Máquinas de Vectores Soporte (SVM)

Alfredo Cuesta Infante

E. T. S. Ingeniería Informática
Universidad Rey Juan Carlos

Master Univ. en Visión Artificial
Reconocimiento de Patrones

Contenidos

Contenidos

Intuición

¿Qué clasificador es mejor?

- Los clasificadores C_1, C_2 y C_3 clasifican PERO...

  ¿qué ocurrirá cuando llegue un nuevo ejemplo próximo a la superficie de decisión?
- Cuanto mayor sea el **margen** menor es el **riesgo** de FPs y FNs

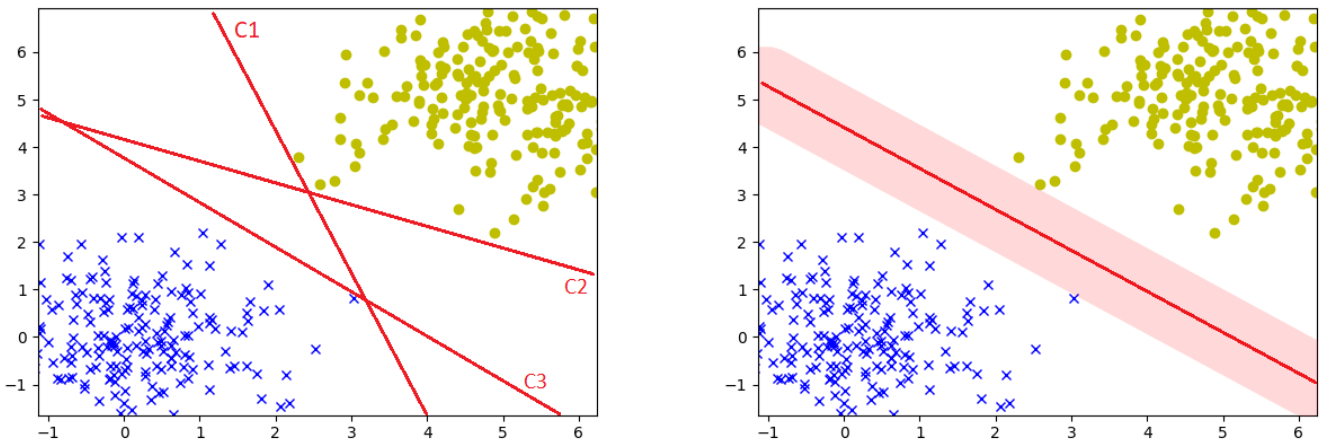

Figura: (Der.) Tres clasificadores que no cometen ningún error en el conjunto de entrenamiento pero seguramente funcionen mal con datos nuevos. (Izq.) Este clasificador que maximiza el margen entre ambas clases. [Fuente: Original de A. Cuesta]

Intuición

"La justicia inflexible es frecuentemente la injusticia más grande"

Publio Terencio Africano (194 a.C. – 159 a.C.)

- Si permitimos algunos ejemplos dentro de los márgenes, entonces podemos aumentarlos y encontrar clasificadores con mayor potencial generalizador

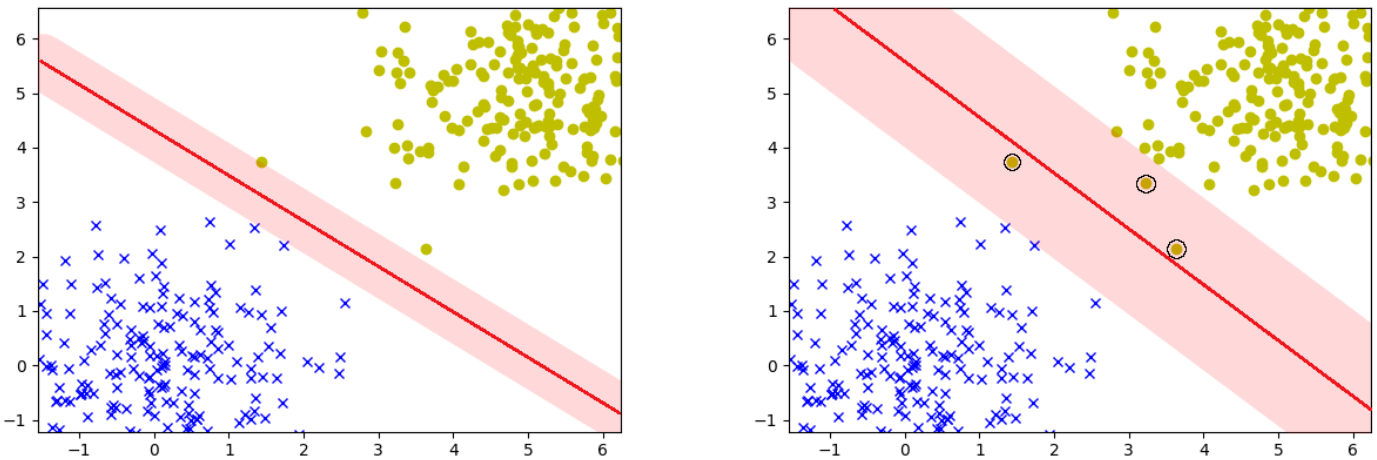

Figura: (Der.) Márgenes *hard*. (Izq.) márgenes *soft* [Fuente: Original de A. Cuesta]

Intuición

Clasificación No-Lineal

- Con SVM **no vamos a necesitar** transformar el vector de características a otro espacio.
- En su lugar *alguien* de ese espacio vendrá al nuestro con el resultado de las operaciones.

Contenidos

Clasificación lineal con márgenes *hard*

Notación

- Ahora los vectores $\mathbf{w}$ y $\mathbf{x}$ no estaran 'expandidos', es decir:

$$\mathbf{w} = (w_1, w_2, \ldots, w_m)^T \quad , \quad \mathbf{x} = (x_1, x_2, \ldots, w_m)^T \ .$$

- Dado que hemos sacado el término independiente del vector de pesos, la expresion del discriminante es ahora:

$$\hat{t} = \begin{cases} +1 & \text{si} \quad \mathbf{w}^T\mathbf{x} + b > 0 \\ -1 & \text{si} \quad \mathbf{w}^T\mathbf{x} + b < 0 \end{cases} \tag{1}$$

  donde b es el término independiente y, al igual que en la semana pasada, las etiquetas son $+1$ y -1 porque serán 'matemáticamente convenientes'.

Recordar también que en 2D

- Una recta se puede definir con un punto y un vector director $\mathbf{v}$
- Un hiperplano se puede definir con un punto y un vector normal $\mathbf{w}$
- El vector director y el normal son ortogonales: $\mathbf{v}^T\mathbf{w} = \mathbf{w}^T\mathbf{v} = 0$
- La ecuación implícita de una recta: $w_1x_1 + w_2x_2 + b = \mathbf{w}^T\mathbf{x} = 0$

Clasificación lineal con márgenes *hard*

Planteamiento geométrico del problema

1. Calcular la distancia de la superficie de decisión al origen

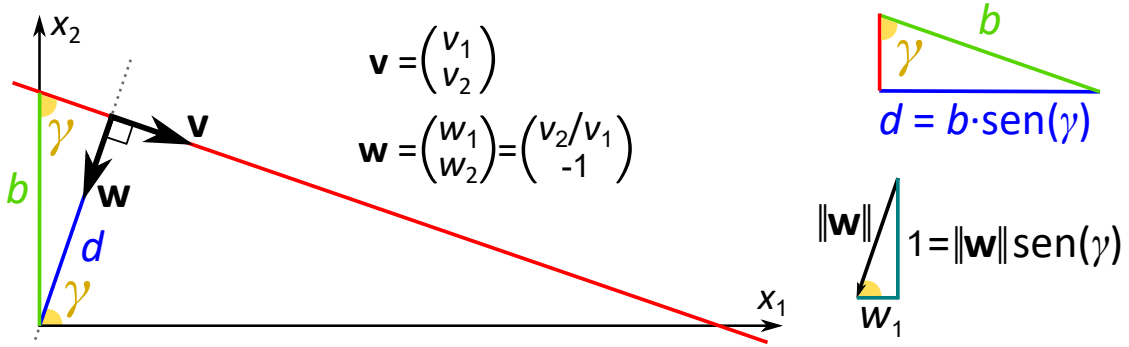

Figura: El clasificador lineal viene dado por un vector director $\mathbf{v}$ y un punto de corte b. El vector $\mathbf{w}$ es perpendicular a $\mathbf{v}$ y sirve para calcular la distancia de la recta al origen. Combinando todo podemos representar d en función de $\mathbf{w}$ y b. [Fuente: Original de A. Cuesta]

En definitiva, la distancia del clasificador al origen es

$$d = \frac{b}{\|\mathbf{w}\|}$$

Clasificación lineal con márgenes *hard*

Planteamiento geométrico del problema

2. Imponer el margen *hard*

Todos los ejemplos de cada clase deben quedar por encima (o por debajo) de los márgenes.

Para ello situamos dos clasificadores H^+ y H^-

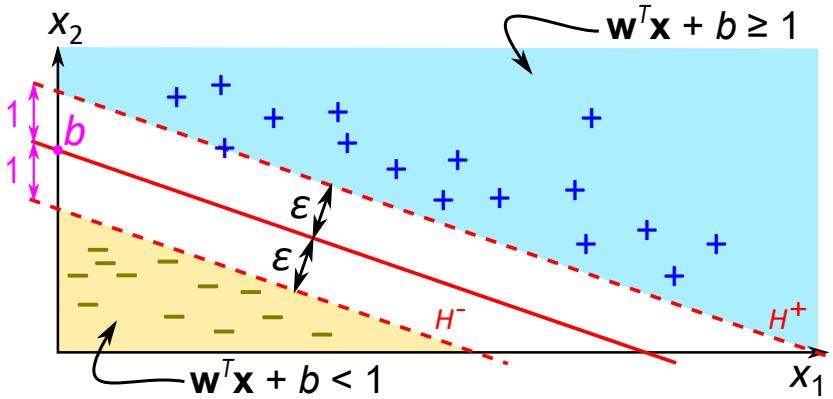

Figura: Para crear el margen se añaden dos planos paralelos entre las clases, sin ningún ejemplo entre ellos. [Fuente: Original de A. Cuesta]

En definitiva

$$2\varepsilon = (d+\varepsilon)-(d-\varepsilon) = \frac{b+1-(b-1)}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}, \quad \text{luego} \quad \varepsilon = \frac{1}{\|\mathbf{w}\|}$$

Clasificación lineal con márgenes *hard*

Planteamiento geométrico del problema

3. En definitiva:

- Cuanto más pequeña sea la norma del vector de pesos $\|\mathbf{w}\|$, mayor será el márgen ε; es decir, buscamos

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \|\mathbf{w}\|$$

- Las dos restricciones impuestas se pueden escribir en una sola:

$$\left.\begin{array}{lll} \text{Si} & \hat{t}^{(i)} = +1 & \text{entonces} \quad \mathbf{w}^{(i)T}\mathbf{x} + b \geq +1 \\ \text{Si} & \hat{t}^{(i)} = -1 & \text{entonces} \quad \mathbf{w}^{(i)T}\mathbf{x} + b < -1 \end{array}\right\} =$$

$$\left(\hat{t}^{(i)}\right)\left(\mathbf{w}^{(i)T}\mathbf{x} + b\right) \geq 1.$$

- Por tanto el problema de optimización es:

$$\begin{array}{ll} \text{minimizar} & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{sujeto a} & \left(\hat{t}^{(i)}\right)\left(\mathbf{w}^{(i)T}\mathbf{x} + b\right) \geq 1, \\ & \text{para } i = 1, 2, \ldots, m. \end{array}$$

Recordar que $\|\mathbf{w}\| = \mathbf{w}^T\mathbf{w}$

Clasificación lineal con márgenes *soft*

Introducción de variables *Slack*

- Permitimos que haya ejemplos en el margen utilizando un peso ζ que incorporamos a la función de coste

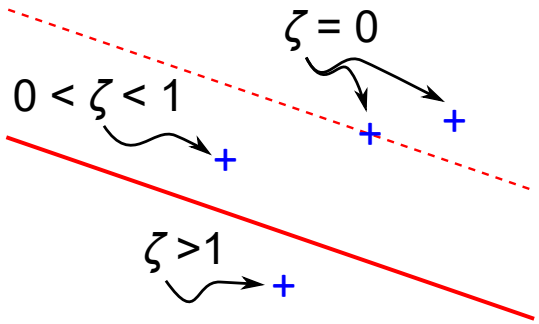

Figura: *Slack variables* para ejemplos que violan la restricción. [Fuente: Original de A. Cuesta]

1. Compromiso
   - Queremos ampliar el margen
   - Queremos minimizar el número de ejemplos que hay dentro del margen

Para gestionarlo se añade el hiperparámetro C:

$$\min_{\mathbf{w},b,\zeta} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m}\zeta^{(i)}$$

Clasificación lineal con márgenes *soft*

Introducción de variables *Slack*

2. Modificación de las restricciones

$$\left(\hat{t}^{(i)}\right)\left(\mathbf{w}^{(i)T}\mathbf{x}+b\right) \geq 1-\zeta^{(i)}$$

3. En definitiva

$$\begin{aligned} &\underset{\mathbf{w},b,\zeta}{\text{minimizar}} \quad \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m}\zeta^{(i)}, \\ &\text{sujeto a} \quad \begin{cases} \left(\hat{t}^{(i)}\right)\left(\mathbf{w}^{(i)T}\mathbf{x}+b\right) \geq 1-\zeta^{(i)} \\ \zeta^{(i)} \geq 0 \\ i = 1, 2, \ldots, m. \end{cases} \end{aligned} \tag{2}$$

- C es una medida de cuanto queremos evitar que haya ejemplos en el márgen. Cuanto más pequeño, menos queremos evitarlo, es decir más *permisivos* somos.
- Problema de optimización cuadrática con restricciones !!
- Solución mediante **multiplicadores de Lagrange**

El problema dual

Multiplicadores de Lagrange

- Técnica mediante la cual el problema se formula en términos de unas nuevas variables $\alpha^{(i)}$ con restricciones más sencillas.

$$\begin{aligned} \underset{\alpha}{\text{minimizar}} \quad & \tfrac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha^{(i)}\alpha^{(j)}t^{(i)}t^{(j)}\mathbf{x}^{(i)T}\mathbf{x}^{(j)} - \sum_{i=1}^{m}\alpha^{(i)} \\ \text{sujeto a} \quad & \alpha^{(i)} > 0 \\ & i = 1, 2, \ldots, m. \end{aligned} \tag{3}$$

  - Cada ejemplo $\mathbf{x}^{(i)}$ tiene un multiplicador de Lagrange $\alpha^{(i)}$.
  - El problema se puede resolver mediante métodos computacionales.
  - Una vez resuelto, algunos (muchos) multiplicadores se anulan. Los ejemplos asociados a los multiplicadores distintos de cero son los **vectores soporte**.
  - El vector de pesos óptimo es:

$$\mathbf{w}^* = \sum_{i=1}^{m}\alpha^{*(i)}t^{(i)}\mathbf{x}^{(i)} \;;\; b^* = \frac{1}{n_s}\sum_{i:\alpha^{*(i)}>0}\left((1-t^{(i)})(\mathbf{w}^{*T}\mathbf{x}^{(i)})\right) \tag{4}$$

    n_s es el número de vectores soporte, o sea el número de $\alpha^{*(i)} > 0$.

Contenidos

Clasificación no lineal

El truco del Kernel

- En la formulación dual, la función de coste depende

  del producto escalar de cada dos vectores transformados

- Pero si en vez de tener la transformación ϕ tuvieramos una función K (Kernel) que nos devuelva el resultado de dicho producto escalar en el espacio al que nos lleva ϕ, **¿podríamos resolver el problema?**

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha^{(i)} \alpha^{(j)} t^{(i)} t^{(j)} \phi(\mathbf{x}^{(i)})^T \phi(\mathbf{x}^{(j)}) - \sum_{i=1}^{m} \alpha^{(i)}$$

$$\Downarrow$$

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha^{(i)} \alpha^{(j)} t^{(i)} t^{(j)} K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) - \sum_{i=1}^{m} \alpha^{(i)}$$

- Aparentemente **NO** porque la solución $\{\mathbf{w}^*, b^*\}$ depende de $\phi(\mathbf{x}^{(\mathbf{i})})$

$$\mathbf{w}^* = \sum_{i=1}^{m} \alpha^{*(i)} t^{(i)} \phi(\mathbf{x}^{(i)}) \; ; \; b^* = \frac{1}{n_s} \sum_{i:\alpha^{*(i)}>0} \left((1 - t^{(i)})(\mathbf{w}^{(*)T} \phi(\mathbf{x}^{(i)})) \right)$$

- Pero ¿para qué queremos saber $\{\mathbf{w}^*, b^*\}$?

  ¿No es para clasificar nuevos ejemplos?

  $\rightarrow$ Vamos a ver qué aspecto tiene la función discriminante

Clasificación no lineal

El truco del Kernel

Sea $\mathbf{z}$ un nuevo ejemplo. Entonces su etiqueta estimada es:

$$\hat{t} = \mathbf{w}^{*T}\phi(\mathbf{z}) + b^* = \left(\sum_{i=1}^{m} \alpha^{*(i)} t^{(i)} \phi(\mathbf{x}^{(i)})\right)^T \phi(\mathbf{z}) + b^*$$

$$= \sum_{i:\alpha^{*(i)}>0} \alpha^{*(i)} t^{(i)} \left(\phi(\mathbf{x}^{(i)})^T \phi(\mathbf{z})\right) + b^*$$

$$= \sum_{i:\alpha^{*(i)}>0} \alpha^{*(i)} t^{(i)} K(\mathbf{x}^{(i)}, \mathbf{z}) + b^*$$

- **¡¡ Podemos clasificar ejemplos sin necesitar ϕ !!**

  Necesitamos:

  - Resolver el problema dual planteado con un Kernel para obtener los multiplicadores no nulos
  - Utilizar la expresión de arriba para clasificar nuevos ejemplos, donde el término independiente óptimo b^* es

$$b^* = \frac{1}{n_s} \sum_{i:\alpha^{*(i)}>0} \left(\left(1 - t^{(i)}\right) \sum_{j:\alpha^{*(j)}>0} \left(\alpha^{*(j)} t^{(j)} K(\mathbf{x}^{(j)}, \mathbf{z})\right)\right)$$

Algunos Kernels importantes

Nombre	$K(\mathbf{a}, \mathbf{b}) =$	Hiperparámetros
Lineal	$\mathbf{a}^T\mathbf{b}$	— — —
Polinomial	$\left(\gamma\mathbf{a}^T\mathbf{b} + r\right)^d$	γ, r, d
Sigmoide	$\tanh\left(\gamma\mathbf{a}^T\mathbf{b} + r\right)$	γ, r
RBF Gaussiano	$\exp\left(-\gamma\|\mathbf{a} - \mathbf{b}\|^2\right)$	γ

RBF = Radial Basis Function