Bloque 3: Neurología-Electroencefalograma

Máquinas de Vectores Soporte (SVM)

Luis Bote Curiel
Francisco Manuel Melgarejo Meseguer

DTSC

Curso 24-25

Índice

Introducción

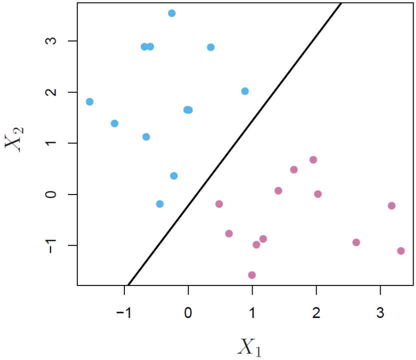
Hiperplanos

Hiperplano

- Dado un espacio $\mathbb{R}^P$, podemos definir un hiperplano como un subespacio afín y plano ($\mathbb{R}^{P-1}$).
- En $\mathbb{R}^3$ el hiperplano es un plano y en $\mathbb{R}^2$ sería una recta.
- Matemáticamente, podemos definirlo como

$$\boldsymbol{w}^T\boldsymbol{x} + \beta = 0 \rightarrow w_1x_1 + w_2x_2 + \beta = 0$$

- Esta ecuación nos indica que los puntos que están sobre el hiperplano tendrán como solución 0 y los que no tengan 0 estarán a un lado o al otro del hiperplano.



Clasificación mediante hiperplanos

Objetivo

- Partimos de una matriz $\boldsymbol{X}$ de N muestras y P dimensiones, que está dividida en sujetos de dos clases que pueden ser perfectamente divididos mediante un clasificador lineal.
- **Objetivo:** Encontrar el hiperplano que mejor divida estas dos clases.

Óptimo

- Supongamos que existe y somos capaces de encontrarlo, de manera que utilizando la ecuación anterior podemos escribir

$$\left.\begin{array}{l} \boldsymbol{w}^T \boldsymbol{x}_i + \beta > 0 \text{ si } y_i = 1 \\ \boldsymbol{w}^T \boldsymbol{x}_i + \beta > 0 \text{ si } y_i = -1 \end{array}\right\} y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \beta) > 0$$

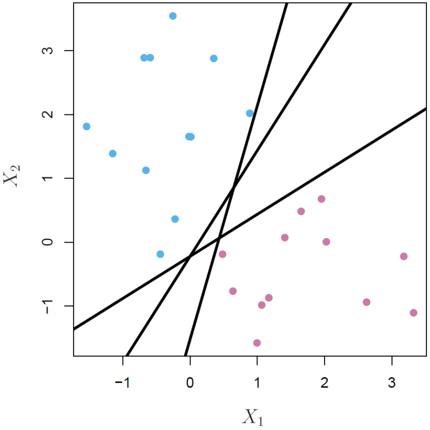
- Además de clasificar, este hiperplano nos permite saber *como de fiable* es una clasificación, valores absolutos más altos de $\boldsymbol{w}^T \boldsymbol{x}_i + \beta >$ nos dicen que más sólida es la clasficación.

Clasificador de Vectores de Soporte

Problema

Problema

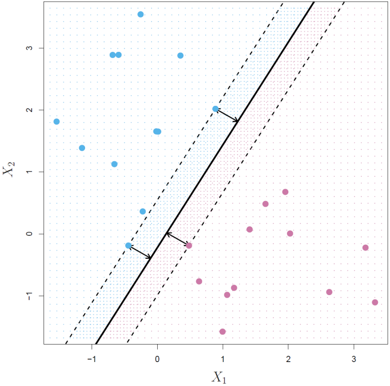
- Todo esto está muy bien, pero el número de hiperplanos que pueden separar dos clases binarias es infinito.
- ¿Cómo sabemos cuál es el óptimo?



Margen máximo

Margen

- Llamaremos margen a la distancia entre cada punto y el hiperplano separador.
- Diremos que el hiperplano óptimo será aquel cuya distancia a los puntos más cercanos al hiperplano a ambos lados de él sea igual y la llamaremos M.
- Ahora el problema será buscar la combinación $(\boldsymbol{w}, \beta)$ que haga esa distancia máxima.



Vectores soporte

Soporte

- A los puntos que se encuentran a distancia M del hiperplano, los llamaremos **vectores soporte**, ya que son aquellos que *soportan* el margen máximo, sin ellos, este margen sería otro.
- Cabe destacar que para facilitar todo el proceso, $M = \frac{1}{||\boldsymbol{w}||}$

Problema

$$\begin{aligned} \max_{\boldsymbol{w},\beta} \quad & M \\ s.a: \quad & \sum_{i}^{P} w_i^2 = 1 \\ & y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \beta) \geq M \quad \forall i \end{aligned}$$

$$\begin{aligned} \min_{\boldsymbol{w},\beta} \quad & \frac{1}{2}||\boldsymbol{w}||^2 \\ s.a: \quad & y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \beta) \geq 1 \quad \forall i \end{aligned}$$

Minimización mediante Lagrange-Wolfe

Lagrange-Wolfe

El método de Lagrange-Wolfe se utiliza para resolver problemas de optimización no lineal con restricciones. Como por ejemplo Minimizar $f(x)$ sujeto a $g_i(x) \leq 0$ y $h_j(x) = 0$.

- **Lagrangiano**: $\mathscr{L}_P(x, \alpha, \mu) = f(x) + \sum \alpha_i g_i(x) + \sum \mu_j h_j(x)$.
- **Función Dual**: $\mathscr{L}_D(x, \alpha, \mu)$ derivar el lagrangiano con respecto a x, α y μ y sustituirlas en el Lagrangiano.
- **Problema Dual**: Maximizar $\theta(\alpha, \mu)$ sujeto a $\alpha_i \geq 0$.

Condiciones KKT

Las condiciones de Karush-Kuhn-Tucker (KKT) son necesarias para la optimalidad en problemas de programación no lineal.

- **Factibilidad Primal**: $g_i(x^*) \leq 0$ y $h_j(x^*) = 0$.
- **Factibilidad Dual**: $\alpha_i \geq 0$.
- **Complementariedad**: $\alpha_i g_i(x^*) = 0$.
- **Estacionaridad**: $\nabla \mathscr{L} = 0$.

Resolviendo el problema de minimización

Resolución

- Definimos el problema primal:

$$\min_{\boldsymbol{w},\beta} \quad \frac{1}{2}||\boldsymbol{w}||^2$$
$$s.a: \quad 1 - y_i(\boldsymbol{w}^T\boldsymbol{x} + \beta) \leq 0 \quad \forall i \in \{1,\ldots,N\}$$

- Calculamos el Lagrangiano:

$$\mathcal{L}_P = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_i^N \alpha_i [y_i(\boldsymbol{w}^T\boldsymbol{x}_i + \beta) - 1]$$

- Calculamos las derivadas:

$$\frac{dL}{d\boldsymbol{w}} \to \boldsymbol{w} = \sum_i^N \alpha_i y_i x_i, \quad \frac{dL}{d\beta} \to \sum_i^N \alpha_i y_i = 0, \quad \frac{dL}{d\alpha} \to 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + \beta) \leq 0 \ \ \forall i \in \{1,\ldots$$

Resolviendo el problema de minimización II

Resolución

- Calculamos el dual:

$$\mathscr{L}_D = -\sum_i^N \alpha_i + \frac{1}{2}\sum_i^N\sum_j^N \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

- Planteamos la optimización (se resuelve mediante algún método conocido), ya que el problema es convexo.

$$\begin{aligned}
\max_{\boldsymbol{w},\beta} \quad & \mathscr{L}_D \\
s.a: \quad & \sum_i^N \alpha_i y_i = 0 \\
& \alpha_i > 0 \quad \forall i \in \{1,\dots,N\} \\
& \alpha_i [y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \beta) - 1] = 0 \quad \forall i \in \{1,\dots,N\}
\end{aligned}$$

- Una vez obtenido el resultado deberemos comprobar que cumple las condiciones KKT.

Solución

Solución

- Como hemos podido ver tenemos unas restricciones que nos obligan a que $\alpha_i > 0$ y que $\alpha_i[y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \beta) - 1] = 0$, esto nos fuerza que los puntos solución $\boldsymbol{x}^*$, $y_i[(x_i^*)^T \boldsymbol{w} + \beta] - 1$ sean **vectores soporte** ya que son aquellos que se encuentran a la distancia M.
- Además, fuerza a que todos los demás α_i sean 0.
- Podemos comprobar que los $\boldsymbol{x}_i^*$ cumplen las condiciones de KKT.
- Las soluciones generales tendrán la siguiente forma:

$$\boldsymbol{w} = \sum_i^S \alpha_i y_i \boldsymbol{x}_i^*, \quad f(\boldsymbol{x}) = \sum_i^S \alpha_i y_i \boldsymbol{x}_i^* \boldsymbol{x}^T + \beta$$

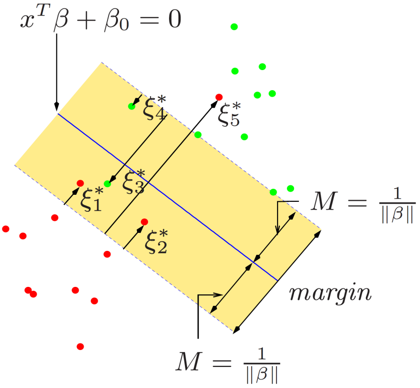
Caso no separable

El Mundo Real

- Consideraremos ahora que no existe una separación perfecta, es decir, que hay algunas observaciones que van a estar mal clasificadas (van a caer en el otro lado del hiperplano).
- De esta manera podemos escribir la función de la zona entre márgenes como

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + \beta) \geq M(1 - \xi_i)$$

- Hay que tener en cuenta que, aunque, podríamos escribir $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + \beta) \geq M - \xi_i$ esto es un dolor de cabeza ya que no es convexo.



Caso no separable II

Resolución

- Hay que tener en cuenta que $\xi_i > 0$ y que $\sum_i^N \xi_i \leq C$, es decir, todos las variables de holgura *fastidian* la clasificación y el error global es constante.
- Problema de minimización

$$
\begin{aligned}
\min_{\boldsymbol{w},\beta} \quad & \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i^N \xi_i \\
s.a: \quad & y_i(\boldsymbol{w}^T\boldsymbol{x} + \beta) \geq 1 - \xi_i \quad \forall \boldsymbol{i} \in \{1, \ldots, N\} \\
& \xi_i \geq 0
\end{aligned}
$$

- Escribimos el primal

$$
\mathscr{L}_p = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i^N \xi_i - \sum_i^N \alpha_i[y_i(\boldsymbol{w}^T\boldsymbol{x}_i + \beta) - (1 - \xi_i)] - \sum_i^N \mu_i \xi_i
$$

Caso no separable III

Resolución

- Escribimos el dual:

$$\mathscr{L}_D = \sum_i^N \alpha_i - \frac{1}{2}\sum_i^N \sum_j^N \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i \boldsymbol{x}_j^T$$

- Planteamos la optimización (se resuelve mediante algún método conocido), ya que el problema es convexo.

$$\begin{aligned}
\max_{\boldsymbol{w},\beta} \quad & \mathscr{L}_D \\
s.a: \quad & \sum_i^N \alpha_i y_i = 0 \\
& 0 \leq \alpha_i \leq C \quad \forall i \in \{1,\dots,N\} \\
& \alpha_i [y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \beta) - (1-\xi_i)] = 0 \quad \forall i \in \{1,\dots,N\} \\
& \mu_i \xi_i = 0 \quad \forall i \in \{1,\dots,N\} \\
& y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \beta) - (1-\xi_i) \geq 0 \quad \forall i \in \{1,\dots,N\}
\end{aligned}$$

Caso no separable IV

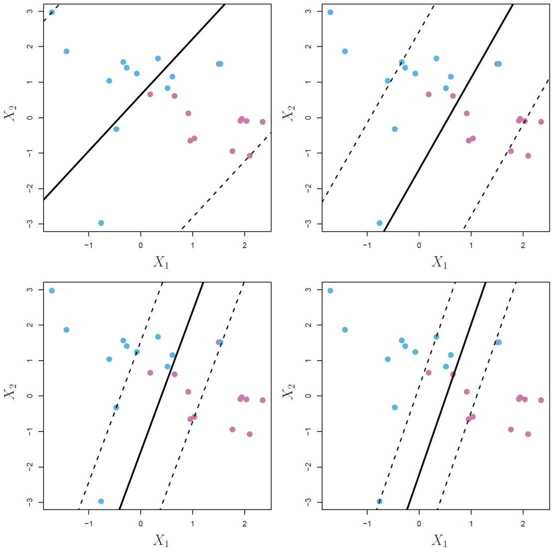
Resolución

- Y para sorpresa de nadie, el resultado vuelve a ser

$$\boldsymbol{w} = \sum_{i}^{S} \alpha_i y_i \boldsymbol{x}_i^*$$

- Al igual que en caso anterior, nos faltaba por definir el valor de β, para ello, debemos tomar una media de todos los posibles valores del parámetro una vez obtenido los vectores soporte.
- ¿Qué ocurre con C? Es un parámetro que le damos al sistema como entrada y que controlará el nivel de error que puede existir.
- Otra forma de controlar el número de errores y el overfitting, es limitando el número máximo de vectores soporte que pueden existir, esa implementación de la SVC se les conoce como ν-SVC/SVM y ese parámetro ν indica la fracción del total de elementos que pueden ser candidatos a vector soporte.

Efecto de C

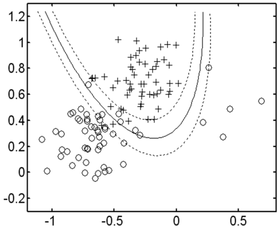


Máquinas de Vectores de Soporte

El mundo real

Fuera de la Linealidad

- En el mundo real, las cosas no suelen ser linealmente separables. Por ello, esta formulación básica adolece bastante frente a esos problemas.
- Sería genial poder tener alguna herramienta que nos permitiese utilizar algo tan fácil como los productos escalares de algunos elementos para poder resolver el problema.
- Una opción sería coger los datos, hacer alguna transportación y que sean linealmente separables en otro espacio distinto al original.



No linealidades

Aumento de Dimensionalidad

- Suponga que cogemos un problema cuyos datos se encuentran en $\mathbb{R}^2$ y nos los llevamos a $\mathbb{R}^6$ mediante esta transformación
$\boldsymbol{x} = (x_1, x_2) \rightarrow z_1 = 1,\ z_2 = \sqrt{2}x_1,\ z_3 = \sqrt{2}x_2,\ z_4 = \sqrt{2}x_1 x_2,\ z_5 = x_1^2,\ z_6 = x_2^2.$
- Si aplicamos este proceso a $\boldsymbol{u} = (u_1, u_2)$ y a $\boldsymbol{v} = (v_1, v_2)$ y realizamos la multiplicación de los resultantes obtenemos

$$1 + 2u_1 v_1 + 2u_2 v_2 + 2u_1 u_2 v_1 v_2 + (u_1 v_1)^2$$

- Si agrupamos términos vemos que es análogo a escribir $(1 + (\boldsymbol{u} \cdot \boldsymbol{v}))^2$, es decir, aplicar este último resultado es equivalente a realizar un producto escalar en $\mathbb{R}^6$
- Entonces podemos buscar *operaciones* que sean análogas a realizar productos escalares en otras dimensiones y así probar si en ese espacio son linealmente separables los datos. A estas operaciones que induce una transformación y un producto vectorial se le conoce como **Kernel**
- Problema: Sería una tarea tremendamente compleja de no haber sido por Mercer.

Teorema de Mercer

Enunciado

Si $K(x, y)$ es una función simétrica y positiva definida, entonces existe una serie de funciones ortonormales $\phi_i(x)$ y valores propios no negativos λ_i tales que:

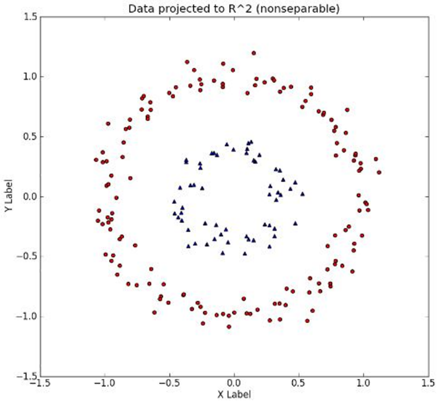
$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y)$$

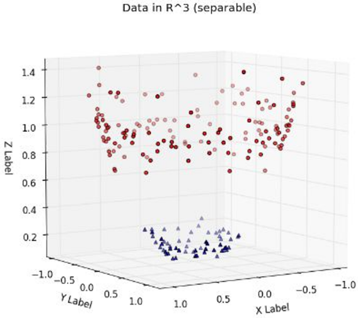
donde la serie converge absolutamente y uniformemente.

Condiciones sobre la matriz de Gram

- Podemos aplicar estas condiciones a la matriz de Gram, formada por evaluar el kernel en todos los pares de puntos de datos $\boldsymbol{K} = K(x, x)$
- La matriz $\boldsymbol{K}$ debe ser simétrica.
- La matriz debe ser definida positiva, es decir, $\mathbf{x}\boldsymbol{K}\mathbf{x}^T > 0$ para cualquier $\mathbf{x}$ no nulo

Ejemplo Kernel





Kernels más utilizados en SVM

1. Kernel Lineal

- **Fórmula**: $K(x, y) = x \cdot y$
- **Uso**: Datos linealmente separables

2. Kernel Polinómico

- **Fórmula**: $K(x, y) = (x \cdot y + c)^d$
- **Uso**: Datos no linealmente separables con relaciones polinómicas

3. Kernel de Función de Base Radial (RBF)

- **Fórmula**: $K(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$
- **Uso**: Datos con patrones complejos y no lineales

4. Kernel Sigmoidal

- **Fórmula**: $K(x, y) = \tanh(\alpha x \cdot y + c)$
- **Uso**: Modelos neuronales y datos con relaciones no lineales

Clasificación One-Vs-One

Descripción

- Se utiliza cuando hay $K > 2$ clases.
- Se construyen $\binom{K}{2}$ clasificadores SVM, cada uno comparando un par de clases.
- Cada SVM compara la clase k (codificada como $+1$) con la clase k' (codificada como -1).

Clasificación

- Se clasifica una observación usando cada uno de los $\binom{K}{2}$ clasificadores.
- Se cuenta el número de veces que la observación es asignada a cada una de las K clases.
- La clasificación final se realiza asignando la observación a la clase a la que fue asignada con mayor frecuencia.

Clasificación One-Vs-All

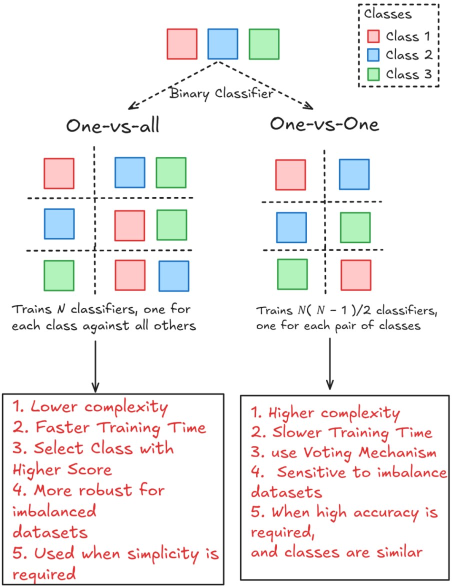
Descripción

- Se utiliza cuando hay $K > 2$ clases.
- Se ajustan K clasificadores SVM, cada uno comparando una clase con las $K - 1$ clases restantes.
- Los parámetros resultantes de ajustar un SVM para la clase k son $\beta_k, w_{1k}, \ldots, w_{pk}$.

Clasificación

- Para una observación d x, se calcula $\beta_k + w_{1k}x_1 + \beta_{2k}x_2 + \cdots + w_{pk}x_p$ para cada clase k.
- Se asigna la observación a la clase para la cual este valor es mayor, indicando un alto nivel de confianza en que la observación pertenece a esa clase.

Resumen


Classes
Class 1
Class 2
Class 3
Binary Classifier
One-vs-all
One-vs-One
Trains N classifiers, one for each class against all others
Trains N(N − 1)/2 classifiers, one for each pair of classes
1. Lower complexity
2. Faster Training Time
3. Select Class with Higher Score
4. More robust for imbalanced datasets
5. Used when simplicity is required
1. Higher complexity
2. Slower Training Time
3. use Voting Mechanism
4. Sensitive to imbalance datasets
5. When high accuracy is required, and classes are similar