$$\psi = (\psi_1, \psi_2, \dots, \psi_i, \dots). \tag{36}$$

Esto implícitamente induce un espacio definido por el vector de características y como una consecuencia una función lineal en F puede ser representada por

$$\sum_{i=1}^{\infty} \lambda_i \psi_i^2 < \infty.$$
$$f(x) = \sum_{i=1}^{\infty} \lambda_i \psi_i \emptyset_i(x) + b = \sum_{j=1}^{1} \alpha_j y_j K\left(x, x_j\right) + b \tag{37}$$

Donde la primera expresión es la representación original y la segunda es el dual. La relación entre los dos está dada por:

$$\psi = \sum_{j=1}^{l} \alpha_j y_j\, \emptyset\left(x_j\right). \tag{38}$$

En la representación original, el número de términos en la suma es igual a la dimensión de calidad en el espacio de características, mientras que en el dual existen l

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

términos. La analogía con el caso finito es similar. La contribución a partir del análisis funcional conduce al problema para ecuaciones integrales de la forma:

$$\int_x K(x,z)\emptyset(z)\,dz = \lambda\emptyset(x) \tag{39}$$

donde K (x, z) es una función kernel acotada, simétrica y positiva y X es un espacio compacto.

Teorema 2 (Mercer) Sea X un subconjunto compacto de Rn. Suponiendo que K es una función simétrica continua tal que el operador integral TK: L2(X) → L2(x),

$$(T_K f)(\cdot) = \int_x K(\cdot,x)f(x)dx, \tag{40}$$

es positivo, esto es

$$\int_{X\,\mathbf{x}\,X} K(x,z)\,f(x)f(z)dxdz \geq 0, \tag{41}$$

para toda f ∈ L2(X). Entonces K (x, z) puede ser expandida en una serie uniformemente convergente (sobre X × X) en términos de las eigenfunciones φj ∈ L2(X), normalizadas

$$K(x,z) = \sum_{j=1}^{\infty} \lambda_i \emptyset_j(x)\emptyset_j(z). \tag{42}$$

3.7.1 Caso linealmente no separable

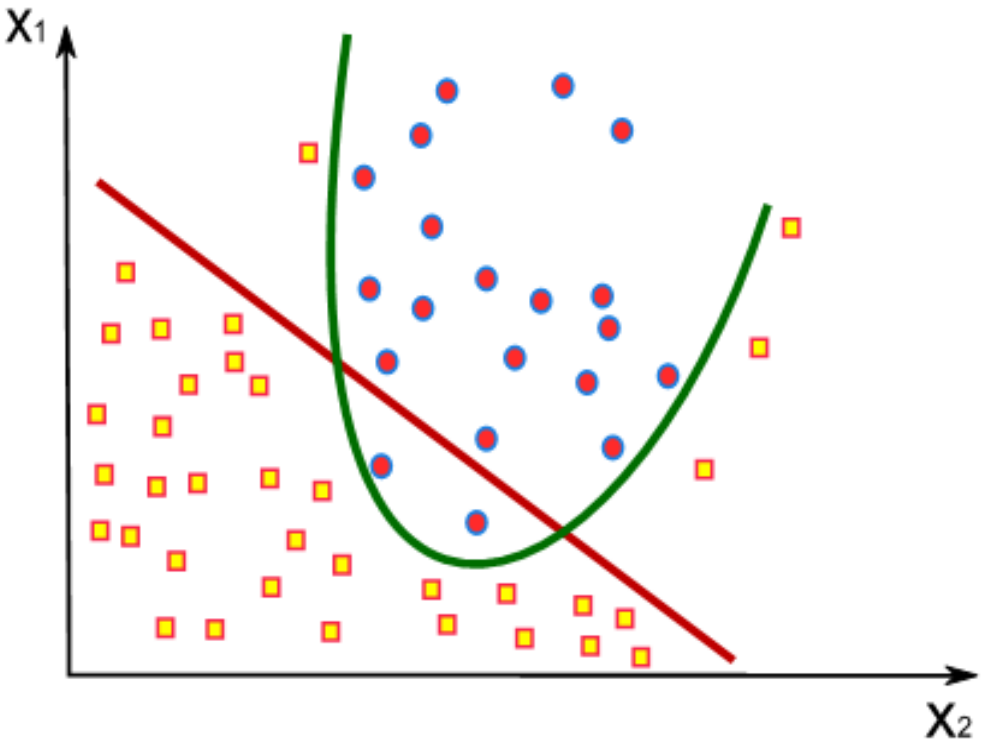
Los clasificadores lineales presentados en las dos secciones anteriores son muy limitados. En la mayoría de las clases, no únicamente se traslapan o interceptan los datos al generar un hiperplano de separación, sino que la separación genuina de estos datos está dada por hipersuperficies no lineales. Una característica del enfoque presentado anteriormente radica en que éste, puede ser fácilmente extendido para crear cotas de decisión no lineal. El motivo de tal extensión es que una SVM puede crear una hipersuperficie de decisión no lineal, capaz de clasificar datos separables no linealmente. Generalmente, para patrones de entrada n-dimensionales, en lugar de una curva no lineal, una SVM creará una hipersuperficie de separación no lineal.

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

El problema de optimización utilizando kernels queda de la siguiente manera.

Figura 11

Clasificador No-Lineal



Nota: elaboración propia.

Proposición 4: Dado un conjunto de datos de entrenamiento S = [(x1, y1) ⋯ (xl, yl)], un espacio de características φ (x) definido por el kernel K (x, z) = hφ (x) · φ (z)i, la solución de

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

$$\max_{\alpha_i} -\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i y_i \alpha_j y_j\left[K(x_i,x_j)+\frac{1}{C}\delta_{ij}\right]+\sum_{i=1}^{l}\alpha_i$$

$$sujeto\ a: \sum_{i=1}^{l}\alpha_i y_i = 0$$

$$es\ \alpha_i^*, f(x) = \sum_{i=1}^{l}\alpha_i^* y_i K(x_i,x) + b^*, donde\ b^*\ es\ elegido\ tal\ que \tag{43}$$

$$y_i f(x_i) = 1-\xi^* = 1-\frac{\alpha^*}{C}$$

$$w^* = \sum_{i=1}^{l}\alpha_i^* y_i K(x,x),$$

La regla de decisión sgn [f (x)] es equivalente al hiperplano en el espacio de características definido por el kernel K (x, z) el cual resuelve el problema de optimización. Luego, el margen geométrico está dado por:

$$\gamma^* = \left(\sum_{i\in\delta v}\alpha_i^* - \frac{1}{C}\langle\alpha^*\cdot\alpha^*\rangle\right)^{-\frac{1}{2}} \tag{44}$$

Utilizando el Kernel

El margen blando en L1

$$\begin{matrix}\min\limits_{w,b,\xi_i}\langle w\cdot w\rangle + C\sum_{i=1}^{l}\xi_i \\ sujeto\ a:\ y_i(\langle w\cdot x_i\rangle + b)\geq 1-\xi_i \\ \xi_i\geq 0\end{matrix} \tag{45}$$

Donde el Lagrangiano original es:

$$K'(x,z) = K(x,z) + \frac{1}{C}\delta_x(z)\gamma^* = \left(\sum_{i\in\delta v}\alpha_i^* - \frac{1}{C}\langle\alpha^*\cdot\alpha^*\rangle\right)^{-\frac{1}{2}} \tag{46}$$

De donde se tiene que el dual está dado por:

$$L(w,b,\xi_i,\alpha) = \frac{1}{2}\langle w\cdot w\rangle - \sum_{i=1}^{l}\alpha_i[y_i(\langle w\cdot x_i\rangle + b) - 1 + \xi_i] + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\gamma_i\xi_i$$

$$w(\alpha) = -\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i y_i\,\alpha_j y_j\,\langle x_i\cdot x_j\rangle + \sum_{i=1}^{l}\alpha_i \tag{47}$$

Este es el mismo que el margen máximo, pero

$$C-\alpha_i-\gamma_i = 0, \gamma_i\geq 0 \Rightarrow C\geq\alpha_i \tag{48}$$

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896

Con las condiciones de Kuhn-Tucker

$$\gamma_i \xi_i = 0 \text{ ó } (\alpha_i - C)\xi_i = 0$$
$$\alpha_i [y_i (\langle w \cdot x_i \rangle + b) - 1 + \xi_i = 0 \quad (49)$$

Donde ξi = 0, γi = 0, ⇒ C = αi, con ξi = 0 el margen es máximo, αi es positivo y puede incrementarse hasta C, por lo tanto, C ≥ αi ≥ 0

Proposición 5: Dado un conjunto de datos de entrenamiento S = [(x1, y1) (xl, yl)], un espacio de características φ (x) definido por el kernel K (x, z) = hφ (x) φ (z)i, la solución de:

$$\max_{\alpha_i} -\frac{1}{2}\sum_{i,j=1}^{l} \alpha_i y_i \alpha_j y_j K(x_i, x_j) + \sum_{i=1}^{l} \alpha_i$$
$$sujeto\ a: \sum_{i=1}^{l} \alpha_i y_i = 0, C \geq \alpha_i \geq 0 \quad (50)$$

hiperplano en el espacio de características definido por el Kernel K (x, z), el cual resuelve el problema de optimización. El margen geométrico está dado por:

$$\gamma^* = \left(\sum_{i \in \delta v} \alpha_i^* \right)^{-\frac{1}{2}} \quad (51)$$

Cuando la cota de αi es C, se origina el problema del máximo margen. Elegir C es lo mismo que obtener v en:

$$\max_{\alpha_i} -\frac{1}{2}\sum_{i,j=1}^{l} \alpha_i y_i \alpha_j y_j K(x_i, x_j)$$
$$sujeto\ a: \sum_{i=1}^{l} \alpha_i y_i = 0,$$
$$\sum_{i=1}^{l} \alpha_i \geq v, \quad (52)$$
$$\frac{1}{l} \geq \alpha_i \geq 0$$

ISBN: 978-628-95636-8-9 - DOI:https://doi.org/10.5281/zenodo.8298896