UNIVERSITAS Miguel Hernández

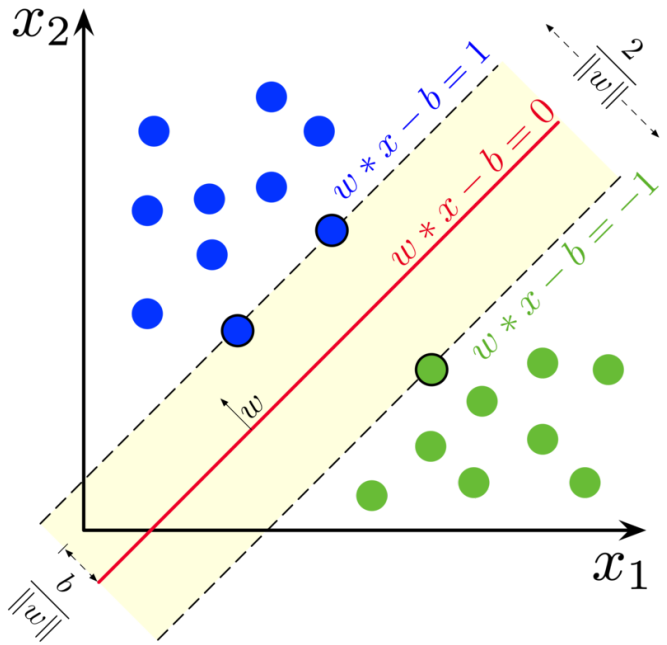
Tema 4. Support Vector Machine (SVM).

José L. Sainz-Pardo Auñón

TÉCNICAS ESTADÍSTICAS PARA EL APRENDIZAJE II

Máster Universitario en Estadística Computacional y Ciencia de Datos para la Toma de Decisiones.

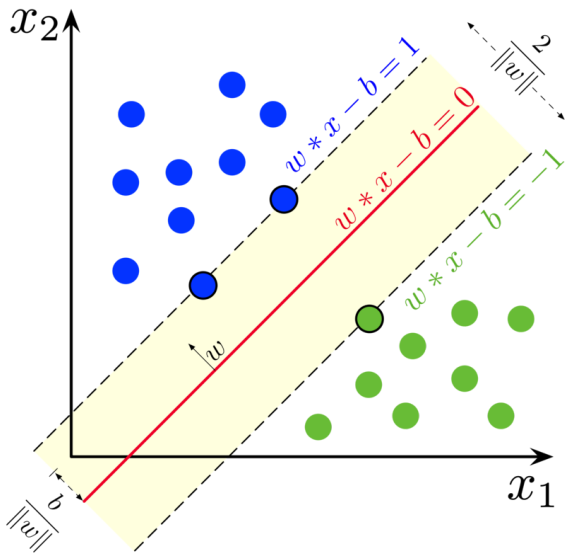
Ejemplo del problema



- El objetivo es separar dos tipos de datos mediante un hiperplano.
- Queremos que los puntos de diferentes clases queden en lados opuestos del hiperplano.
- Además, la distancia de los puntos extremos al hiperplano debe estar balanceada.

¿Cuál es el hiperplano que logra esto?

Definición formal del problema



- **Support Vector Machine (SVM)** busca formular un hiperplano que separe óptimamente dos clases.
- Se maximiza el **margen** ($\frac{2}{||w||}$), es decir, la distancia entre los puntos más cercanos al hiperplano.
- El problema de optimización asociado es minimizar $||w||$, sujeto a las restricciones que garantizan que los puntos se encuentren correctamente clasificados por el hiperplano $wx - b$.
- Los puntos más cercanos al hiperplano, conocidos como **vectores de soporte**. Determinan la posición y orientación del hiperplano.

Definición formal del problema

Cálculo del margen:

- El hiperplano separador se define como:

$$w^T x + b = 0$$

- La distancia de un punto x al hiperplano $w^T x + b = 0$ es:

$$\text{Distancia} = \frac{\text{Proyección del vector } x_0 \text{ sobre el vector normal } w}{\text{Norma del vector normal } ||w||} = \frac{|w^T x + b|}{||w||}$$

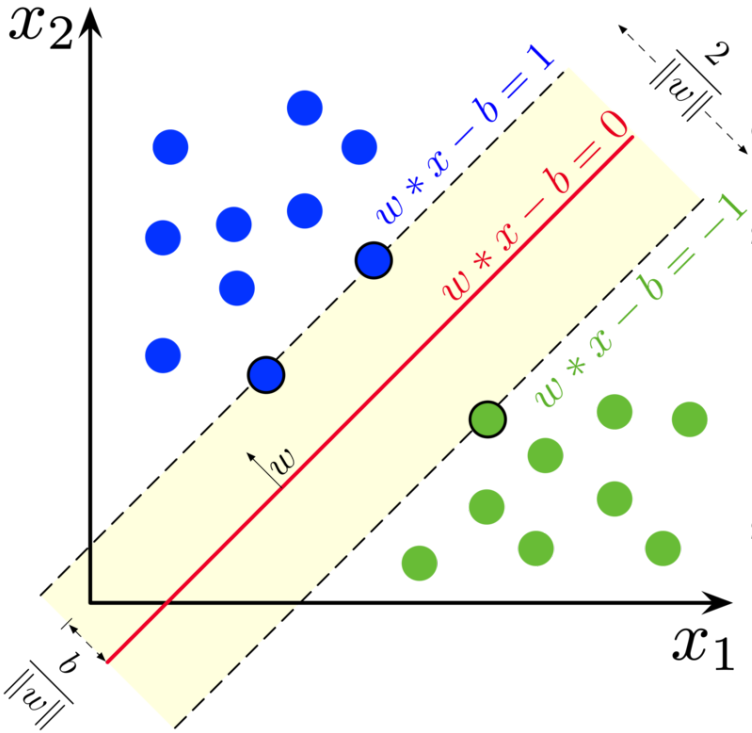
- Dado que, para las clases $+1$ y -1, los hiperplanos vienen definidos por:

$$w^T x + b = +1 \quad \text{y} \quad w^T x + b = -1$$

- Finalmente, la distancia entre los hiperplanos es:

$$\frac{|1 - (-1)|}{||w||} = 2\frac{1}{||w||} = \frac{2}{||w||}$$

Definición formal del problema



Dado que maximizar el margen, es decir, $\max \frac{2}{||w||_k}$ es equivalente a minimizar $||w||_k$, podemos definir el problema de obtención del Vector Soporte $wx - b$ como el siguiente problema de optimización:

$$\min_{w,b,\xi} \quad ||w||_k + C \sum_{i=1}^{n} \xi_i$$

s.a:

$$y_i \left(w^T x_i + b \right) \geq 1 - \xi_i, \quad i = 1, \ldots, n$$

$$\xi_i \geq 0, \quad i = 1, \ldots, n$$

Definición formal del problema

Variables:

- $w \in \mathbb{R}^d$ son los coeficientes del hiperplano.
- $b \in \mathbb{R}$ es el término de sesgo o desplazamiento.
- $\xi_i \geq 0$ son las variables de holgura que permiten violaciones del margen (para márgenes suaves).

Coeficientes y parámetros:

- C es un parámetro que controla la penalización de los errores de clasificación.
- $y_i \in \{-1, 1\}$ son las etiquetas de las clases para cada observación $x_i \in \mathbb{R}^d$.
- $x_i \in \mathbb{R}^d$ son los puntos de datos de entrenamiento.

Generalized SVM

- El SVM se puede generalizar para diferentes normativas L_k.
- Para la norma L_1, o distancia de Manhattan, el problema es lineal y puede resolverse mediante **programación lineal**.

$$\min_{w,b,\xi} \quad \sum_{j=1}^{p} |w_j| + C \sum_{i=1}^{n} \xi_i$$

s.a:

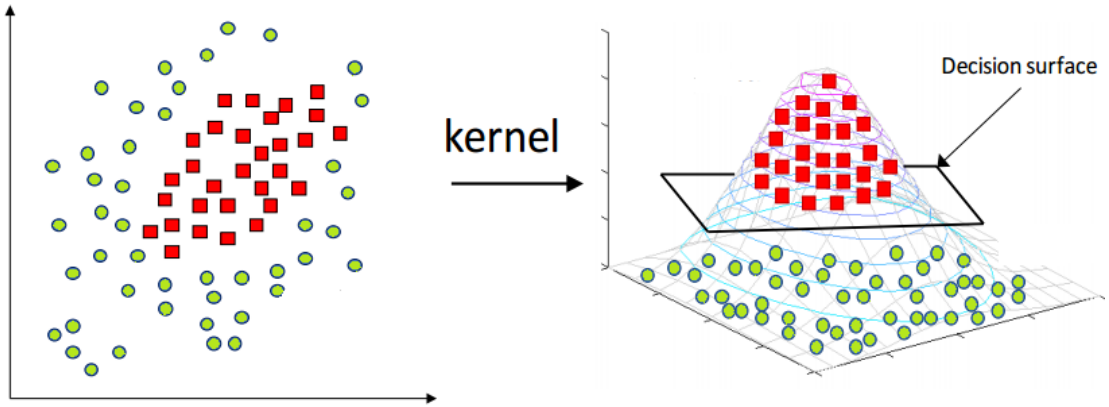
$$y_i \left(w^T x_i + b \right) \geq 1 - \xi_i, \quad i = 1, \ldots, n$$

$$\xi_i \geq 0, \quad i = 1, \ldots, n$$

(faltaría desdoblar en dos variables las variables $|w_j|$ en $|w_j^+|$ y $|w_j^-|$ para su completa linealización).

- Para la norma L_2, o distancia euclídea, el problema ya no es lineal y puede resolverse mediante métodos como los operadores de **Lagrange** o usando el **truco del núcleo**.

Truco del Núcleo (Kernel Trick)



- En problemas donde los datos no son linealmente separables, podemos proyectarlos en un espacio de mayor dimensión donde sí lo sean.
- Esto se logra mediante el **truco del núcleo**, que permite proyectar los puntos a un espacio de características de mayor dimensión sin hacer explícita la transformación.
- **Ejemplos de núcleos**:
  - **Núcleo Lineal**: $K(x, x') = x^T x'$
  - **Núcleo Polinomial**: $K(x, x') = (x^T x' + c)^d$
  - **Núcleo Gaussiano (RBF)**: $K(x, x') = \exp(-\gamma ||x - x'||^2)$

Ejercicio: SVM con Distancia de Manhattan

Edad	Salario	Compra
20	1000	1
30	1500	1
40	2000	0

- Se tiene una tabla con datos de edad y salarios de clientes, así como si compraron un producto (compra=1) o no (compra=0).
- Plantea el problema utilizando la **distancia de Manhattan** para obtener la función discriminante SVM que diferencia a los clientes que compran de los que no.
- Ahora plantea el problema utilizando la **distancia euclídea**.

Algunas observaciones

- Si pudiera realizarse una clasificación perfecta, es decir, una división totalmente correcta entre los datos, no serían necesarios los términos de error.
- En caso de no incluirse los términos de error y no poder realizarse una clasificación perfecta, el problema no tendría solución.
- Por tanto, las variables ϵ_i miden los errores de clasificación. Aquel individuo i cuyo $\epsilon_i = 0$ están correctamente clasificados.

Algunas observaciones

- El parámetro C (coste) controla el balance entre maximizar el margen y minimizar los errores.
- Un valor bajo de C significa que el modelo permite un margen más amplio, priorizando una mejor generalización, incluso si se permiten algunos errores de clasificación en los datos de entrenamiento. En este caso, el SVM es más tolerante a los errores de clasificación (sobreajuste bajo pero también potencial subajuste).
- Un valor alto de C implica una penalización más fuerte para los errores de clasificación, lo que lleva a un margen más estrecho, es decir, el SVM intentará clasificar todos los puntos correctamente, incluso si el margen es más pequeño. Esto puede resultar en un ajuste excesivo (sobreajuste).
- Para calibrar el parámetro C se realizan pruebas con diferentes valores de C para encontrar el que mejor ajuste realiza (una vez más se puede utilizar validación cruzada).

SVM con más de dos categorías

- Para clasificar en $k > 2$ categorías, se necesitan $k - 1$ hiperplanos.
- Un método común es el de **uno frente al resto**, donde se entrenan $k - 1$ clasificadores binarios. Para clasificar en tres categorías, se construyen dos clasificadores y se resuelve un único problema de optimización.
- Si un individuo es asignado a más de un grupo, se utiliza el valor absoluto más alto del discriminador.

SVM con más de dos categorías

Ejemplo. Si para clasificar en tres categorías, se han construido dos con la siguiente codificación:

Clase	Clasificador 1	Clasificador 2
1	1	-1
2	-1	1
3	-1	-1

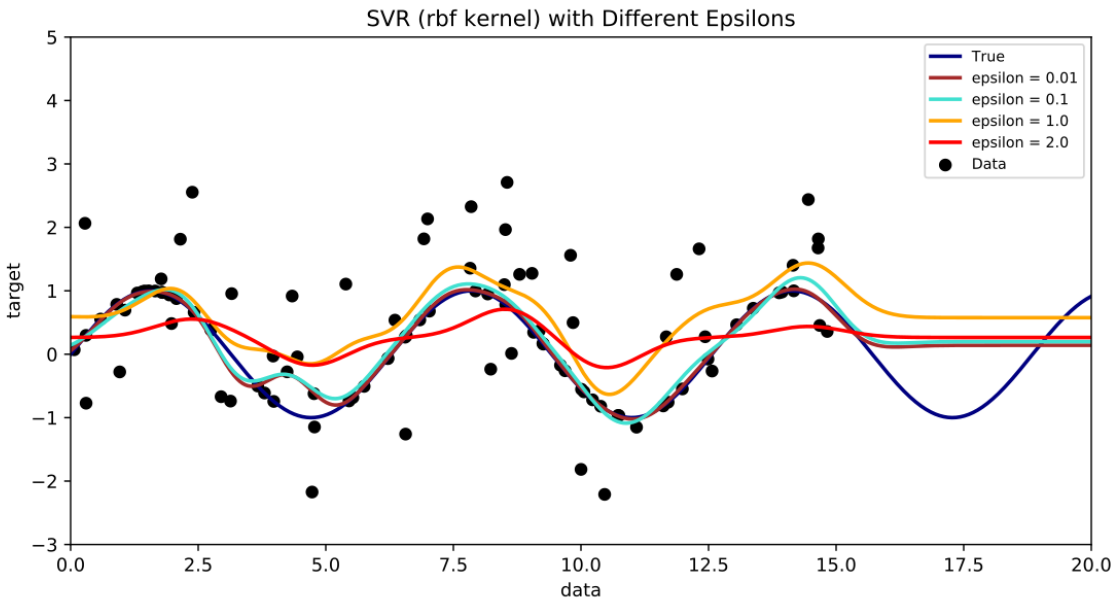
Obtén la asignación de los siguientes 5 individuos cuyos valores en los clasificadores es:

Individuo	Clasificador 1	Clasificador 2	Clase
1	2.5	3.0	
2	2.5	2.0	
3	-2	3.5	
4	2	2.5	
5	-3.5	-4.0	

Support Vector Regression

- El utilizar SVM como un método de regresión (predecir un valor continuo), es conocido como SVR (Support Vector Regression).
- SVR intenta encontrar una función que tenga un margen de error aceptable (ϵ).
- En lugar de encontrar el mejor separador, se busca una función que esté dentro de este margen de error para la mayoría de los puntos de datos.
- El objetivo de SVR es minimizar la norma del vector de pesos, lo que permite que se ignore el error dentro del margen ϵ. Fuera de este margen, los errores se penalizan.
- Además de calibrar el parámetro C se precisa calibrar también el parámetro ϵ. Una vez más, la validación cruzada puede sernos útil para ello.
- Al igual que con SVM, pueden utilizarse distintos núcleos para transformar los datos.

Support Vector Regression



El modelo general de programación matemática de SVR es:

$$\min_{w,b,\xi} \quad ||w||_k + C\sum_{i=1}^{n} |\xi_i|$$

s.a:

$$y_i - \left(w^T x_i + b\right) \leq \epsilon + \xi_i, \quad i = 1, \ldots, n$$

$$\xi_i \geq 0, \quad i = 1, \ldots, n$$


UNIVERSITAS
Miguel Hernández