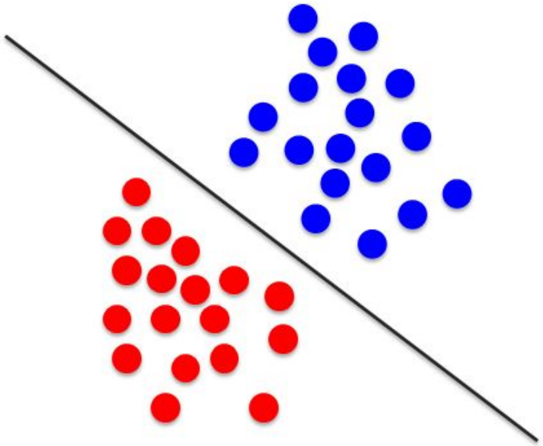
Ayudantía 8: SVM

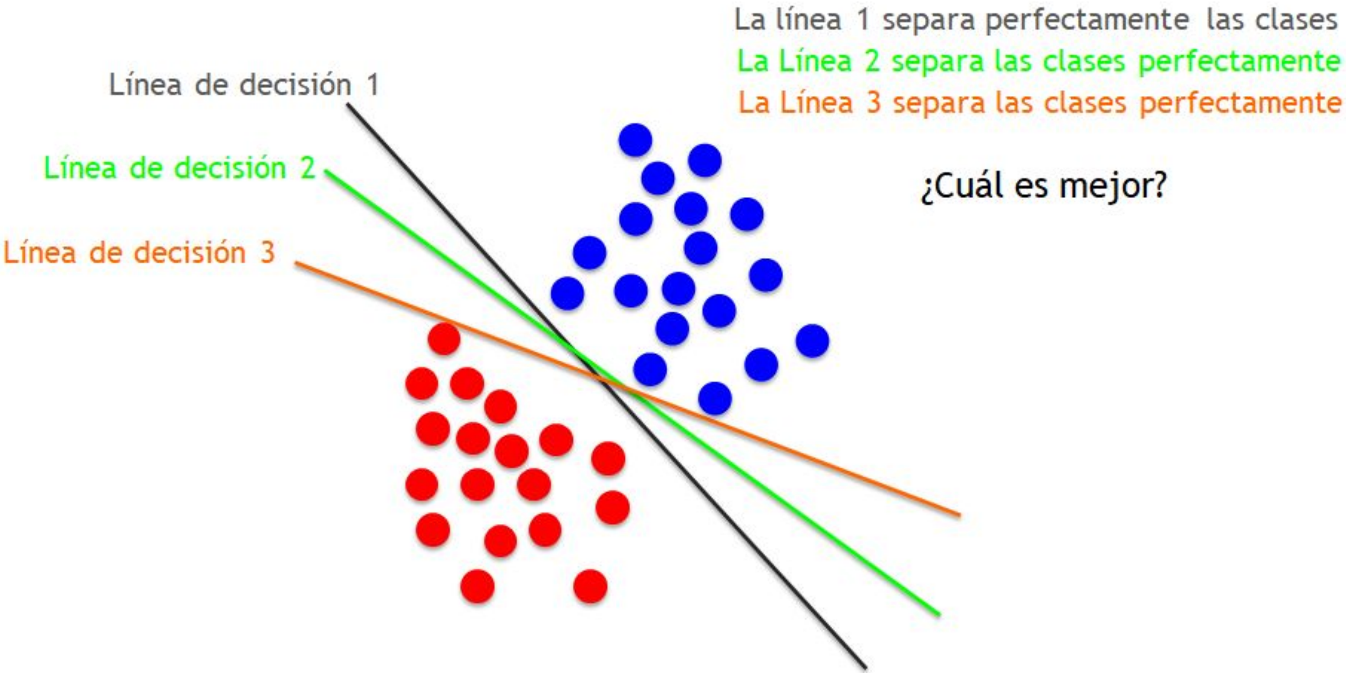
José Manuel Opaso
Bernardita Morris

Support Vector Machines (SVM)

Técnica de aprendizaje supervisado generalmente usada para la clasificación de datos basada en sus rótulos, dividiendo las distintas clases a los que éstos pertenecen a través de un hiperplano óptimo.

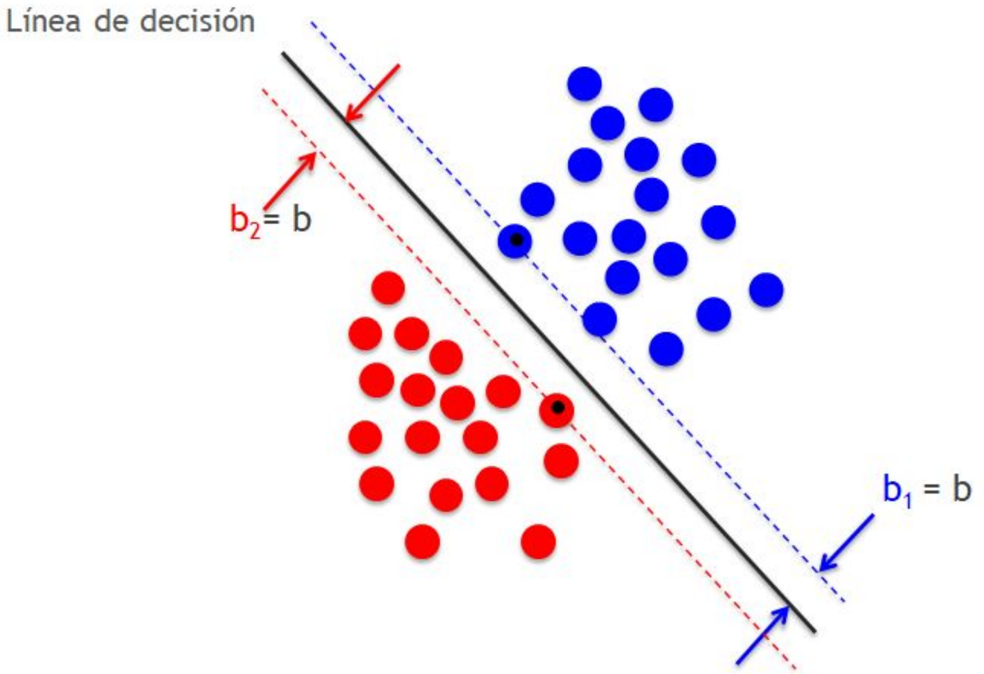


¿Cómo encontramos el hiperplano óptimo?



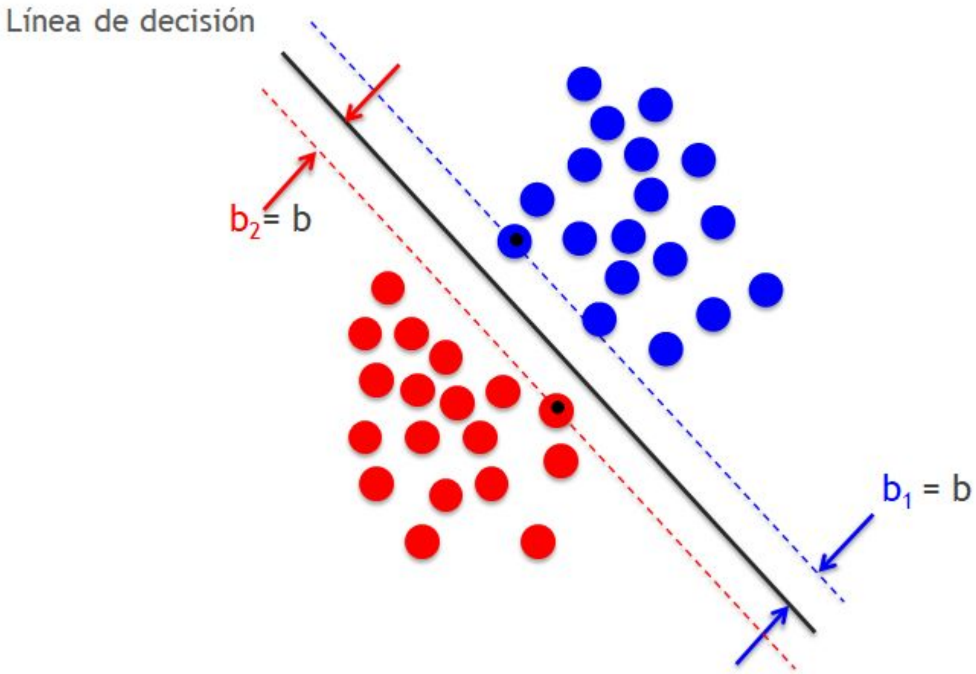
SVM: dos clases

- Margen (b): distancia perpendicular entre el hiperplano y los registros más cercanos a cada uno de sus lados.



SVM: dos clases

- Vectores de soporte: conjunto de datos de entrenamiento más cercanos a la superficie de decisión. Son los más difíciles de clasificar



Sobre el algoritmo

Condiciones:

1. Hiperplano g(x) debe **clasificar correctamente** los registros del set de entrenamiento
2. Hiperplano g(x) debe **maximizar el margen** a registros más cercanos a la superficie de decisión

Estas condiciones se resumen en el problema de optimización:

$$\underset{w,c}{\operatorname{argmax}} \frac{1}{||w||}$$

$$\text{sujeto a: } z_k(w\, x_k + c) \geq 1, k = 1 \ldots n.$$

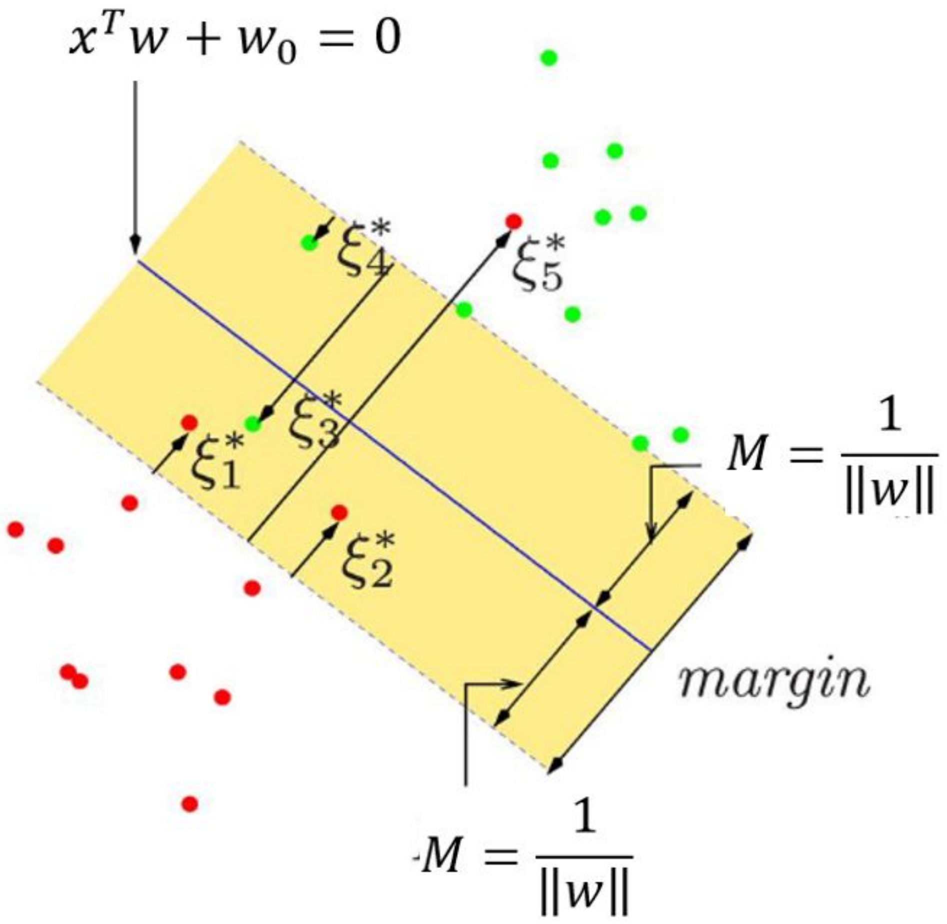
Solución al problema

$$g(x) = \sum_{i=1}^{n} \alpha_i z_i < x_i^T, x > +w_0$$

- Para clasificar un vector nuevo (x), hay que calcular su similaridad respecto a los vectores de soporte

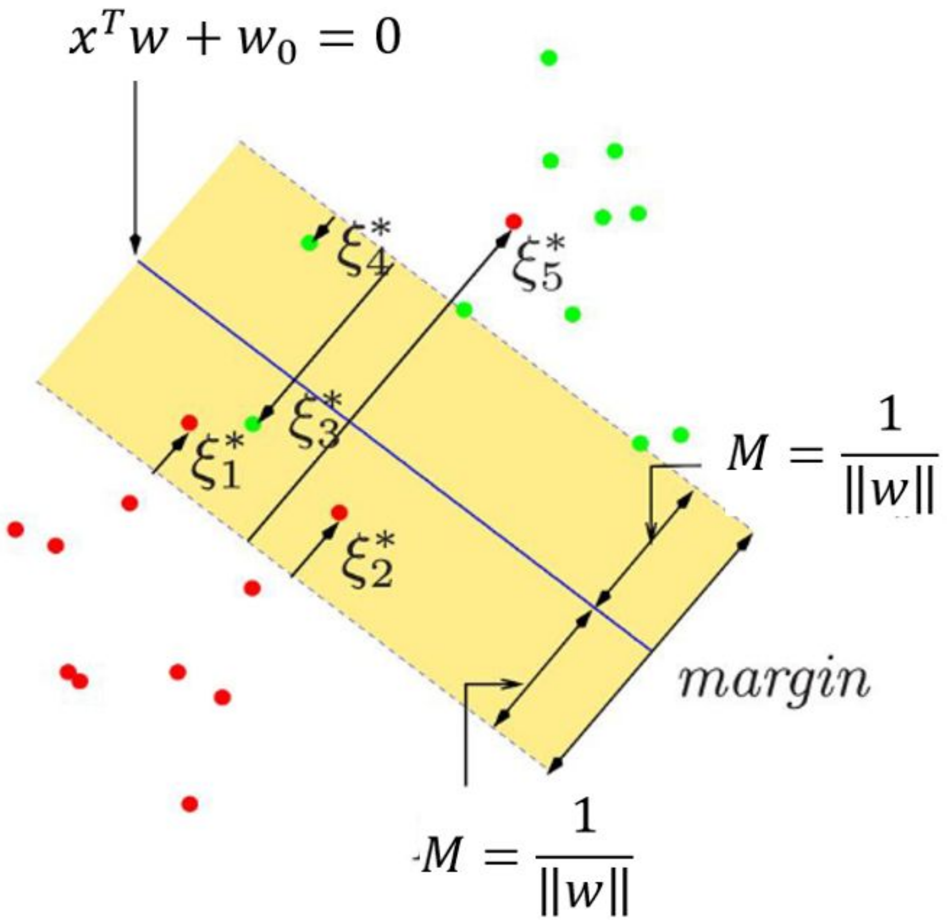
No linealidad

Problema relajado



Problema relajado

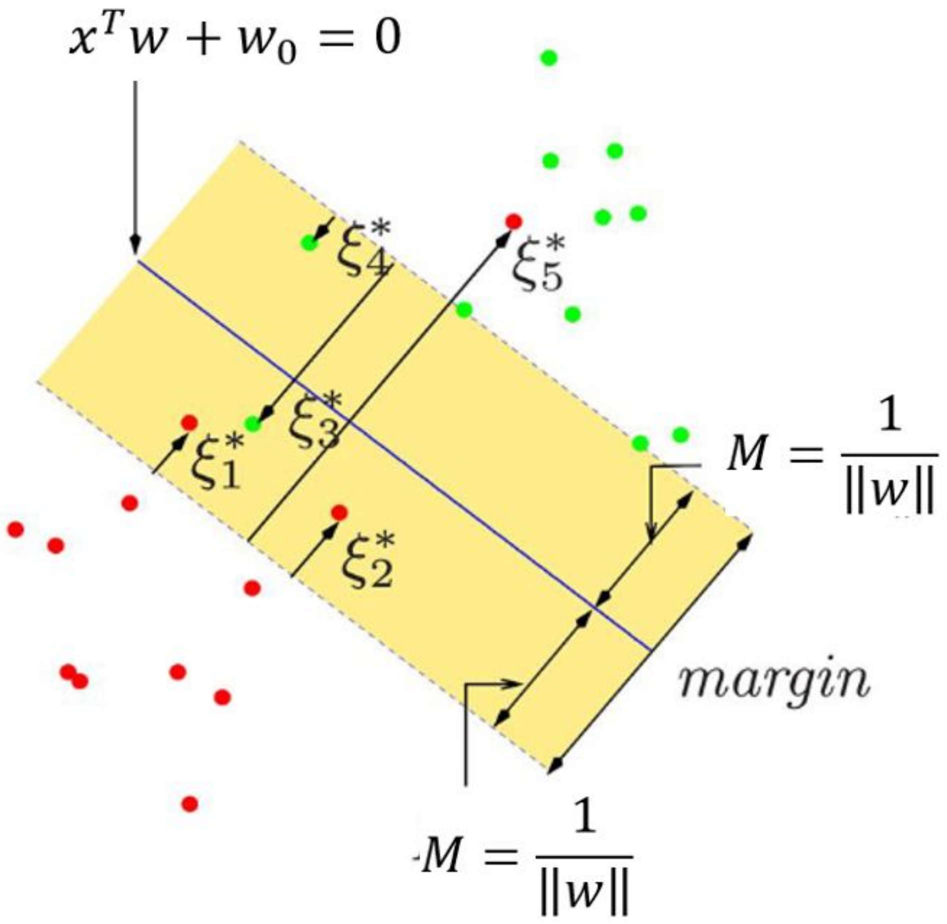
$$\frac{1}{\|w\|} y_i\left(x_i^T w + w_0\right) > M - \xi_i^*$$



Problema relajado

$$\frac{1}{\|w\|}y_i(x_i^T w + w_0) > M - \xi_i^*$$

$$\frac{1}{\|w\|}y_i(x_i^T w + w_0) > M(1 - \xi_i)$$

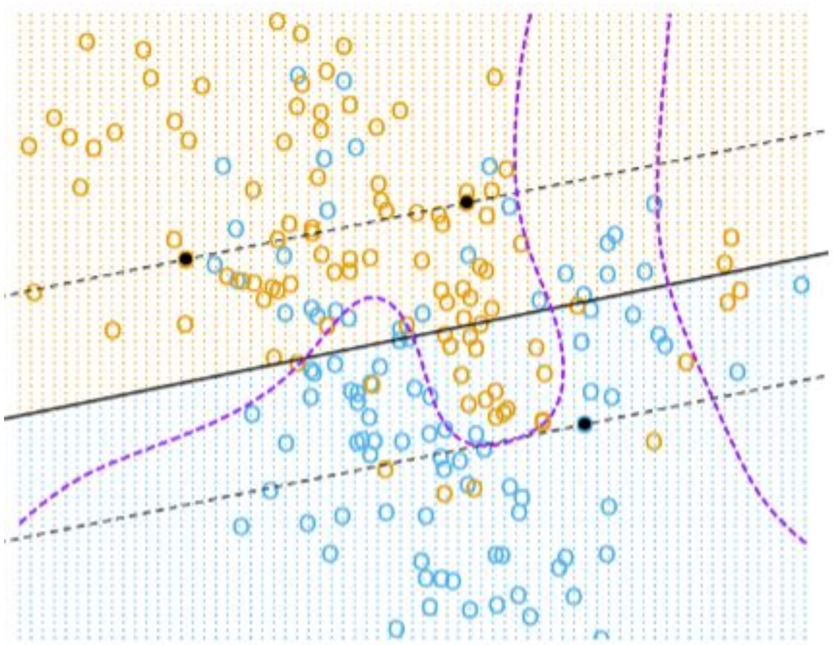


$$\min \ \|w\| \ \ subject\ to \left\{ y_i\left(x_i^T w + w_0\right) \geq 1 - \xi_i, \ \xi_i \geq 0 \right\}$$

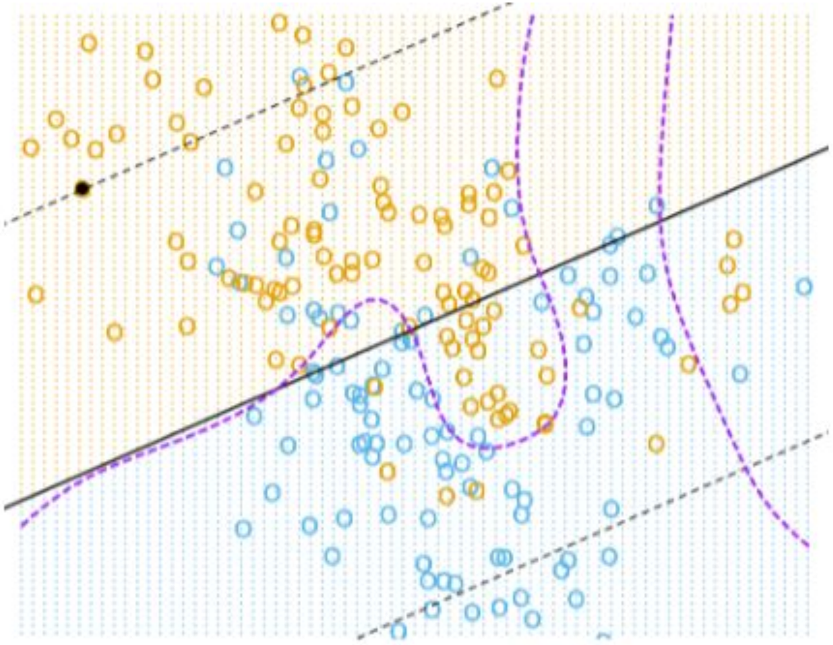
$$\min \ \|w\| \ subject\ to \left\{y_i\left(x_i^T w + w_0\right) \geq 1 - \xi_i, \ \xi_i \geq 0\right\}$$

$$\min \frac{1}{2}\|w\|^2 + C \sum_{i}^{N} \xi_i$$

$$subject\ to \ \ \xi_i \geq 0, y_i\left(x_i^T w + w_0\right) > 1 - \xi_i$$

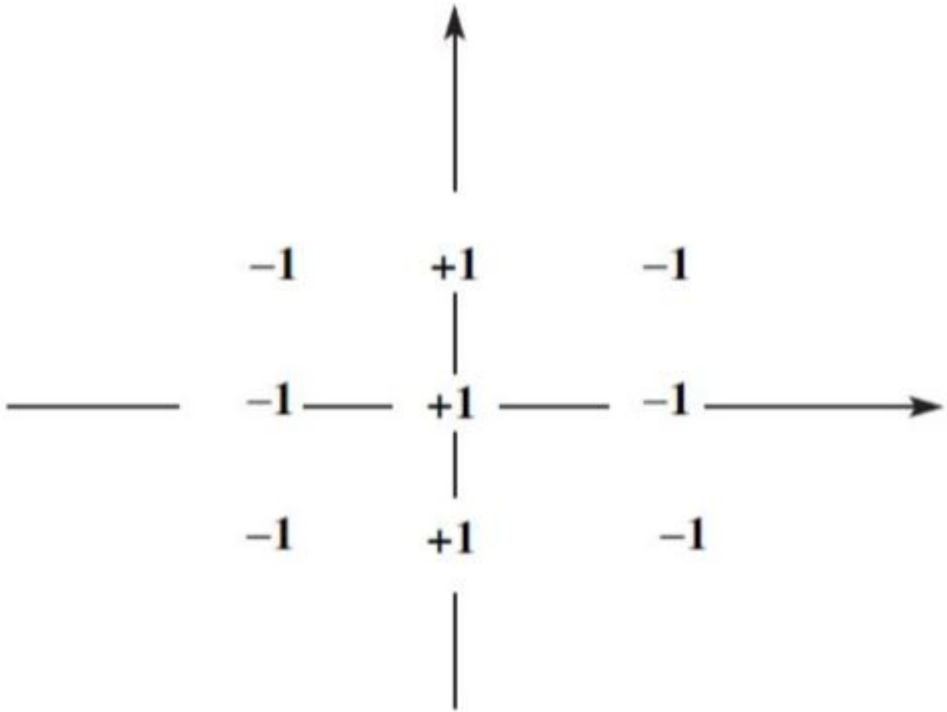


C = 1000

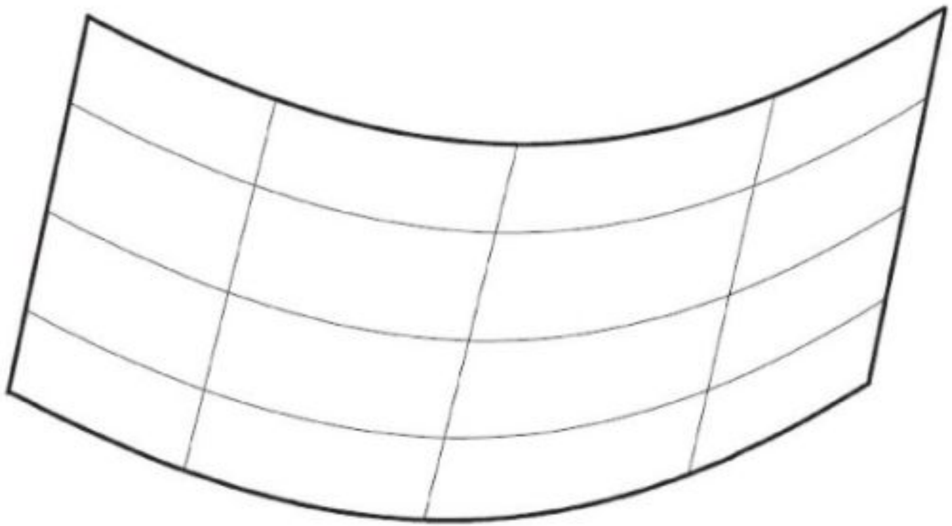


C = 0.01

Super bien, pero sigue siendo un clasificador lineal ..

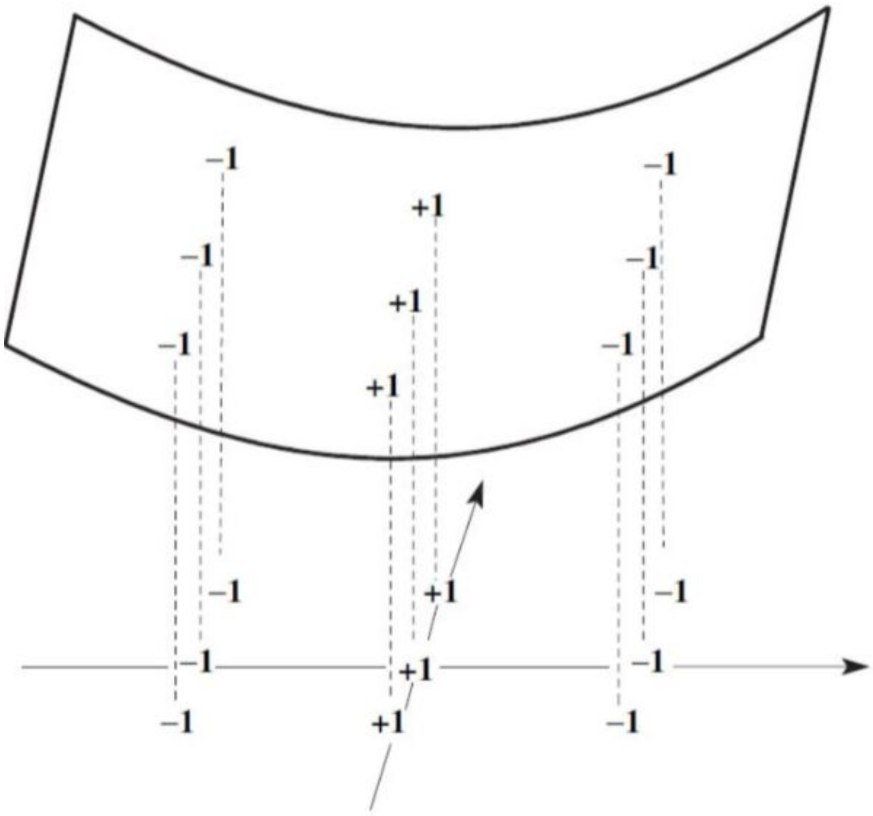


Transformación de espacio de características

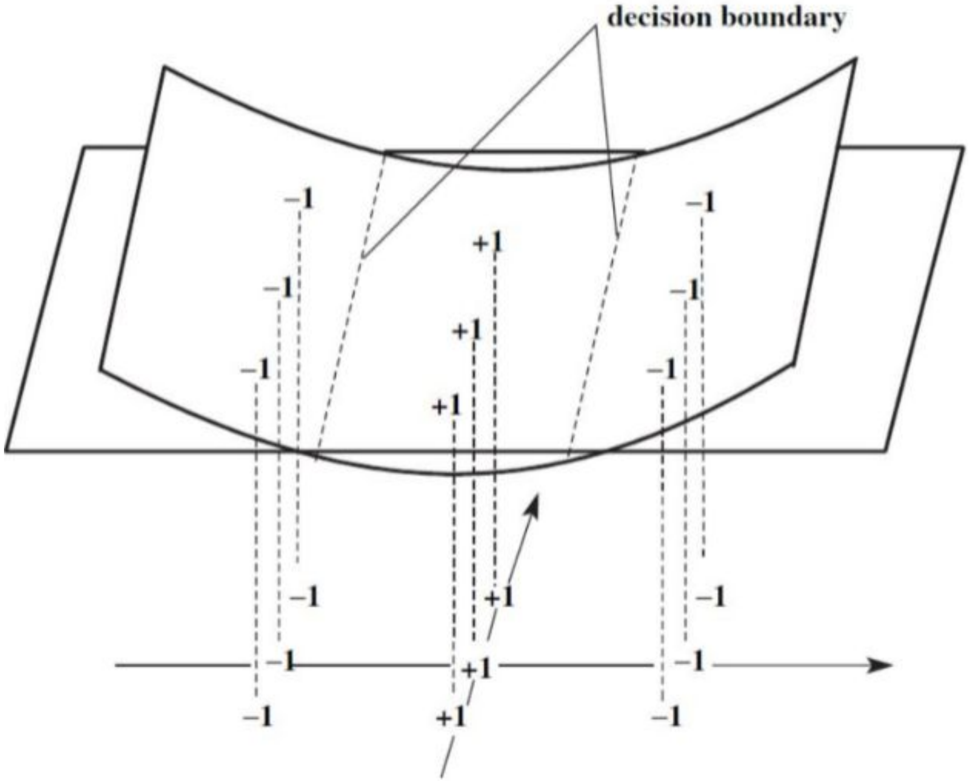


Espacio de características $f(x_1, x_2) = x_1^2$

Transformación espacio características



Transformación espacio características



Cómo generalizamos esto..

- Ir probando dimensión a dimensión no es una buena opción
- Dimensionalidad muy grande difícil de manejar
- ¿Estamos destinados a vivir en espacios de superficies de decisión lineales ?

Kernel

- Supongamos que tenemos una función ϕ, que lleva un vector x a un nuevo espacio de dimensionalidad arbitraria (potencialmente infinita).
- Imaginemos que para cada par de vectores, ϕ(xi) y ϕ(xj), calculamos su producto punto y lo guardamos en la posición i, j de una matriz M.
- Acá viene la magia: Si M es semidefinida positiva, entonces define un **kernel** válido.
- Qué significa esto: Que existe una función K(xi, xj), que toma dos vectores xi y xj y retorna el producto punto de h(xi) y h(xj), sin necesidad de conocer la función h(x).

Incorporarlo en SVM: Kernel Trick

Incorporarlo en SVM: Kernel Trick

Recordemos la solución del problema de optimización por KKT:

$$g(x) = \sum_{i=1}^{n} \alpha_i z_i < x_i^T, x > + w_0$$

Incorporarlo en SVM: Kernel Trick

Recordemos la solución del problema de optimización por KKT:

$$g(x) = \sum_{i=1}^{n} \alpha_i z_i < x_i^T, x > + w_0$$

$$g(x) = \sum_{i=1}^{n} \alpha_i z_i < \phi(x_i)^T, \phi(x) > + w_0$$

¿Cómo determinar el Kernel?

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + \beta_0$$

$$\begin{aligned} d\text{th-Degree polynomial: } & K(x, x') = (1 + \langle x, x' \rangle)^d, \\ \text{Radial basis: } & K(x, x') = \exp(-\gamma \|x - x'\|^2), \\ \text{Neural network: } & K(x, x') = \tanh(\kappa_1 \langle x, x' \rangle + \kappa_2). \end{aligned}$$

SVM Multiclase

Generalizando

- En vez de maximizar el margen absoluto, maximizamos el margen relativo entre distintos planos.
  - Plano asociado a una clase debe ser mejor que el resto de los planos.

$$y_i(w^T\phi(x_i) + \beta)$$

Generalizando

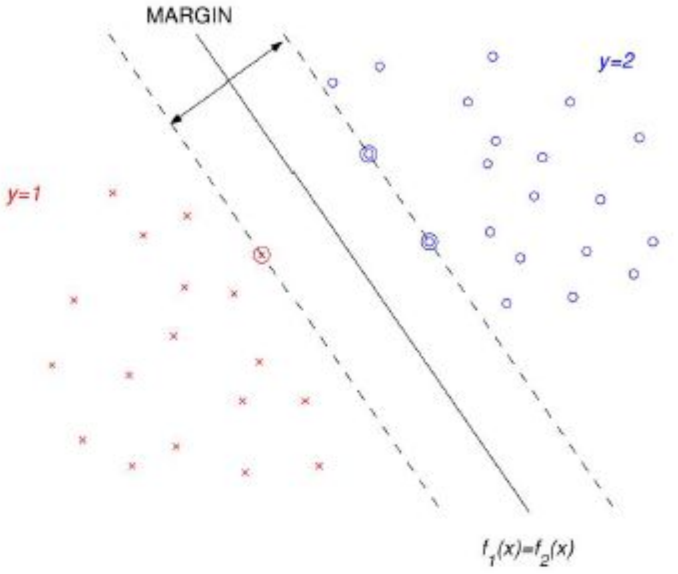
- En vez de maximizar el margen absoluto, maximizamos el margen relativo entre distintos planos.
    - Plano asociado a una clase debe ser mejor que el resto de los planos.

$$\cancel{y_i(w^T \cdots + \beta)}$$

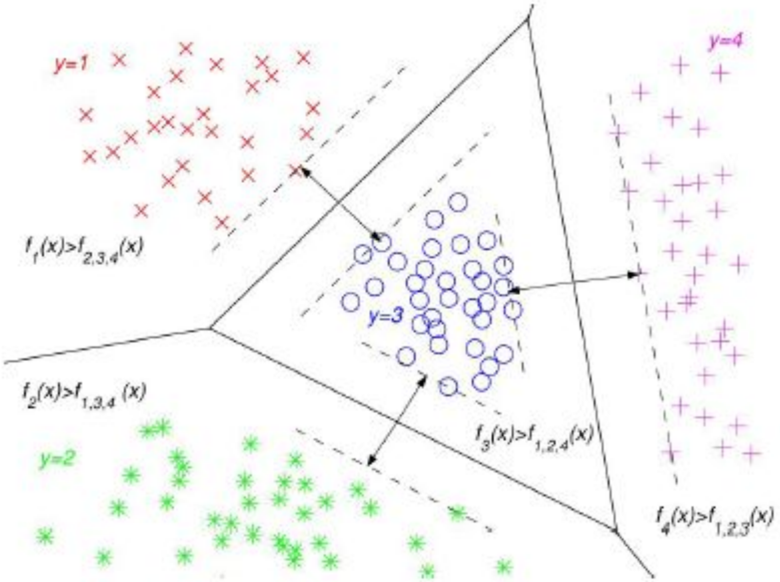
$$(w_{y_i}^T \phi(x_i) + \beta_{y_i}) - (w_k^T \phi(x_i) + \beta_k)$$
$$\forall (x_i, y_i) \in \text{Training Set (TS)},\ y_i \in \mathcal{Y}, k \in \mathcal{Y} \setminus y_i$$

Gráficamente

Two-class case



N-class case



Finalmente

- Deberíamos verificar nC-n restricciones.

- Finalmente, fijando el margen en 1:

$$\arg\min_{w_k,\beta_k} \frac{1}{2} \sum_{k \in \mathcal{Y}} ||w_k||^2$$

$$\text{sujeto a: } (w_{y_i}^T \phi(x_i) + \beta_{y_i}) - (w_k^T \phi(x_i) + \beta_k) > 1$$

$$i \in TS,\ k \in \mathcal{Y} \setminus y_i$$