Máquinas de Vectores de Soporte

Luis Norberto Zúñiga Morales

7 de enero de 2024

Índice

1. Motivación **2**

2. Máquina de Vectores de Soporte: Caso Lineal **3**
2.1. Teoría 3
2.2. Implementación Práctica 7

3. Máquina de Vectores de Soporte: Margen Suave **8**
3.1. Teoría 8
3.2. Implementación Práctica 9

4. Máquinas de Vectores de Soporte: Caso No Lineal **10**
4.1. Teoría 10
4.2. Implementación Práctica 12

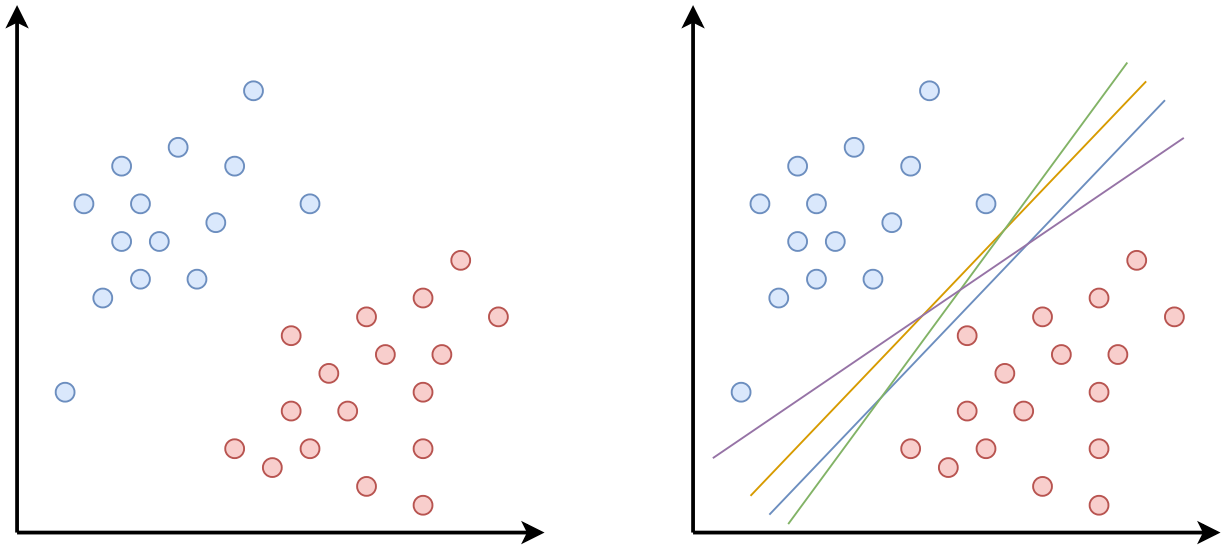
5. Máquinas de Vectores de Soporte para el Caso Multiclase **13**

6. Ejercicios **14**

7. Registro de Actualizaciones **14**

Resumen

La Máquina de Vectores de Soporte (MVS) es un poderoso modelo de aprendizaje automático que es utilizado en aplicaciones que cubren un amplio espectro de áreas de estudio, tales como medicina, lingüistica computacional, cómputo financiero, psicología, entre muchos otras. La MVS [4] es un algoritmo de clasificación binaria (i.e., separa objetos de dos clases distintas) que puede adaptarse a un problema de clasificación multiclase (más de dos clases) y que, gracias al truco del kernel, es posible salir de la idea inicial de separar clases por medio de funciones lineales, y utilizar funciones no lineales para realizar separaciones con formas más complejas. Esto vuelve a las MVS en un modelo de clasificación muy flexible.



(a) Puntos de dos clases distintas en $\mathbb{R}^2$.

(b) Posibles planos para la separación.

Figura 1: Ejemplo sencillo del problema de separación de clases. ¿Cuántos hiperplanos se pueden construir? ¿Cuál de ellos se debe elegir?

1. Motivación

La idea básica que surge al considerar el problema de separar elementos de dos clases distintas de datos, los cuales se pueden representar como vectores en $\mathbb{R}^n$, es la forma de realizar tal separación. Lo más sencillo es la construcción de un hiperplano que separe los puntos de tal manera que de un lado se encuentren aquellos de las clase A y del otro aquellos de la clase B[1]. Para ejemplificar el problema anterior, consideremos el problema de clasificación de la Fig. (1a) que muestra un caso a modo para nosotros, cosa que rara vez se observa en práctica. Al momento de separar las clases, ¿de cuántas maneras es posible construir un hiperplano que separe puntos de ambas clases? La Fig. (1b) muestra algunos hiperplanos propuestos que resuelven perfectamente el problema planteado. ¿Cuál de ellos se debe elegir? ¿Existe alguno que se pueda denominar «mejor»?

Después de esta breve discusión, es natural preguntarse si existe una manera ordenada que permita construir un plano único que separe, de alguna manera, óptimamente los puntos que representan a cada miembro de una clase en particular. Por supuesto, faltaría definir con qué se relaciona ese óptimo y condiciones adicionales para encontrarlo. Las Máquinas de Vectores de Soporte abordan esta pregunta planteando un problema que busca al hiperplano que maximiza la distancia entre este y ciertos puntos del conjunto de datos de una manera elegante y relativamente sencilla.

[1]Para este modelo, hasta que se diga lo contrario, se considerará el caso de clasificación binaria, el cual permitirá formular el problema básico a atacar, y que posteriormente se expandirá a problemas de clasificación multiclase.

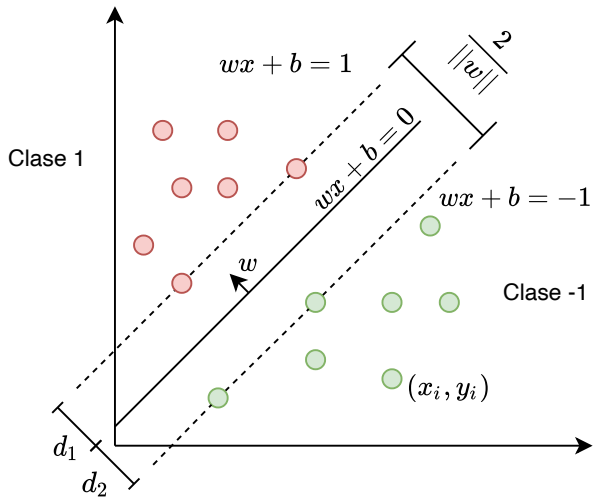


Figura 2: Diagrama que muestra la idea básica de la Máquina de Vectores de Soporte.

2. Máquina de Vectores de Soporte: Caso Lineal

2.1. Teoría

Para formular la idea básica de la MVS es útil comenzar con el caso más sencillo, el cual se presenta al considerar funciones lineales para construir la función que separa los puntos de ambas clases de manera perfecta. A lo anterior se le conoce como un problema de **clasificación binaria linealmente separable**.

Supongamos que se tienen L puntos de entrenamiento $\{\boldsymbol{x}_i, y_i\}$, donde $\boldsymbol{x}_i \in \mathbb{R}^D$ representa un vector de entrada de dimensión D (el vector de atributos o características) el cual se asocia a una de dos clases posibles: $y_i = +1$ o $y_i = -1$. Al asumir un problema de clasificación separable linealmente, la función que separa los puntos de ambas clases es una recta cuando $D = 2$, y un hiperplano cuando $D > 2$. Para no perder generalidad, dicho hiperplano se puede definir de la siguiente manera:

$$\boldsymbol{w}^T \cdot \boldsymbol{x} + b = 0 \tag{1}$$

donde:

- $\boldsymbol{w}$ es un vector normal al hiperplano
- $\frac{b}{||\boldsymbol{w}||}$ es la distancia perpendicular del hiperplano al origen

Para determinar el hiperplano separador, es posible utilizar a los puntos $\boldsymbol{x}_i$ que se encuentren más cercanos a este, de tal forma que la separación se encuentre lo más lejos posible de los puntos más cercanos de ambas clases. La Figura 2 muestra la idea anterior, donde los hiperplanos formados por los vectores de soporte (línea punteada) forman un margen donde se ubicará el hiperplano separador (linea sólida), idealmente en medio de este.

Para encontrar el hiperplano separador mediante la MVS, se necesita encontrar aquellos $\boldsymbol{w}$ y b que forman los hiperplanos de apoyo de tal forma que

$$\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b \geq +1 \quad \text{para } y_i = +1 \tag{2}$$
$$\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b \leq -1 \quad \text{para } y_i = -1 \tag{3}$$

Las Ecuaciones 2 y 3 se pueden combinar convenientemente en una sola expresión:

$$y_i(\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b) \geq 1 \qquad \forall i \tag{4}$$

mientras que las ecuaciones de los hiperplanos H_1 y H_2 de soporte (uno para cada clase) se encuentran dadas por:

$$\boldsymbol{w}^T \cdot \boldsymbol{x} + b = +1 \qquad \text{para } H_1 \tag{5}$$
$$\boldsymbol{w}^T \cdot \boldsymbol{x} + b = -1 \qquad \text{para } H_2 \tag{6}$$

Una forma de interpretar la Ecuación 4 es que evita que los puntos del conjunto de datos queden dentro del margen de separación y, a lo más, formen parte de la recta o hiperplano de los márgenes de apoyo, que es el caso de las Ecuaciones 5 y 6.

En la Figura 2 se muestra que el hiperplano separador se encuentra en medio del espacio formado por H_1 y H_2, el cual se llama **margen de la MVS**, donde d_1 y d_2 denotan la distancia de H_1 y H_2 al hiperplano separador, respectivamente. Este margen debe ser lo más amplio posible, por lo que se presenta un problema de maximización. Sin embargo, ¿cuánto mide el margen de la MVS?

Sea $\boldsymbol{x}_j$ un punto en el hiperplano $\boldsymbol{w}^T \cdot \boldsymbol{x} + b = -1$. Para medir la distancia entre los hiperplanos $\boldsymbol{w}^T \cdot \boldsymbol{x} + b = +1$ y $\boldsymbol{w}^T \cdot \boldsymbol{x} + b = -1$ es necesario calcular la longitud del segmento perpendicular desde $\boldsymbol{x}_j$ hasta el plano $\boldsymbol{w}^T \cdot \boldsymbol{x} + b = +1$. Denotemos como z dicha cantidad. La clave aquí es denotar z en términos de $\boldsymbol{w}$, lo cual se logra al considerar el vector unitario $\frac{\boldsymbol{w}}{||\boldsymbol{w}||}$ del hiperplano $\boldsymbol{w}^T \cdot \boldsymbol{x} + b = +1$, y multiplicarlo por z. Por lo tanto, se tiene que

$$\boldsymbol{w}^T(\boldsymbol{x}_j + z\frac{\boldsymbol{w}}{||\boldsymbol{w}||}) + b = 1 \tag{7}$$

ya que $\boldsymbol{x}_j + z\frac{\boldsymbol{w}}{||\boldsymbol{w}||}$ es un punto que yace en $\boldsymbol{w}^T \cdot \boldsymbol{x} + b = +1$ En otras palabras, si al punto $\boldsymbol{x}_j$, que por definición se encuentra en el hiperplano $\boldsymbol{w}^T \cdot \boldsymbol{x} + b = -1$, le sumamos la distancia z, terminamos con un punto en el hiperplano $\boldsymbol{w}^T \cdot \boldsymbol{x} + b = +1$. Expandiendo la Ecuación 7:

$$
\begin{aligned}
\boldsymbol{w}^T(\boldsymbol{x}_j + z\frac{\boldsymbol{w}}{||\boldsymbol{w}||}) + b &= 1 \\
\boldsymbol{w}^T \cdot \boldsymbol{x}_j + z\frac{\boldsymbol{w}^T\boldsymbol{w}}{||\boldsymbol{w}||} + b &= 1 \\
\boldsymbol{w}^T \cdot \boldsymbol{x}_j + z\frac{||\boldsymbol{w}||^2}{||\boldsymbol{w}||} + b &= 1 \\
\boldsymbol{w}^T \cdot \boldsymbol{x}_j + z||\boldsymbol{w}|| + b &= 1 \\
\boldsymbol{w}^T \cdot \boldsymbol{x}_j + b &= 1 - z||\boldsymbol{w}|| \\
-1 &= 1 - z||\boldsymbol{w}|| \\
2 &= z||\boldsymbol{w}|| \\
z &= \frac{2}{||\boldsymbol{w}||}
\end{aligned}
\tag{8}
$$

Para maximizar $\frac{2}{||\boldsymbol{w}||}$, la distancia entre los hiperplanos H_1 y H_2, se debe minimizar[2] $||\boldsymbol{w}||$. Por lo tanto, el problema de optimización resultante es el siguiente:

$$
\begin{aligned}
\text{mín} \quad & ||\boldsymbol{w}|| \\
\text{sujeto a} \quad & y_i(\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b) \geq 1
\end{aligned}
\tag{9}
$$

Al minimizar $||\boldsymbol{w}||$, notemos que esta no es diferenciable en 0, por lo que en su lugar[3] se minimiza $\frac{1}{2}||\boldsymbol{w}||^2$. Oportunamente, con esta forma es posible solucionar el problema de optimización cuadrática más adelante. En consecuencia, es necesario encontrar la solución a:

$$
\begin{aligned}
\text{mín} \quad & \frac{1}{2}||\boldsymbol{w}||^2 \\
\text{sujeto a} \quad & y_i(\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b) \geq 1
\end{aligned}
\tag{10}
$$

El siguiente paso es resolver el problema de minimización, por lo que vamos a utilizar multiplicadores de Lagrange $\boldsymbol{\alpha}$, donde $\alpha_i \geq 0 \ \forall i$.

$$
\mathbb{L}_P = \frac{1}{2}||\boldsymbol{w}||^2 - \boldsymbol{\alpha}[y_i(\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b) - 1] \qquad \forall i \tag{11}
$$

$$
= \frac{1}{2}||\boldsymbol{w}||^2 - \sum_{i=1}^{L} \alpha_i[y_i(\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b) - 1] \tag{12}
$$

$$
= \frac{1}{2}||\boldsymbol{w}||^2 - \sum_{i=1}^{L} \alpha_i y_i(\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b) + \sum_{i=1}^{L} \alpha_i \tag{13}
$$

Como se mencionó anteriormente, se debe encontrar $\boldsymbol{w}$ y b tales que minimicen la Ecuación 10 y los $\alpha_i \geq 0$ que maximicen la Ecuación 13. Para lograrlo, diferenciemos $\mathbb{L}_P$ con respecto de $\boldsymbol{w}$ y b, igualando las derivadas parciales a cero:[4]

[2]Entre más pequeño sea el denominador en una fracción, mayor es el resultado final.

[3]Los algoritmos de optimización suelen funcionar mejor cuando se aplican en funciones diferenciables en cualquier punto.

[4]Es decir, el clásico problema de optimizar mediante derivadas.

$$\frac{\partial \mathbb{L}_P}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{L} \alpha_i y_i \boldsymbol{x}_i = 0 \Rightarrow \boldsymbol{w} = \sum_{i=1}^{L} \alpha_i y_i \boldsymbol{x}_i \tag{14}$$

$$\frac{\partial \mathbb{L}_P}{\partial b} = \sum_{i=1}^{L} \alpha_i y_i = 0 \tag{15}$$

Al sustituir las Ecuaciones 14 y 15 en la Ecuación 13, se obtiene una nueva expresión que depende de $\boldsymbol{\alpha}$, por lo que hay que maximizar:

$$\mathbb{L}_D = \sum_{i=1}^{L} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j, \quad \alpha_i \geq 0, \quad \sum_{i=1}^{L} \alpha_i y_i = 0 \tag{16}$$

$$= \sum_{i=1}^{L} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i H_{ij} \alpha_j, \quad H_{ij} = y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j \tag{17}$$

$$= \sum_{i=1}^{L} \alpha_i - \frac{1}{2} \boldsymbol{\alpha}^T \boldsymbol{H} \boldsymbol{\alpha}, \quad \alpha_i \geq 0, \quad \sum_{i=1}^{L} \alpha_i y_i = 0 \tag{18}$$

La Ecuación 18 se le conoce como el **dual** del **primal** $\mathbb{L}_P$[5] En este caso, el dual tiene la característica de requerir únicamente el producto punto de los vectores de entrada $\boldsymbol{x}_i$, el cual será útil más adelante cuando se introduzca el truco del kernel.

Durante este proceso, se partió de minimizar $\mathbb{L}_P$ a maximizar $\mathbb{L}_D$. Es decir, el problema de optimización luce de la siguiente forma:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{L} \alpha_i - \frac{1}{2} \boldsymbol{\alpha}^T \boldsymbol{H} \boldsymbol{\alpha} \\ \text{sujeto a} \quad & \alpha_i \geq 0 \\ & \sum_{i=1}^{L} \alpha_i y_i = 0 \end{aligned} \tag{19}$$

Al problema descrito en la expresión anterior se le conoce como un problema de **optimización cuadrático convexo**, el cual se puede resolver por medio de algoritmos implementados en diversos paquetes de software. En resumen, al resolver el problema de optimización cuadrático se obtendrá $\boldsymbol{\alpha}$, y de la Ecuación 14 se puede obtener $\boldsymbol{w}$. Solo faltaría calcular b.

¿Recuerdan que durante al planteamiento de la idea se mencionó que los margenes se construyen mediante vectores de apoyo del conjunto de datos? Cualquier punto x_s que funcione como vector de soporte satisface la ecuación

$$y_s(\boldsymbol{x}_s \cdot \boldsymbol{w} + b) = 1 \tag{20}$$

Sustituyendo en la Ecuación 14:

$$y_s(\sum_{k \in S} \alpha_k y_k \boldsymbol{x}_k \cdot \boldsymbol{x}_s + b) = 1 \tag{21}$$

[5]De ahí los subíndices P y D en $\mathbb{L}$.

donde S denota el conjunto de índices de los vectores de soporte, el cual se construye al encontrar los i tales que $\alpha_i > 0$. Multiplicando por y_s, observando que $y_s^2 = 1$ de la Ecuación 2 y 3:

$$\begin{aligned} & y_s^2(\sum_{k\in S} \alpha_k y_k \boldsymbol{x}_k \cdot \boldsymbol{x}_s + b) = y_s \\ \Rightarrow \quad & b = y_s - \sum_{k\in S} \alpha_k y_k \boldsymbol{x}_k \cdot \boldsymbol{x}_s \end{aligned} \tag{22}$$

Ahora, surge la pregunta sobre cuál o cuáles vectores de soporte $\boldsymbol{x}_s$ utilizar para calcular b. En la práctica, se utiliza un promedio de todos los vectores en S:

$$b = \frac{1}{N_S} \sum_{s\in S} (y_s - \sum_{k\in S} \alpha_k y_k \boldsymbol{x}_k \cdot \boldsymbol{x}_s) \tag{23}$$

Finalmente se tiene $\boldsymbol{w}$ y b para construir el hiperplano óptimo que separe los datos y, en consecuencia, la Máquina de Vectores de Soporte.

2.2. Implementación Práctica

1. Crear $\boldsymbol{H}$, donde $H_{ij} = y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j$.

2. Encontrar las α_i que maximicen

$$\sum_{i=1}^{L} \alpha_i - \frac{1}{2} \boldsymbol{\alpha}^T \boldsymbol{H} \boldsymbol{\alpha}$$

   sujeto a las restricciones

$$\alpha_i \geq 0, \quad \sum_{i=1}^{L} \alpha_i y_i = 0.$$

   En práctica, se utilizan librerías en distintos lenguajes de programación que resuelvan el problema de optimización cuadrática, como qpsolvers en Python.

3. Calcular $\boldsymbol{w} = \sum_{i=1}^{L} \alpha_i y_i \boldsymbol{x}_i$.

4. Determinar el conjunto de vectores de soporte S mediante la identificación de los índices i tales que $\alpha_i > 0$.

5. Calcular el valor de b mediante la ecuación

$$b = \frac{1}{N_S} \sum_{s\in S} (y_s - \sum_{k\in S} \alpha_k y_k \boldsymbol{x}_k \cdot \boldsymbol{x}_s).$$

6. Cada elemento del conjunto de prueba $\boldsymbol{x}_t$ se clasifica evaluando

$$y_t = sgn(\boldsymbol{w}^T \cdot \boldsymbol{x}_t + b).$$

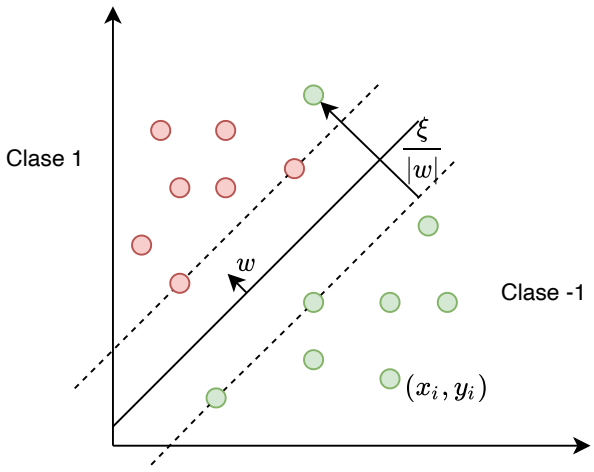


Figura 3: Idea de la MVS con penalización.

3. Máquina de Vectores de Soporte: Margen Suave

El caso de clasificación de datos linealmente separable permite crear los fundamentos de la MVS. La idea discutida en la sección anterior introduce el concepto de **margen duro**, donde los datos deben ser clasificados perfectamente. Esto conlleva a que, en ciertos casos, los márgenes que se generan sean muy estrechos, lo cual puede llevar al sobreajuste. Con este problema en mente, se introduce el concepto del **margen suave** en la MVS que produce un modelo más flexible capaz de generalizar mejor los datos.

3.1. Teoría

Para resolver el problema que genera el margen duro, la idea es relajar las limitantes dadas en las Ecuaciones 2 y 3 para permitir puntos mal clasificados y, con este sacrificio, permitir que el modelo aprenda mejor de los datos. Lo anterior se logra introduciendo un variable de holgura positiva ξ_i, $i = 1, \ldots, L$:

$$\boldsymbol{x}_i \cdot \boldsymbol{w} + b \geq +1 - \xi_i \qquad y_i = +1 \tag{24}$$

$$\boldsymbol{x}_i \cdot \boldsymbol{w} + b \leq -1 + \xi_i \qquad y_i = -1 \tag{25}$$

$$\xi_i \geq 0 \qquad \forall i \tag{26}$$

Similar a la Ecuación 4, las Ecuaciones 24, 25 y 26 se pueden combinar en una sola:

$$y_i(\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i \quad \text{donde} \quad \xi_i \geq 0 \ \forall i \tag{27}$$

El margen suave penaliza los puntos mal clasificados (del lado equivocado) al agregar cierto pérdida en la función a optimizar. Dicha penalización incrementa conforme aumenta la distancia entre el punto mal clasificado y su margen de decisión correcto. De esta forma, se busca encontrar un balance entre el número de clasificaciones incorrectas y el tamaño del

margen de separación. Lo anterior se adapta en la Ecuación 10 de la siguiente manera:

$$\begin{aligned} \text{mín} \quad & \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{L}\xi_i \\ \text{sujeto a} \quad & y_i(\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i \quad \forall i \end{aligned} \tag{28}$$

donde el parámetro C controla la razón de intercambio entre la penalización y el tamaño del margen[6].

De manera similar a la MVS Lineal, es momento de calcular el Lagrangiano, el cual debe ser minimizado con respecto a $\boldsymbol{w}$, b y ahora ξ_i; y maximizado con respecto a $\boldsymbol{\alpha}$ y $\boldsymbol{\mu}$:

$$\mathbb{L}_P = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{L}\xi_i - \sum_{i=1}^{L}\alpha_i[y_i(\boldsymbol{w}^T \cdot \boldsymbol{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^{L}\mu_i\xi_i \tag{29}$$

Diferenciando con respecto a $\boldsymbol{w}$, b y ξ_i e igualando a cero:

$$\frac{\partial \mathbb{L}_P}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{L}\alpha_i y_i \boldsymbol{x}_i = 0 \Rightarrow \boldsymbol{w} = \sum_{i=1}^{L}\alpha_i y_i \boldsymbol{x}_i \tag{30}$$

$$\frac{\partial \mathbb{L}_P}{\partial b} = \sum_{i=1}^{L}\alpha_i y_i = 0 \tag{31}$$

$$\frac{\partial \mathbb{L}_P}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \Rightarrow C = \alpha_i + \mu_i \tag{32}$$

Sustituyendo en la Ecuación 29, $\mathbb{L}_D$ tiene la misma forma que la ecuación (16) en el caso lineal. Sin embargo, al considerar la Ecuación 32 y el hecho que $\mu_i \geq 0 \ \forall i$[7], se sigue que $\alpha > C$. Por lo tanto, se debe encontrar:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{L}\alpha_i - \frac{1}{2}\boldsymbol{\alpha}^T \boldsymbol{H}\boldsymbol{\alpha} \\ \text{sujeto a} \quad & 0 \leq \alpha_i \leq C \quad \forall_i \\ & \sum_{i=1}^{L}\alpha_i y_i = 0 \end{aligned} \tag{33}$$

Para calcular b, se sigue la misma idea que en el caso lineal dado que la Ecuación 30 y 14 son la misma. En este caso, los vectores de soporte son aquellos cuyos índices i cumplan la desigualdad $0 \leq \alpha_i \leq C$.

3.2. Implementación Práctica

1. Crear $\boldsymbol{H}$, donde $H_{ij} = y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j$.

[6] En la Ecuación 28, el término ξ_i es en realidad una función de pérdida, similar a la idea de regularización que vimos en los modelo de regresión lineal o regresión logística.

[7] Ya que las μ_i son multiplicadores de Lagrange que se añaden al considerar las variables de holgura.

2. Elegir el valor del parámetro C, el cual permitirá penalizar clasificaciones erróneas.

3. Encontrar las α_i que maximicen

$$\sum_{i=1}^{L} \alpha_i - \frac{1}{2}\boldsymbol{\alpha}^T \boldsymbol{H}\boldsymbol{\alpha}$$

sujeto a las restricciones

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{L} \alpha_i y_i = 0.$$

mediante un programa para resolver problemas de optimización cuadrática.

4. Calcular $\boldsymbol{w} = \sum_{i=1}^{L} \alpha_i y_i \boldsymbol{x}_i$.

5. Determinar el conjunto de vectores de soporte S mediante la identificación de los índices i tales que $0 \leq \alpha_i \leq C$.

6. Calcular el valor de b mediante la ecuación

$$b = \frac{1}{N_S} \sum_{s \in S} (y_s - \sum_{k \in S} \alpha_k y_k \boldsymbol{x}_k \cdot \boldsymbol{x}_s).$$

7. Cada elemento del conjunto de prueba $\boldsymbol{x}_t$ se clasifica evaluando

$$y_t = sgn(\boldsymbol{w}^T \cdot \boldsymbol{x}_t + b).$$

4. Máquinas de Vectores de Soporte: Caso No Lineal

El caso de separación lineal que vimos en la primer parte permitió sentar las bases para el modelo de la MVS. Desafortunadamente, en la práctica, esta clase de problemas rara vez se ve en algún conjunto de datos: el caso más común es aquel donde los datos no pueden ser separados fácilmente con un hiperplano y requieren funciones con formas más complejas, las cuales suelen ser no lineales. A pesar de que el margen suave puede ayudar a resolver este tipo de problemas, es posible utilizar un truco que facilita el proceso de clasificación con este tipo de datos.

4.1. Teoría

Durante la construcción del clasificador, en la Ecuación 17 se introdujo la matriz

$$H_{ij} = y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j \tag{34}$$

El producto punto de los vectores de entradas en la matriz anterior se puede representar como una función:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i \cdot \boldsymbol{x}_j = \boldsymbol{x}_i^T \boldsymbol{x}_j \tag{35}$$

La función $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ es un ejemplo de una familia de funciones llamadas **kernels**, las cuales se basan en el calculo de productos puntos de los vectores de entrada del conjunto de datos. La finalidad de los kernels es que son funciones $x \mapsto \phi(\boldsymbol{x})$ que permiten mapear los datos a diferentes dimensiones sin la necesidad de determinar la función ϕ que realiza el mapeo, ya que esta se encuentra dada por nosotros, y únicamente se necesita determinar el producto punto de los vectores.

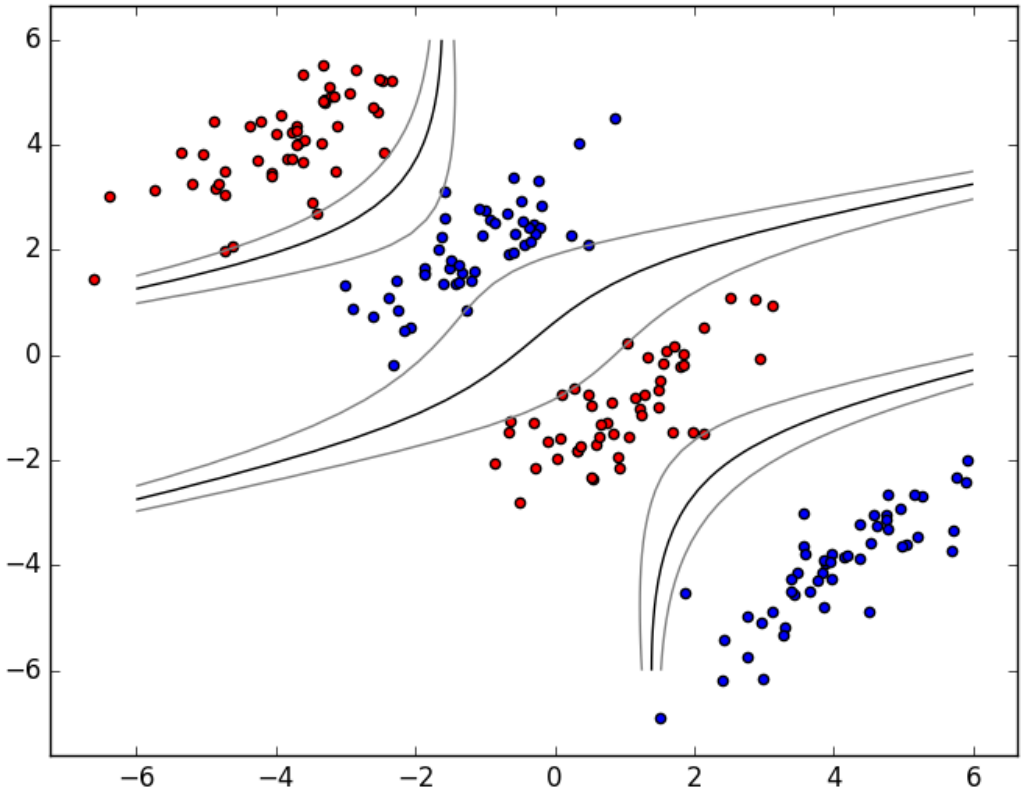


Figura 4: Ejemplo de un conjunto de datos que no es linealmente separable y una MVS que se ajusta para separar los puntos de una forma no lineal.

El **truco del kernel** permite atacar problemas que no son linealmente separables en el espacio en turno, y al realizar el mapeo por medio del kernel, es posible que en otro espacio sí sea separable. La Fig. (4) muestra un ejemplo de un problema de separación donde los datos de ambas clases no pueden separarse por medio de una MVS lineal, por lo que se recurre a una MVS con el truco del kernel.

Los kernels más comunes en la práctica son:

- Kernel Lineal: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i \cdot \boldsymbol{x}_j$
- Kernel Polinomial: $[\gamma(\boldsymbol{x}_i \cdot \boldsymbol{x}_j) + r]^d$

- Kernel Función de Base Radial: $\exp\left(-\gamma \cdot |\boldsymbol{x}_i - \boldsymbol{x}_j|^2\right)$
- Kernel Sigmoide: $\tanh\left(\boldsymbol{x}_i \cdot \boldsymbol{x}_j + r\right)$

donde $\gamma > 0$ y $r, d \in \mathbb{R}$.

Expresado en la formulación del clasificador en la Ecuación 28:

$$\begin{aligned} \text{mín} \quad & \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{L}\xi_i \\ \text{sujeto a} \quad & y_i(\boldsymbol{w}^T \cdot \phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i \quad \forall i \end{aligned} \tag{36}$$

y su dual en la Ecuación 33

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{L}\alpha_i - \frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{H}_k\boldsymbol{\alpha} \\ \text{sujeto a} \quad & 0 \leq \alpha_i \leq C \ \ \forall_i \\ & \sum_{i=1}^{L}\alpha_i y_i = 0 \end{aligned} \tag{37}$$

donde

$$\boldsymbol{H}_k = y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{38}$$

4.2. Implementación Práctica

1. Elegir de antemano cual es el kernel que se aplicará en la MVS y la función de mapeo $\phi(\boldsymbol{x})$. En práctica, el kernel de función de base radial funciona mejor.
2. Crear $\boldsymbol{H}_k$, donde $H_{ij} = y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.
3. Elegir el valor del parámetro C, el cual permitirá penalizar clasificaciones erróneas.
4. Encontrar las α_i que maximicen

   $$\sum_{i=1}^{L}\alpha_i - \frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{H}_k\boldsymbol{\alpha}$$

   sujeto a las restricciones

   $$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{L}\alpha_i y_i = 0.$$

   mediante un programa para resolver problemas de optimización cuadrática.
5. Calcular $\boldsymbol{w} = \sum_{i=1}^{L}\alpha_i y_i \phi(\boldsymbol{x}_i)$.
6. Determinar el conjunto de vectores de soporte S mediante la identificación de los índices i tales que $0 \leq \alpha_i \leq C$.

7. Calcular el valor de b mediante la ecuación

$$b = \frac{1}{N_S} \sum_{s \in S} (y_s - \sum_{k \in S} \alpha_k y_k k(\boldsymbol{x}_k, \boldsymbol{x}_s)).$$

8. Cada elemento del conjunto de prueba $\boldsymbol{x}_t$ se clasifica evaluando

$$y_t = sgn(\boldsymbol{w}^T \cdot \phi(\boldsymbol{x}_t) + b).$$

5. Máquinas de Vectores de Soporte para el Caso Multiclase

Originalmente las máquinas de vectores de soporte se plantearon considerando un modelo de clasificación binaria, es decir, solo separan dos clases. Múltiples enfoques se han considerado para extender el algoritmo para el caso de clasificación multiclase. Se consideran dos formas de atacar el problema: uno contra todos y uno contra uno.

El enfoque **uno contra todos** [2] construye n modelos de MVS, uno para cada clase considerada. El i-ésimo modelo considera la j-ésima clase y sus elementos como la clase positiva y las $n-1$ clases restantes se consideran como las negativas. Dado el conjunto de datos $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_D, y_D)$, la j-ésima MVS resuelve el siguiente problema:

$$\begin{aligned} \text{mín} \quad & \frac{1}{2}||\boldsymbol{w}^j||^2 + C \sum_{i=1}^{L} \xi_i^j \\ \text{sujeto a} \quad & (\boldsymbol{w}^j)^T \cdot \phi(\boldsymbol{x}_i) + b^j \geq 1 - \xi_i^j \quad \text{si} \quad y_j = i \\ & (\boldsymbol{w}^j)^T \cdot \phi(\boldsymbol{x}_i) + b^j \leq -1 + \xi_i^j \quad \text{si} \quad y_j \neq i \\ & \xi_i^j \geq 0 \end{aligned} \tag{39}$$

Al resolver la Ecuación 39 se tienen n funciones de decisión:

$$(\boldsymbol{w}^1)^T \phi(\boldsymbol{x}) + b^1$$

$$\vdots$$

$$(\boldsymbol{w}^n)^T \phi(\boldsymbol{x}) + b^n$$

$\boldsymbol{x}_i$ pertenece a la clase con el mayor valor en la función de decisión:

$$y = \underset{j \in \{1,...,n\}}{\text{argmax}} \ (\boldsymbol{w}^j)^T \phi(\boldsymbol{x}_i) + b^j. \tag{40}$$

El segundo método es llamado **uno contra uno**. Dicho enfoque construye $n(n-1)/2$ clasificadores donde cada uno se entrena con información de dos clases. Considerando los datos de la clase j y la clase k se resuelve el siguiente problema:

$$
\begin{aligned}
\text{mín} \quad & \frac{1}{2}||\boldsymbol{w}^{jk}||^2 + C\sum_{i=1}^{L}\xi_i^{jk} \\
\text{sujeto a} \quad & (\boldsymbol{w}^{jk})^T \cdot \phi(\boldsymbol{x}_i) + b^{jk} \geq 1 - \xi_i^{jk} \quad \text{si} \quad y_i = j \\
& (\boldsymbol{w}^{jk})^T \cdot \phi(\boldsymbol{x}_i) + b^{jk} \leq -1 + \xi_i^{jk} \quad \text{si} \quad y_j = k \\
& \xi_i^{jk} \geq 0
\end{aligned} \tag{41}
$$

Hsu y Lin [2] deciden usar la siguiente estrategia basada en votos: si $\text{sign}((\boldsymbol{w}^{jk})^T\phi(\boldsymbol{x})+b^{jk})$ dice que $\boldsymbol{x}$ pertenece a la j-ésima clase, se suma un voto a esa clase. Si dice que pertenece a la clase k, entonces se da el voto a la clase k. La clase que se elige para $\boldsymbol{x}$ es aquella que tenga más votos.

6. Ejercicios

1. La MVS también funciona como un modelo para problemas de regresión. Determinar cuáles son las modificaciones que se hacen para ajustar la MVS para este problema. Incluir ideas gráficas, modelos, la solución del problema de optimización y demás detalles que consideren importantes.

2. Cuando se construyo el modelo de la MVS con margen suave, se introdujo una variable de holgura ξ_i que penaliza los errores de clasificación en el primal del problema de optimización. Tradicionalmente, la función de pérdida que se usa es la pérdida de Hinge. De esta forma, se puede resolver la MVS con gradiente descendiente de forma iterativa para encontrar los parámetros del modelo.

   a) Escriban la expresión de la función de pérdida de Hinge y su gráfica.

   b) Investiguen como se incrusta esta función de pérdida en el desarrollo de la MVS y resuelvan el problema de optimización resultante.

   c) ¿Cómo se interpreta la función de pérdida de Hinge al darle un valor a cierto punto mal o bien clasificado?

3. Investigar qué es el teorema de Mercer y cuál es su relación con la MVS.

4. Investigar sobre los siguientes kernels; su formulación, usos, ventajas y desventajas.

   a) Kernel de la función de Bessel.

   b) Kernel radial ANOVA.

   c) String kernel.

7. Registro de Actualizaciones

12/2023 Corrección de errores ortográficos y homogeneidad de las expresiones matemáticas. Ajuste de las secciones:

- Se separó el caso del margen suave del caso no lineal.
- Ajuste en los ejercicios para los estudiantes.

01/2022 Primera versión del escrito.

Referencias

[1] W. Chao. A tutorial for support vector machine. Disponible en: http://disp.ee.ntu.edu.tw/ pujols/Support %20Vector %20Machine.pdf.

[2] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

[3] C. K. Sahu and M. Sharma. Hinge loss in support vector machines. Disponible en: https://www.niser.ac.in/ smishra/teach/cs460/23cs460/lectures/lec11.pdf.

[4] V. Vapnik and C. Cortes. Support-vector networks. *Machine Learning*, 20:273–297, 1995.