Machine Learning 101

Support Vector Machines

Felipe Alonso Atienza

Data Scientist @BBVA

Introducción

- Máquinas de vectores (de) soporte, del inglés, *Support Vector Machines*

- Inicialmente concebidas para problemas de clasificación, y posteriormente extendidas a regresión.
  - SVC: *Support Vector Classification*
  - SVR: *Support Vector Regression*

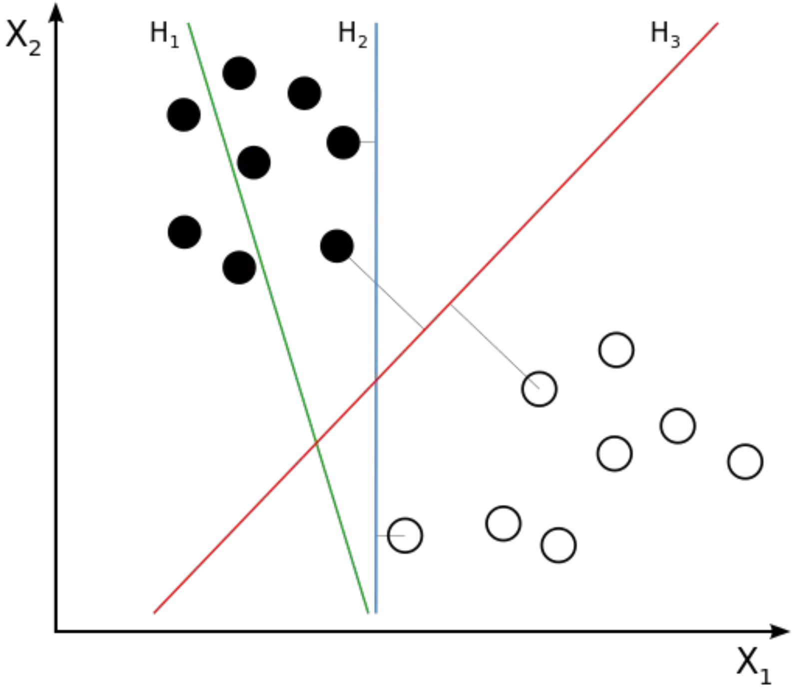
- Se definen como **clasificadores lineales de máximo margen**

- Propuestas a mediados-finales de los 90s, con mucho auge en los 2000s
  - Grandes prestaciones en aprendizaje supervisado
  - Métodos Kernel

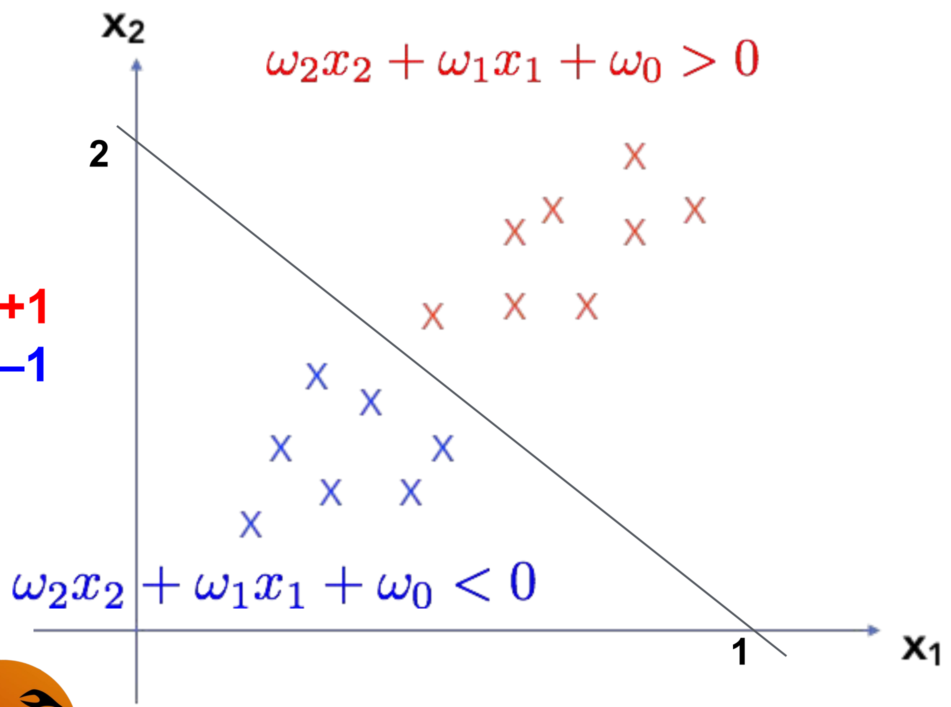
Índice

Intuición

- Clasificador lineal definido por un hiperplano separador de máximo margen



Plano separador



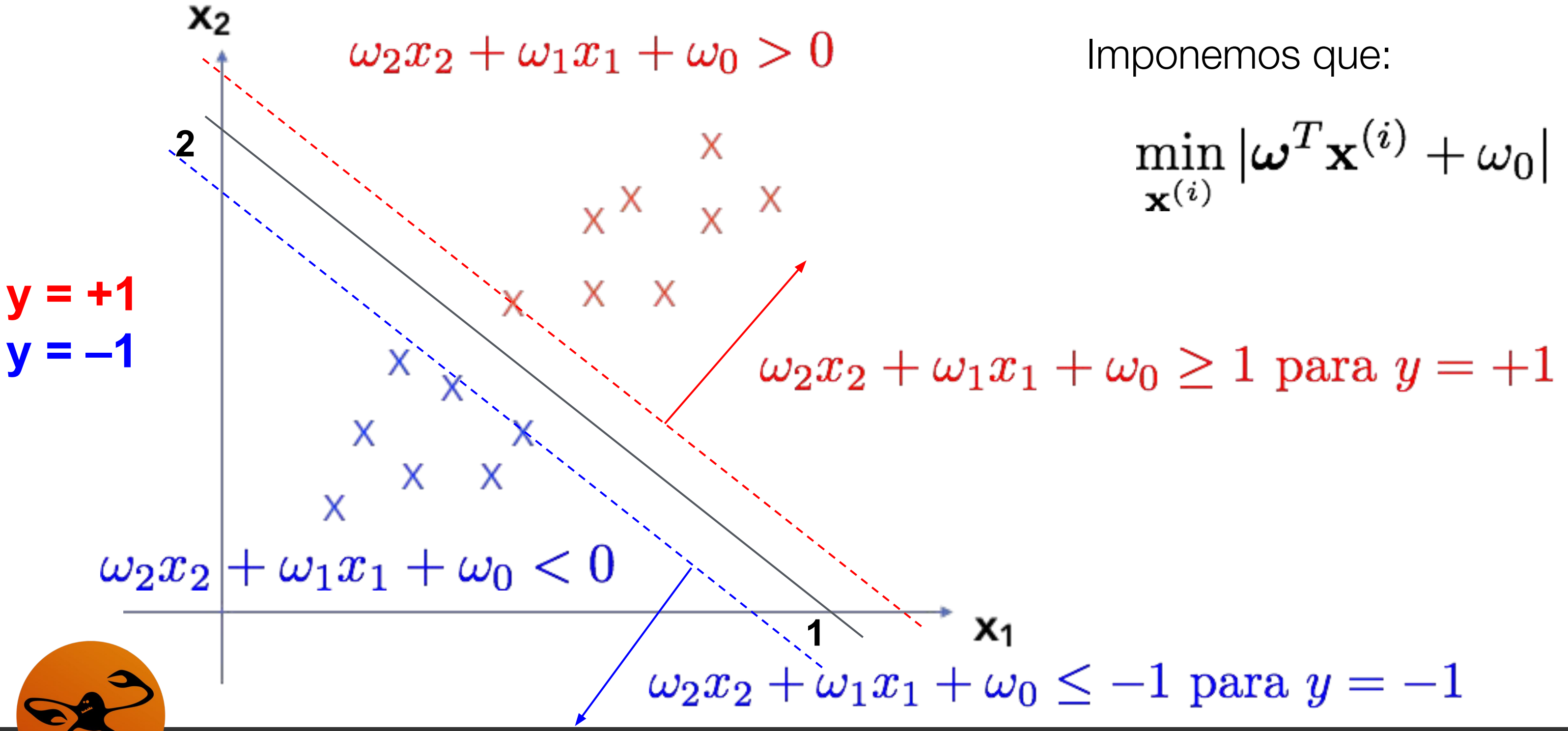
y = +1
y = –1

$$x_2 = m x_1 + n$$
$$x_2 = -2x_1 + 2$$
$$x_2 + 2x_1 - 2 = 0$$

$$\omega_2 x_2 + \omega_1 x_1 + \omega_0 = 0$$
$$\boldsymbol{\omega}^T \mathbf{x} + \omega_0 = 0$$

¿y cómo fuerzo a que sea de máximo margen?

Plano separador: condición

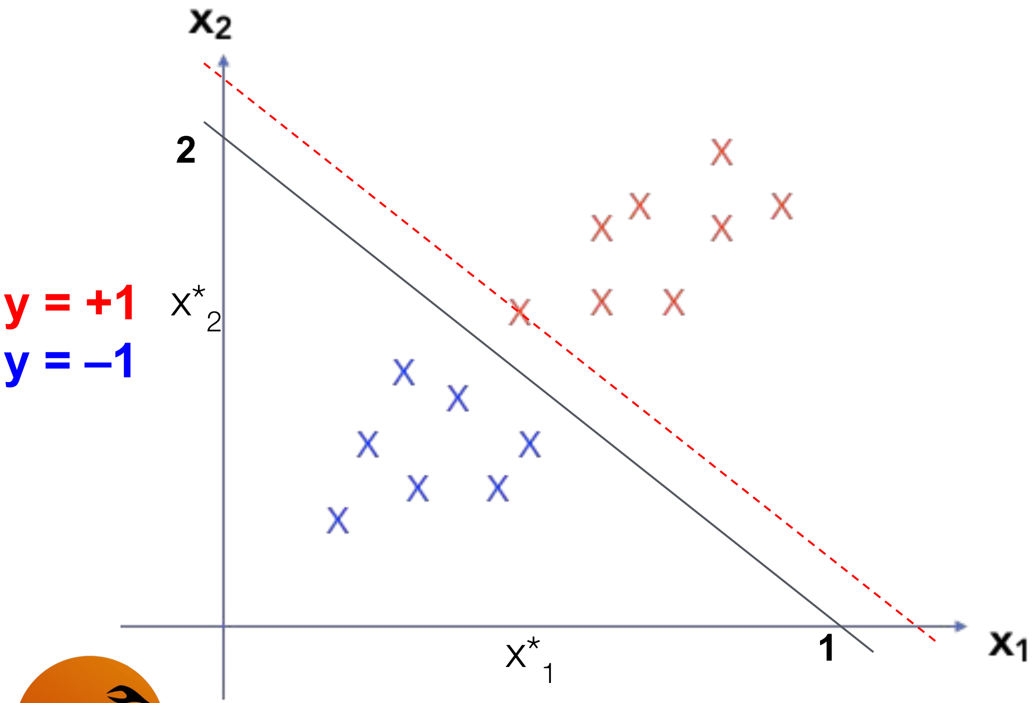


$$\omega_2 x_2 + \omega_1 x_1 + \omega_0 = 0$$

Imponemos que:

$$\min_{\mathbf{x}^{(i)}} \left| \boldsymbol{\omega}^T \mathbf{x}^{(i)} + \omega_0 \right| = 1$$

Plano separador: justificación



Si

$$\omega_2 x_2 + \omega_1 x_1 + \omega_0 = 0$$

es un plano separador, entonces

$$c \cdot (\omega_2 x_2 + \omega_1 x_1 + \omega_0) = 0$$

también lo es. Así, escojo *c* para que se cumpla la condición que me interesa.

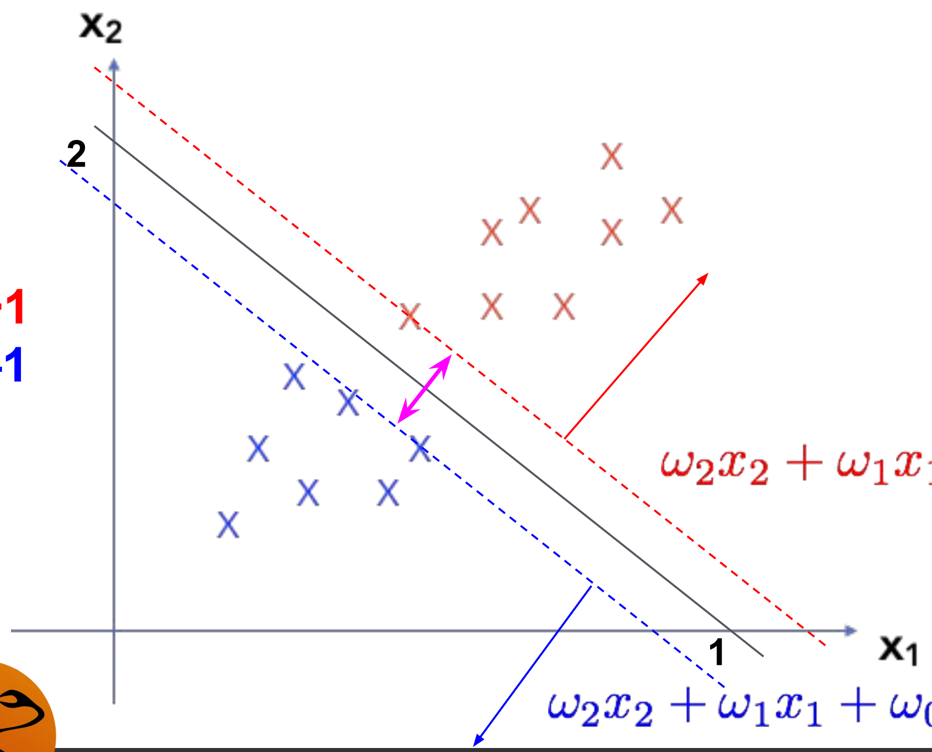
Ej: supongamos que x^*_1=0.5, y x^*_2 = 1.1, para el que quiero que

$$c \cdot (\omega_2 x_2^* + \omega_1 x_1^* + \omega_0) = 1$$

Entonces ¿cuánto vale c?

Plano separador: margen

y = +1
y = –1



El objetivo es **maximizar** el margen

$$\text{margen} = \frac{2}{||\boldsymbol{\omega}||}$$

bajo las condiciones anteriores, que pueden expresarse como

$$y\left(\omega_2 x_2 + \omega_1 x_1 + \omega_0\right) \geq 1$$

$$y\left(\boldsymbol{\omega}^T \mathbf{x} + \omega_0\right) \geq 1$$

$$\omega_2 x_2 + \omega_1 x_1 + \omega_0 \geq 1 \text{ para } y = +1$$

$$\omega_2 x_2 + \omega_1 x_1 + \omega_0 \leq -1 \text{ para } y = -1$$

Función de coste

- Clasificador lineal definido por un hiperplano separador de máximo margen

- Así, dado un conjunto de datos etiquetados

$$\{\mathbf{x}^{(i)}, y^{(i)}\} \text{ con } \mathbf{x}^{(i)} \in \mathbb{R}^d, y^{(i)} \in \{-1, +1\}$$

- El funcional a **minimizar** es

$$\min_{\boldsymbol{\omega}, \omega_0} \frac{1}{2} ||\boldsymbol{\omega}||_2^2 \text{ s.to } y^{(i)} \left(\boldsymbol{\omega}^T \mathbf{x}^{(i)} + \omega_0 \right) \geq 1, i = 1, \ldots, N$$

  - Problema de optimización **convexa** (solución única) ...
  - … con restricciones (Lagrangiano)

SVMs vs Regresión logística

- Clasificadores lineales los dos
- Máximo margen vs Mínimo error
- Optimización convexa vs Máxima verosimilitud
- ...

Índice

Support Vectors

- Se puede demostrar que la solución es

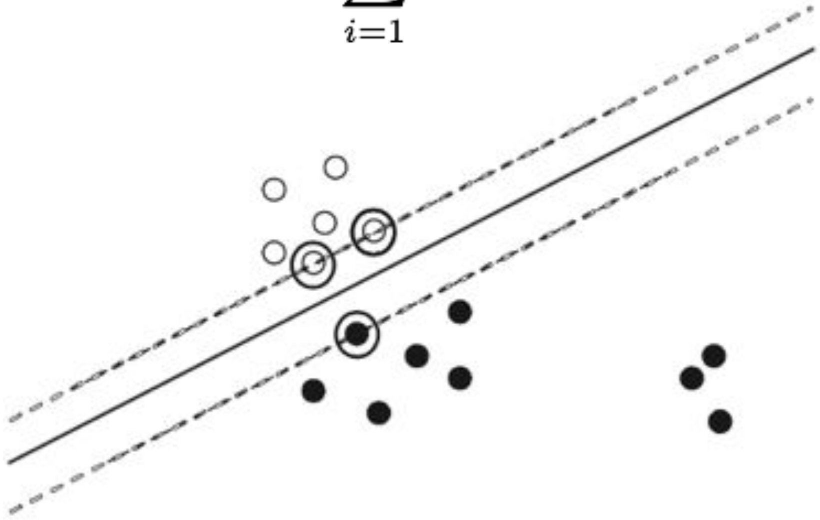
$$\hat{\omega} = \sum_{i=1}^{N} \alpha_i y^{(i)} \mathbf{x}^{(i)}$$

    - Una combinación lineal de las muestras de entrenamiento (recuerda este resultado cuando veas redes neuronales
    - $\alpha_i \geq 0$ pero para muchas muestras se cumple que $\alpha_i = 0$(**solución dispersa**)
    - Vectores soporte: muestras para las que $\alpha_i \neq 0$

Support Vectors

- Se puede demostrar que la solución es

$$\hat{\omega} = \sum_{i=1}^{N} \alpha_i y^{(i)} \mathbf{x}^{(i)}$$

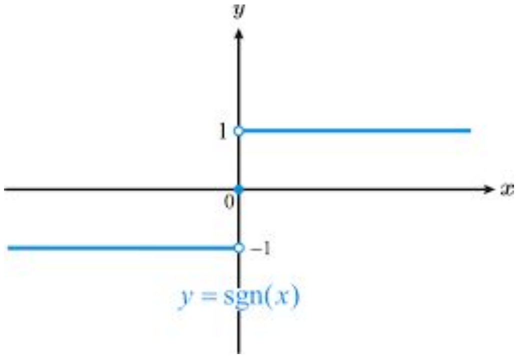


Frontera de separación

- Conocidos los pesos, la predicción se realiza a través de la fórmula

$$\hat{y} = f(\mathbf{x}) = \text{sign}\left(\hat{\boldsymbol{\omega}}^T\mathbf{x} + \hat{\omega}_0\right) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y^{(i)} \mathbf{x}^T \mathbf{x}^{(i)} + \hat{\omega}_0\right)$$

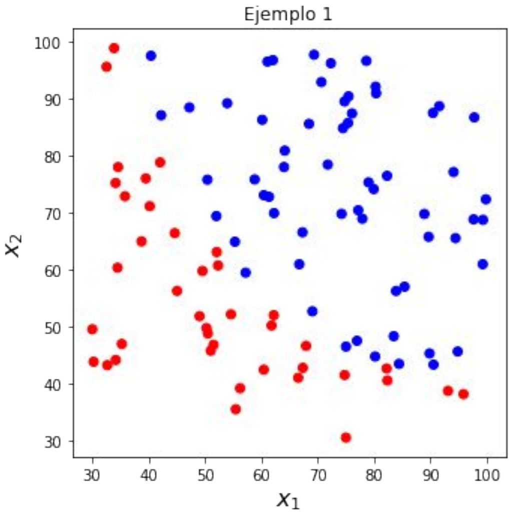
  - Producto escalar entre las muestras de entrenamiento y la muestra sobre la que quiero realizar la predicción



Índice

Caso linealmente no separable

- Hasta ahora hemos trabajado con un caso en el que las clases son claramente separables, esto es, no hay solapamiento entre ellas
- No hablamos de fronteras no lineales, seguimos considerando que existe un <u>hiperplano</u> capaz de separar las clases, aunque con <u>errores</u>

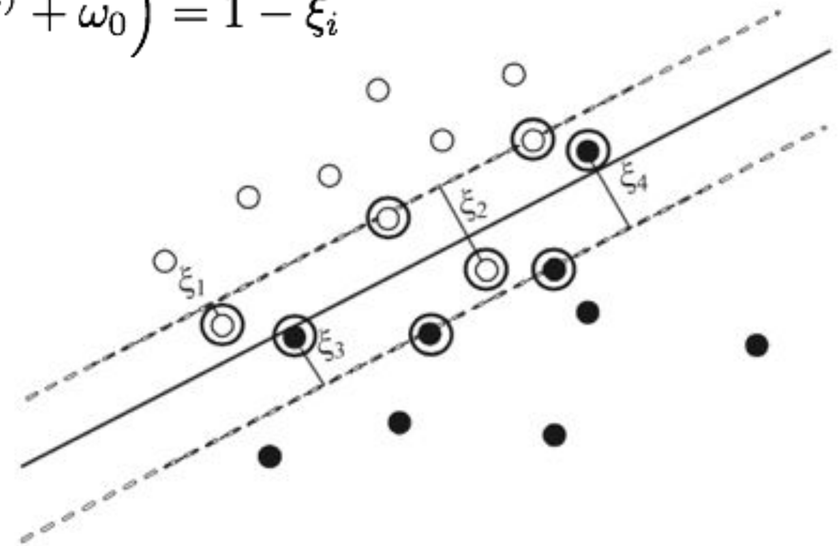


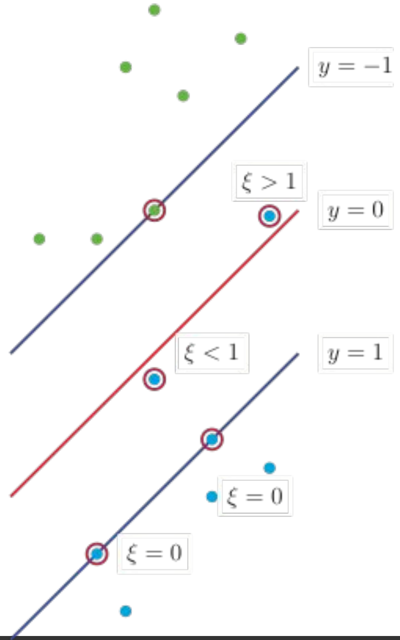
Caso linealmente no separable

- Voy a permitir errores: muestras **dentro del margen o mal clasificadas**
- Exclusivamente esas muestras les asigno un error (*slack* variable)

$$y^{(i)}\left(\boldsymbol{\omega}^T\mathbf{x}^{(i)}+\omega_0\right)=1-\xi_i$$

con $\xi_i \geq 0$





Caso linealmente no separable

- … pero penalizo los errores, con un coste C, ¿os suena?

$$\min_{\boldsymbol{\omega},\omega_0,\xi_i} \frac{1}{2}||\boldsymbol{\omega}||_2^2 + C\sum_{i=1}^{N}\xi_i$$

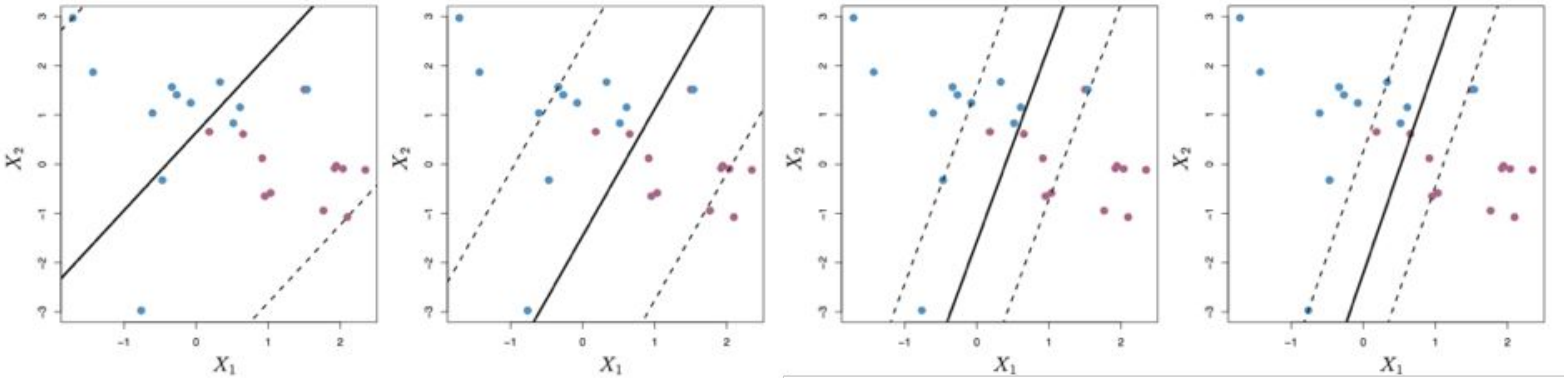
$$\text{s.to } y^{(i)}\left(\boldsymbol{\omega}^T\mathbf{x}^{(i)} + \omega_0\right) \geq 1 - \xi_i, i = 1,\ldots,N$$

$$\xi_i \geq 0.$$

- … regularización

Parámetro de regularización C

- C: cota superior al número de errores
- Compromiso entre margen y errores en la solución



- Si C elevado, margen estrecho, más peso a los errores. Alta complejidad
- Si C pequeño, margen ancho, menos peso a los errores. Baja complejidad

Caso linealmente no separable

- La solución no cambia con respecto al caso anterior

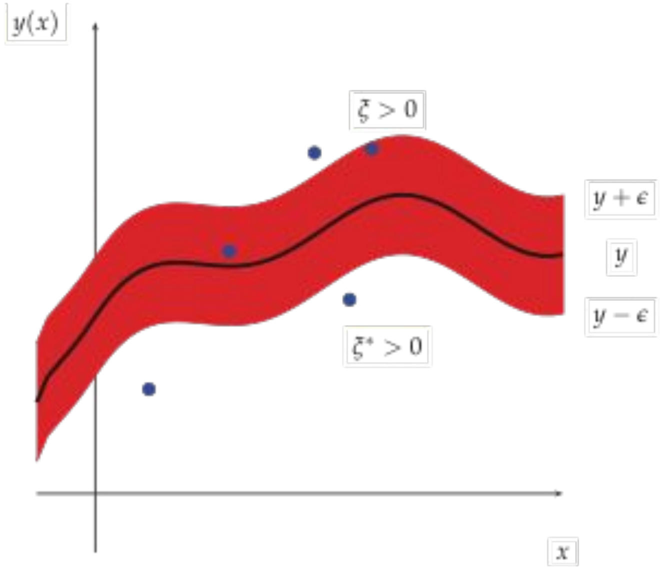
$$\hat{\boldsymbol{\omega}} = \sum_{i=1}^{N} \alpha_i y^{(i)} \mathbf{x}^{(i)}$$

$$\hat{y} = f(\mathbf{x}) = \text{sign}\left(\hat{\boldsymbol{\omega}}^T \mathbf{x} + \hat{\omega}_0\right) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y^{(i)} \mathbf{x}^T \mathbf{x}^{(i)} + \hat{\omega}_0\right)$$

Índice

SVR: intuición

- Buscar el hiperplano que mejor se ajuste a los datos y permita un tolerancia a los errores
  - En otras palabras, regresión lineal, con restricciones



SVR: formulación

- Queremos que nuestra solución sea de la forma

$$y = f(\mathbf{x}) = \boldsymbol{\omega}^T \mathbf{x} + \omega_0$$

y permitir errores dentro del “margen”: $[y - \epsilon, y + \epsilon]$, así que el funcional a minimizar es similar al problema de clasificación

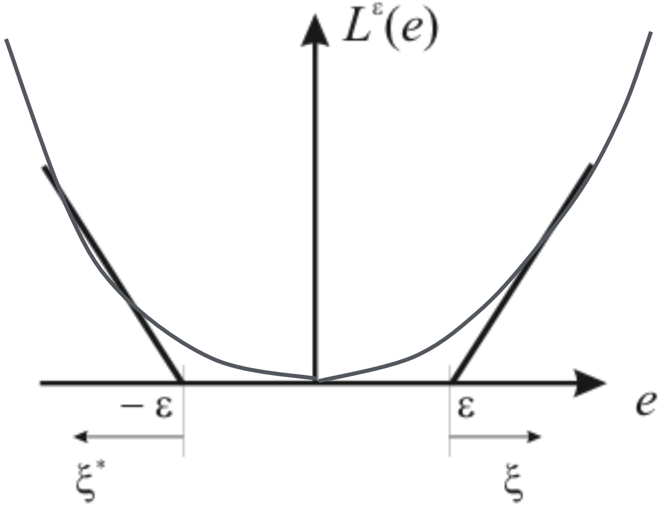
$$\min_{\boldsymbol{\omega}, \omega_0} \frac{1}{2} ||\boldsymbol{\omega}||_2^2$$

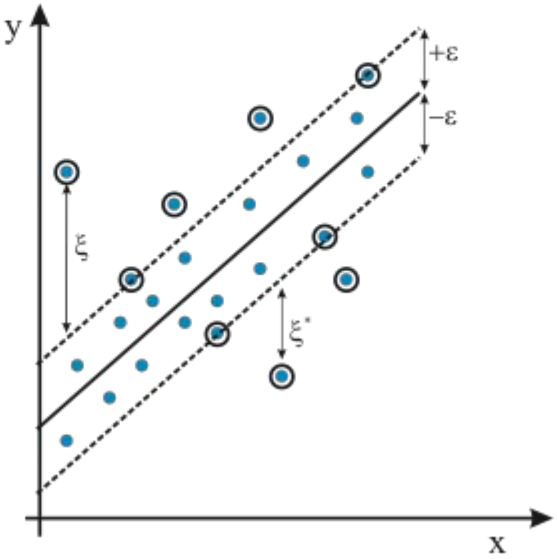
$$\text{s.to } y^{(i)} - \left(\boldsymbol{\omega}^T \mathbf{x}^{(i)} + \omega_0\right) \leq \epsilon$$

$$-y^{(i)} + \left(\boldsymbol{\omega}^T \mathbf{x}^{(i)} + \omega_0\right) \leq \epsilon$$

SVR: formulación

- pero, ¿qué hago con las muestras que caen fueran del margen?
- Las penalizo





SVR: formulación

- Regresión lineal, con función de coste e-insensible

$$\min_{\boldsymbol{\omega},\omega_0,\xi_i,\xi_i^*} \frac{1}{2}||\boldsymbol{\omega}||_2^2 + C\sum_{i=1}^{N} L^{\epsilon}(y^{(i)} - \hat{y}^{(i)}) = \min_{\boldsymbol{\omega},\omega_0,\xi_i,\xi_i^*} \frac{1}{2}||\boldsymbol{\omega}||_2^2 + C\sum_{i=1}^{N} (\xi_i + \xi_i^*)$$

sujeto a

$$y^{(i)} - \left(\boldsymbol{\omega}^T \mathbf{x}^{(i)} + \omega_0\right) \leq \epsilon + \xi_i$$

$$-y^{(i)} + \left(\boldsymbol{\omega}^T \mathbf{x}^{(i)} + \omega_0\right) \leq \epsilon + \xi_i^*$$

$$\xi_i, \xi_i \geq 0$$

SVR: solución

- Coeficientes: $\hat{\boldsymbol{\omega}} = \sum_{i=1}^{N} \alpha_i \mathbf{x}^{(i)}$
  - Combinación lineal de las muestras de entrenamiento
- Regresión: $\hat{y} = f(\mathbf{x}) = \hat{\boldsymbol{\omega}}^T \mathbf{x} + \hat{\omega}_0 = \sum_{i=1}^{N} \alpha_i \mathbf{x}^T \mathbf{x}^{(i)} + \hat{\omega}_0$
  - Producto escalar entre las muestras de entrenamiento y la muestra sobre la que quiero realizar la predicción

Índice

Recursive Feature Elimination

- Método *wrapper*
- Originalmente propuesto para SVM, analizando los coeficientes del modelo

¿Entiendes el algoritmo?

- En sklearn, extendido a otros algoritmos con indicadores de relevancia, como coeficientes o importancia de variables
    - Regresión lineal, logística, Ridge, Lasso
    - Algoritmos basados en árboles

Referencias

- Felipe Alonso-Atienza, Tesis Doctoral (uc3m)

- Machine Learning, a probabilistic perspective
    - Capítulo 14

- Support Vector Machines, Chris Williams, Universidad de Edimburgo

¿Preguntas?