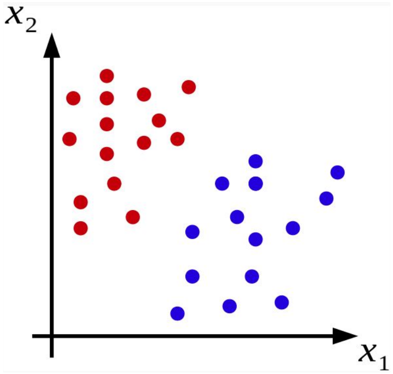
Aprendizaje Automático

Clasificadores basados en vectores soportes - Parte 1

2023-2Q

Support Vector Machines (SVM)

Es un algoritmo de **aprendizaje supervisado** utilizado tanto para problemas de **clasificación** como de **regresión**.



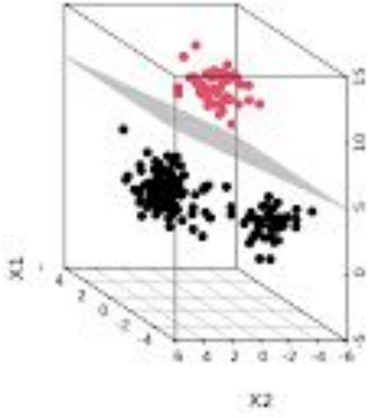
Su **principal objetivo** es encontrar el **hiperplano óptimo** que mejor **separa las diferentes clases**.

Concepto de hiperplano

Ecuación de un hiperplano

En un espacio **p-dimensional** un hiperplano está definido por

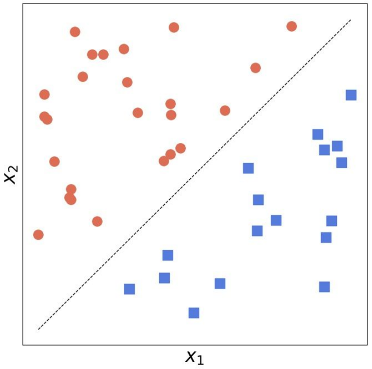
$$b_0 + b_1x_1 + b_2x_2 + .. + b_px_p = 0$$



Ecuación de un hiperplano en R2

En **R2**, la recta queda definida por $b_0 + b_1x_1 + b_2x_2 = 0$

- $x = (x_1, x_2)$ → sobre la **recta**
- $x' = (x'_1, x'_2)$ tal que $b_0 + b_1x'_1 + b_2x'_2 > 0$
- $x'' = (x''_1, x''_2)$ tal que $b_0 + b_1x''_1 + b_2x''_2 < 0$



Separabilidad lineal

Supongamos que tenemos un conjunto de **n ejemplos** de **p atributos x_i** y **clase y_i** ($1 \leq i \leq n$):

$$x_1 = (x_{1,1}, x_{1,2}, .., x_{1,p}), y_1$$
$$x_2 = (x_{2,1}, x_{2,2}, .., x_{2,p}), y_2$$
$$\ldots$$
$$x_n = (x_{n,1}, x_{n,2}, .., x_{n,p}), y_n$$

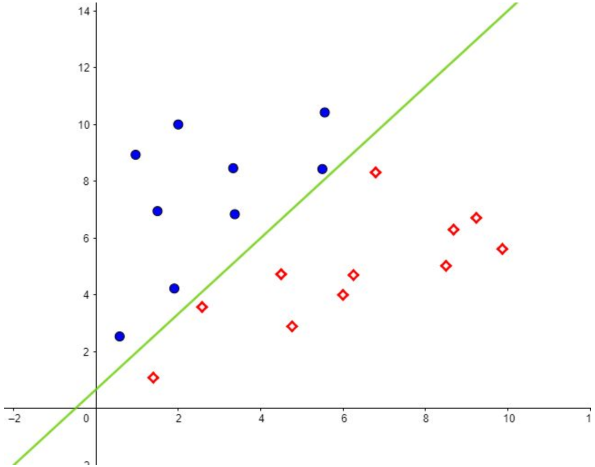
donde cada **y_i pertenece a {1, −1}**.

El objetivo es encontrar un clasificador que clasifique correctamente cada uno de estos ejemplos de acuerdo a su clase.

Separabilidad lineal

Hiperplano de Separación

Si obtenemos un hiperplano de tal forma que todos los ejemplos cuya clase es -1 queden de un lado del hiperplano y todos los ejemplos cuya clase es +1 queden del otro habremos alcanzado el objetivo.



Separabilidad lineal

Entonces ...

Dado x_i , i = 1, . . . n si ocurre que

$$b_0 + b_1x_{i,1} + b_2x_{i,2} + \cdots + b_px_{i,p} > 0 \text{ cuando } y_i = 1$$

y

$$b_0 + b_1x_{i,1} + b_2x_{i,2} + \cdots + b_px_{ip} < 0 \text{ cuando } y_i = -1$$

entonces podemos garantizar que $y_i(b_0 + b_1x_{i,1} + b_2x_{i,2} + \cdots + b_px_{i,p}) > 0$

Separabilidad lineal

$y_i (b_0 + b_1 x_{i,1} + b_2 x_{i,2}) > 0$

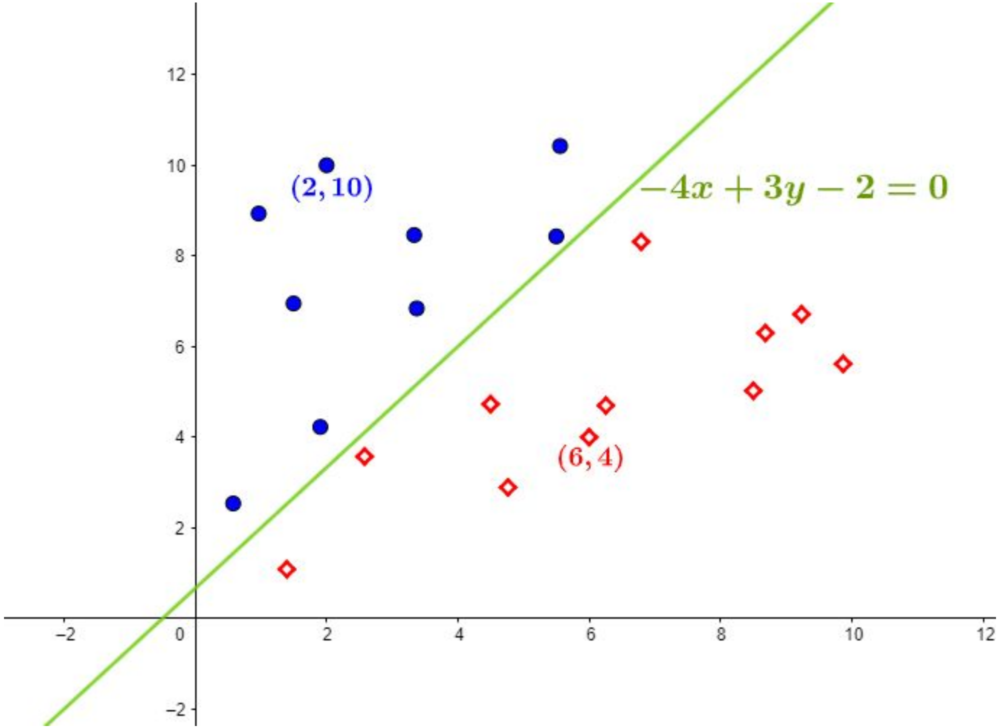
$y_i (-2 -4x_{i,1} + 3x_{i,2}) > 0$

$1 * (-2 -4 * 2 + 3 * 10) > 0$

$1 * (20) > 0$

$-1 * (-2 -4 * 6 + 3 * 4) > 0$

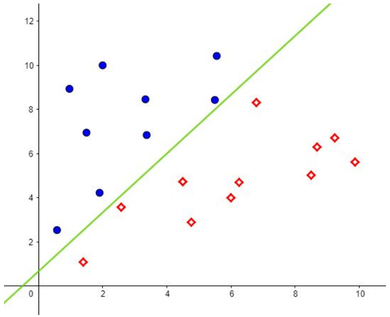
$-1 * (-14) > 0$



Separabilidad lineal

Dada una observación de test **x'** con p atributos podemos clasificarla de acuerdo a:

- **x'** será de la **clase 1** si $b_0 + b_1 x'_1 + b_2 x'_2 + \cdots + b_p x'_p > 0$
- **x'** será de la **clase -1** si $b_0 + b_1 x'_1 + b_2 x'_2 + \cdots + b_p x'_p < 0$



Separabilidad lineal

Sea $\mathbf{f\,(x') = b_0 + b_1x'_1 + b_2x'_2 + .. + b_px'_p}$, además, podemos decir que:

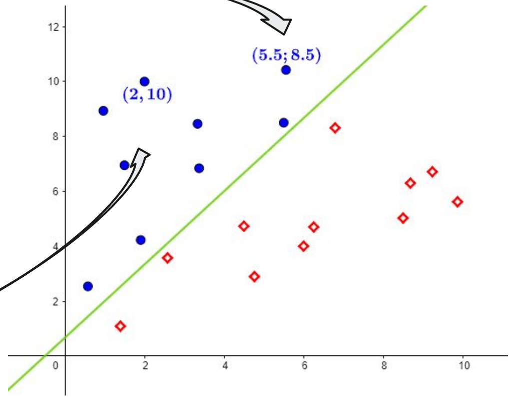
- Si f (x') es un valor cercano a 0, x' estará cerca del hiperplano
- Si f (x') es un valor lejano del 0, x' estará lejos del hiperplano.

$$-4.5,5+3.8,5-2=-1,5$$

-1.5 cercano a 0 → (5.5; 8.5) cerca del hiperplano

$$-4.2+3.10-2=20$$

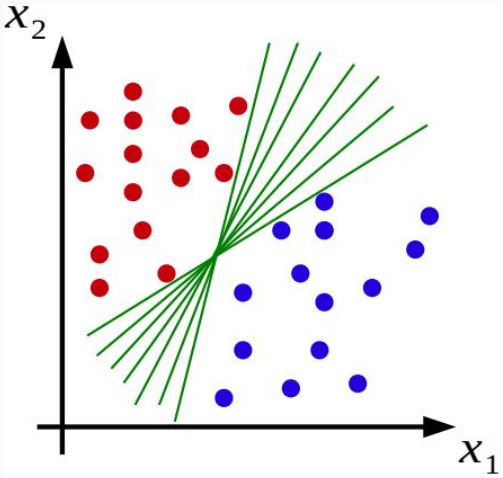
20 lejano a 0 → (2,10) lejos del hiperplano



El clasificador de margen maximal

¿Cómo separamos las clases?

Puede haber más de un hiperplano que los separe, en general, infinitos planos que separan ambas clases.

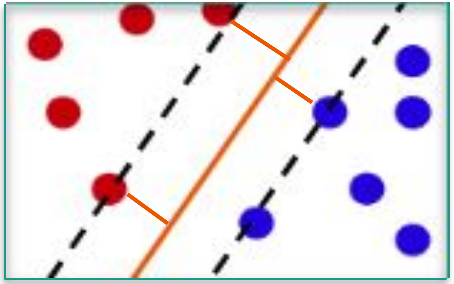


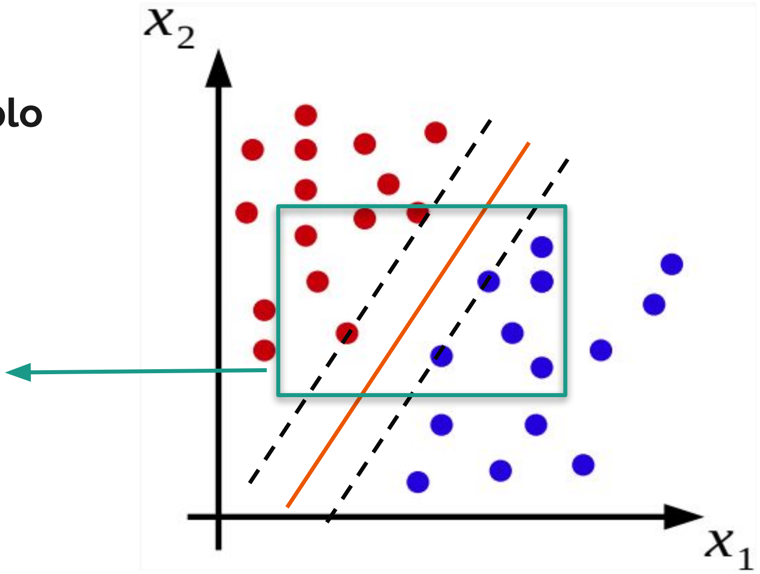
Sin embargo hay un hiperplano que posee una propiedad en particular.

El clasificador de margen maximal

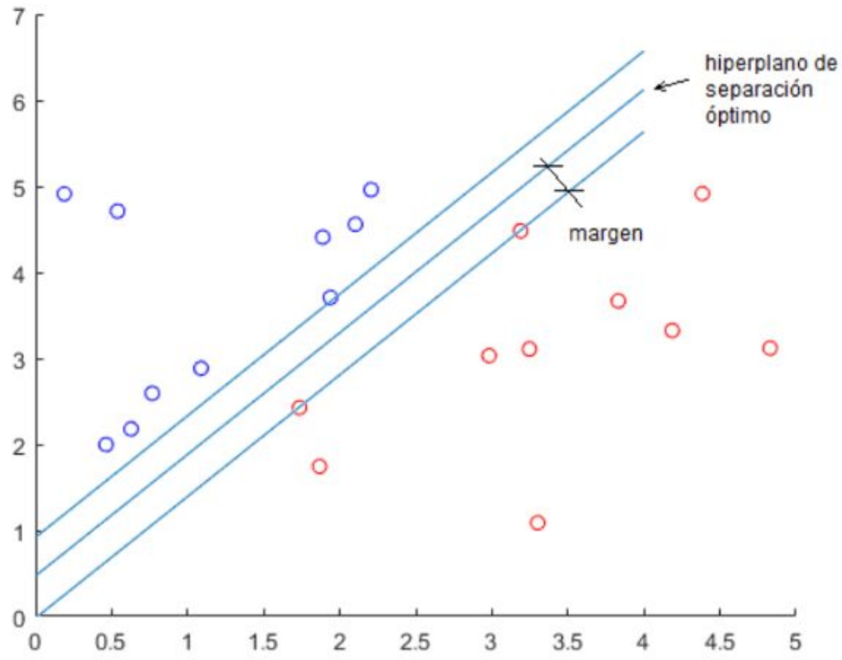
Sea **H** un hiperplano que separa ambas clases, consideremos las distancias de cada uno de los ejemplos a H.

Definiremos **margen** como la **distancia del ejemplo más cercano a H.**



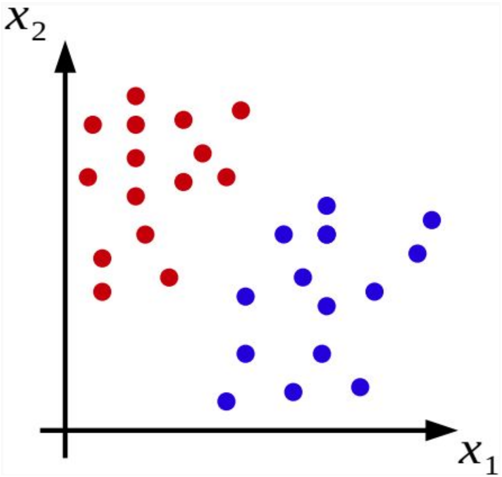


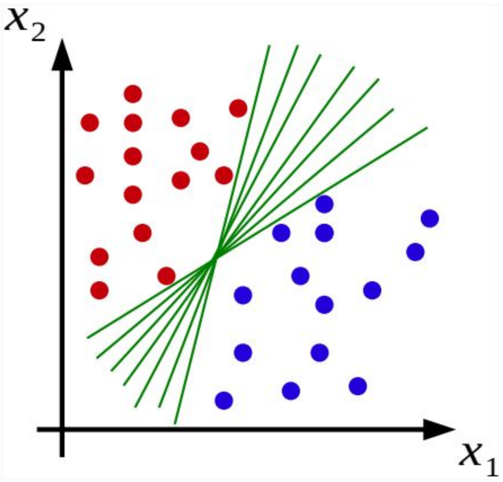
El clasificador de margen maximal

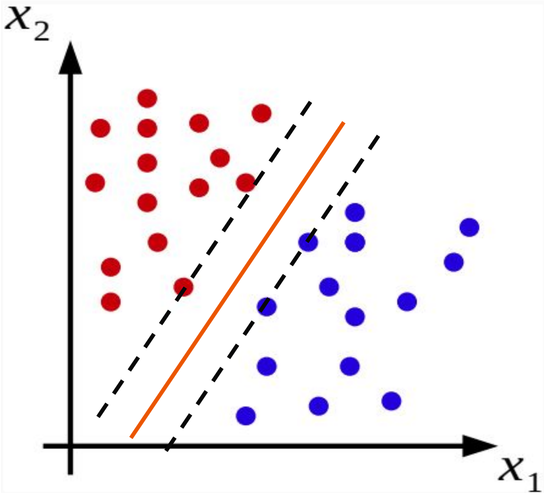


El clasificador de margen maximal

Nos interesa el hiperplano que posea margen máximo, es decir, **Hiperplano de margen maximal** o **Hiperplano de separación óptimo.**





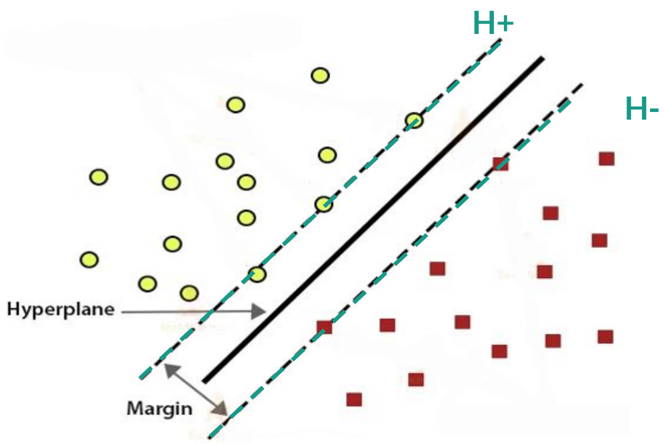


El clasificador de margen maximal

Si usamos el **hiperplano de margen maximal** para separar ambas clases, el clasificador se llama **Clasificador de margen maximal**.

Además quedan definidos dos hiperplanos más, equidistantes de H, **H+ y H–**.

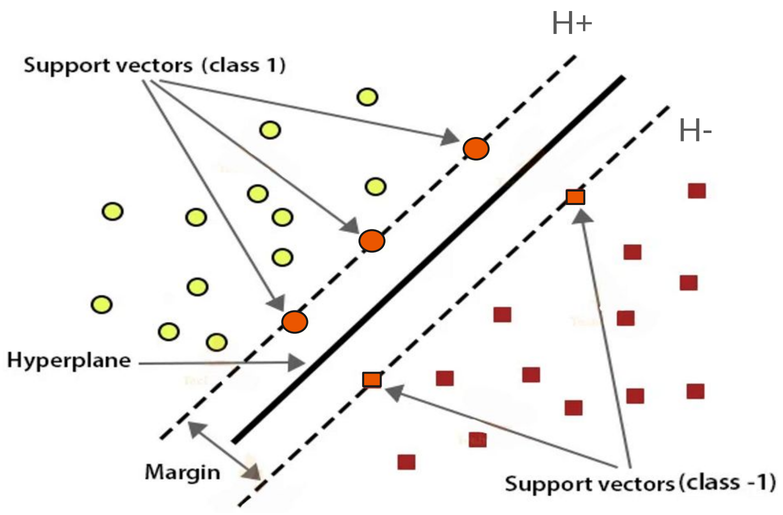
La distancia entre H+ y H, y entre H– y H, es la misma: el margen.



El clasificador de margen maximal

Sobre H+ y H– se pueden observar ejemplos de ambas clases que están sobre ellos, se denominan **vectores de soporte** (*support vectors*).

El **hiperplano de separación óptimo depende solamente de los vectores de soporte** y no del resto de los ejemplos de las clases.



El clasificador de margen maximal

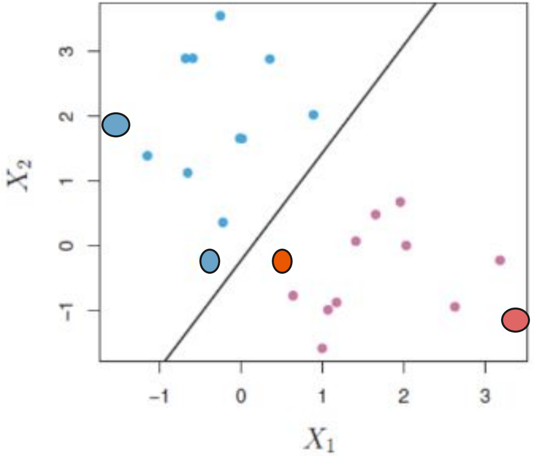
Construcción del Clasificador de margen maximal

Sea M el margen (el cual queremos maximizar) y dado que b define al hiperplano de separación óptimo entonces lo que queremos hacer es encontrar los valores de b tales que maximicen M sujeto a

- $y_i * (b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_p x_{i,p}) \geq M, \forall i, 1 \leq i \leq n,$
- $\sum_{j=1}^{p} b_j^2 = 1.$

Clasificador con vectores de soporte

La distancia de una observación de test al hiperplano de separación óptimo nos da una idea de la **confianza** que podemos tener en la clasificación.



- Si la **distancia es grande** tendremos **más confianza** (la observación está bastante adentro de la clase).

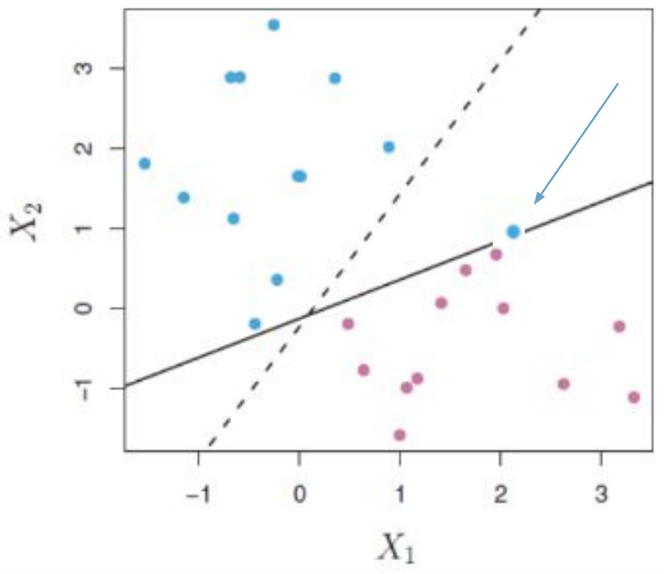
- Si la **distancia es pequeña**, cerca a 0, tendremos **menos confianza** (la observación está cerca del límite de la clase y por lo tanto cerca de la otra clase).

¿Es óptimo siempre el hiperplano de separación óptimo?

Supongamos dos clases que están bien separadas, es decir, tienen un margen considerable.

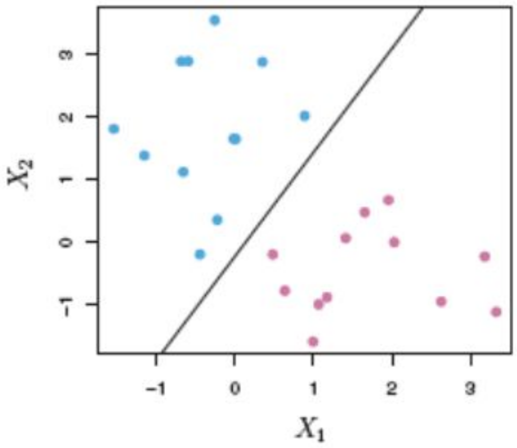
Agregamos un ejemplo que hace que el margen se reduzca considerablemente.

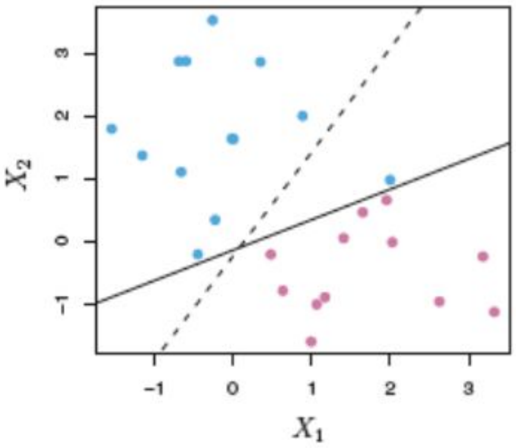
Ejemplos que con el hiperplano inicial hubieran sido clasificados con mayor confianza ahora estarán clasificados con menos confianza.



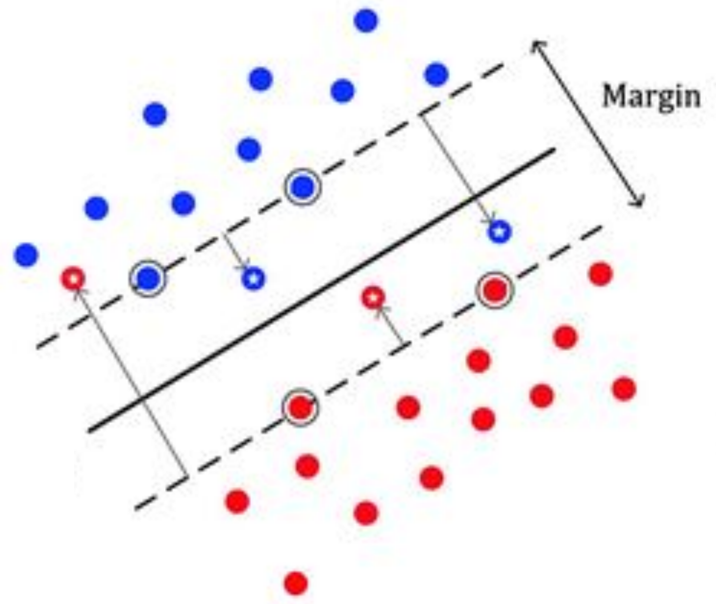
¿Es óptimo siempre el hiperplano de separación óptimo?

¿No valdría la pena usar el hiperplano inicial (el que no tenía en cuenta el ejemplo que hace el margen muy reducido) en vez de usar el nuevo hiperplano que divide pero acarrea un test menos confiable?





Clasificador con margen tolerante



Clasificador con margen tolerante

Objetivos:

- Que las observaciones individuales sean clasificadas con robustez (que la distancia al hiperplano no sea crítica)

- Una mejor clasificación de la mayor ía de los ejemplos de entrenamiento (asumiendo clases no linealmente separables).

Clasificador con margen tolerante

Encontrar los valores de b y ε_1, .., ε_n tales que maximicen M sujeto a

- $y_i * (b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_p x_{i,p}) \geq M * (1 - \epsilon_i), \ \forall i, 1 \leq i \leq n,$
- $\sum_{j=1}^{p} b_j^2 = 1.$
- $\forall i, \epsilon_i \geq 0 \wedge \sum_{i=1}^{n} \epsilon_i \leq C.$

donde C es un parámetro de ajuste del método.

Cada ε_i permite clasificar el ejemplo x_i en un lugar erróneo si fuera necesario.

Podría pasar por:

- estar dentro del margen de la clase
- estar del lado incorrecto del hiperplano.

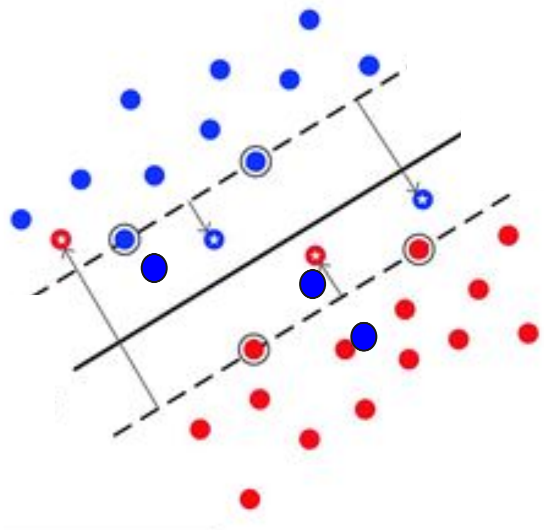
Clasificador con margen tolerante

Dada una observación x_i si

$$y_i * (b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_p x_{i,p}) \geq M * (1 - \epsilon_i), \forall i, 1 \leq i \leq n,$$

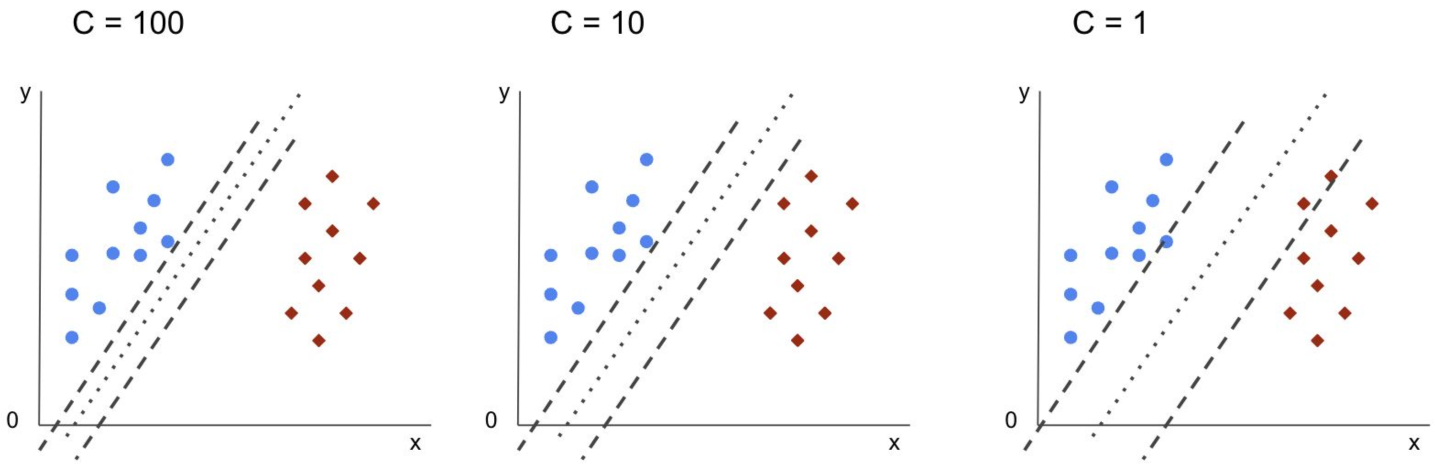
se satisface para

- **$\varepsilon_i = 0$** ⇒ la observación está del **lado correcto** del **margen**
- **$\varepsilon_i > 0$** ⇒ la observación está del **lado incorrecto** del **margen**
- **$\varepsilon_i > 1$** ⇒ la observación está del **lado incorrecto** del **hiperplano**



Clasificador con margen tolerante

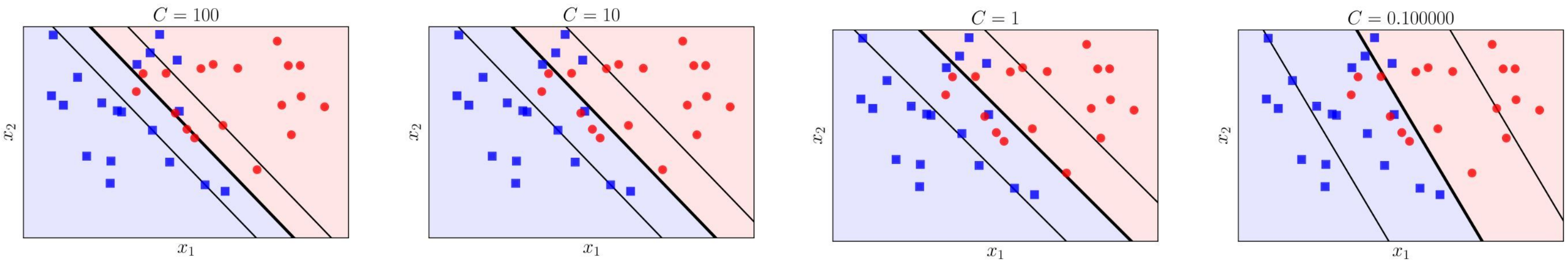
El valor de **C** es un parámetro que dice cuánto se va a permitir que las observaciones (en su conjunto) **violen el margen o el hiperplano**.



Clasificador con margen tolerante

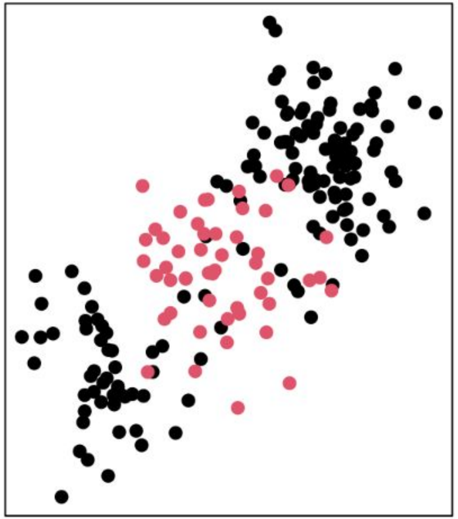
- Si C = 0, el Clasificador con margen tolerante se convierte en un Clasificador de margen maximal.
- Una forma de encontrar C es con validación cruzada.

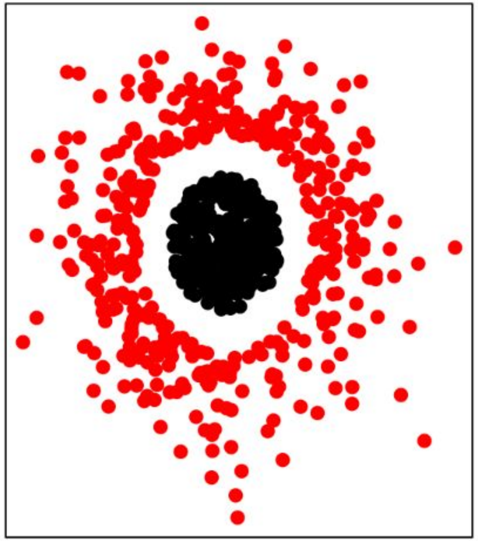
.



Clasificación con límites de decisión no lineales

Lo que vimos hasta el momento es efectivo cuando la separación entre clases es lineal, pero no funciona bien en casos no lineales.





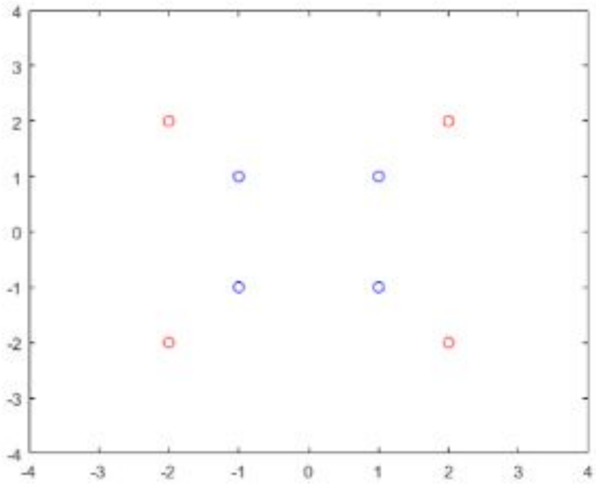
Clasificación con límites de decisión no lineales

Consideremos un conjunto de entrenamiento donde sus ejemplos x_i son de dimensión p = 2 y su clase es $y_i \in \{-1, 1\}$ (-1 en rojo y 1 en azul):

$X = \{(-2, -2),(-2, 2),(2, -2),(2, 2),(-1, -1),(-1, 1),(1, -1),(1, 1)\}$

$Y = \{-1, -1, -1, -1, 1, 1, 1, 1\}$

donde no se puede establecer un hiperplano de separación lineal entre una clase y la otra.

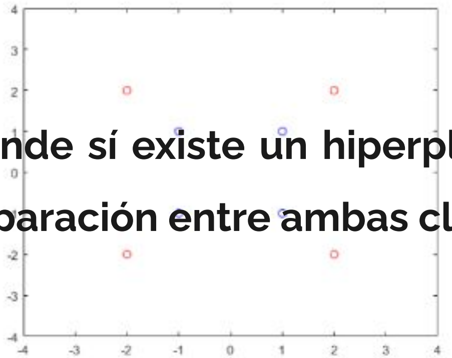


Clasificación con límites de decisión no lineales

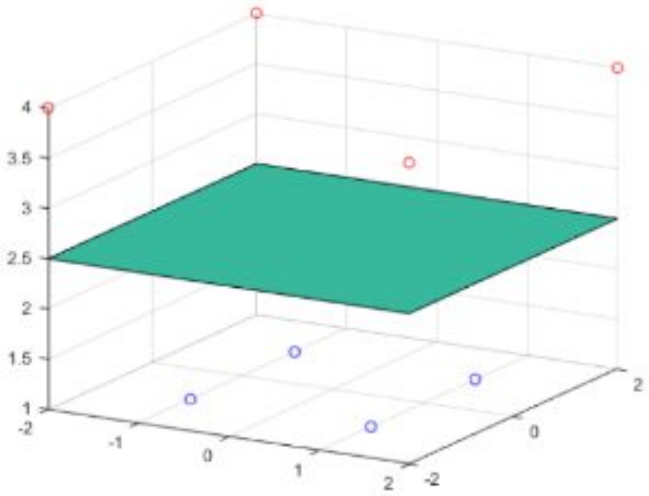
Pero si los ejemplos $x_i = (x_{i1}, x_{i2})$ de X en vez de representarlos en R^2 lo hacemos en R^3 como $x_i = (x_{i1}, x_{i2}, x^2_{i1})$ con el mismo Y:

$X = \{(-2, -2, 4),(-2, 2, 4),(2, -2, 4),(2, 2, 4),(-1, -1, 1),(-1, 1, 1),(1, -1, 1),(1, 1, 1)\}$

$Y = \{-1, -1, -1, -1, 1, 1, 1, 1\}$



Donde sí existe un hiperplano de separación entre ambas clases.



Clasificación con límites de decisión no lineales

Por ejemplo

Si tenemos ejemplos en una dimensión p x_{i1}, x_{i2}, .., x_{ip} podríamos representarlos en una dimensión 2p de acuerdo a

x_{i1}, x^2_{i1}, x_{i2}, x^2_{i2}, .., x_{ip}, x^2_{ip} donde podría haber un hiperplano de dimensión 2p-1 que los separara.

Clasificación con límites de decisión no lineales

En este caso, el problema a resolver es encontrar los valores de $b = b_0$, b_{11}, b_{12}, ..., b_{p1}, b_{p2} y ε_1, ..., ε_n tales que maximicen M sujeto a

- $y_i * (b_0 + \sum_{j=1}^{p} b_{j1} * x_{ij} + \sum_{j=1}^{p} b_{j2} * x_{ij}^2) \geq (M - \epsilon_i)$, $\forall i, 1 \leq i \leq n$
- $\sum_{j=1}^{p} \sum_{k=1}^{2} b_{jk}^2 = 1$.
- $\epsilon_i \geq 0, \forall_i, 1 \leq i \leq n \wedge \sum_{i=1}^{n} \epsilon_i \leq C$.

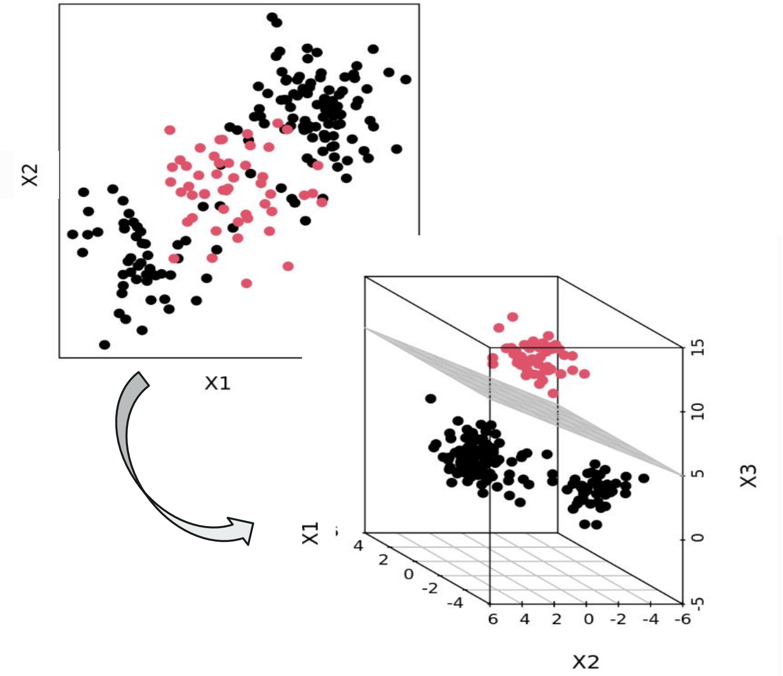
donde C es un parámetro de ajuste del método.

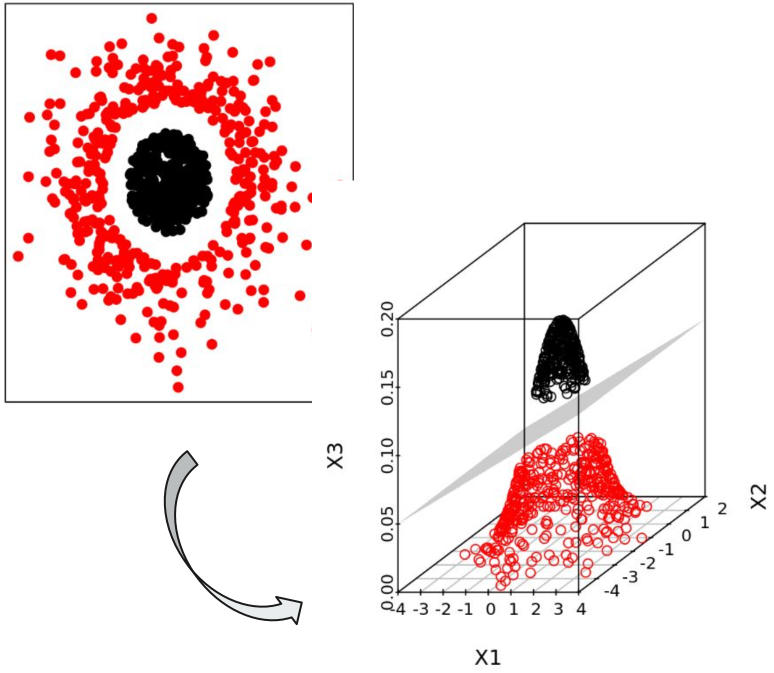
Clasificación con límites de decisión no lineales

En el ejemplo, utilizamos x^2_{ij}, pero podríamos utilizar otro grado u otra función.

La idea es obtener un espacio donde exista separabilidad lineal implica separabilidad no lineal en el espacio original de los ejemplos de dimensión p.

Clasificación con límites de decisión no lineales





Máquina basada en vectores de soporte

Esta propuesta es una generalización de la Clasificación con líımites de decisión no lineales y la forma de generalizar esta idea es mediante la introducción del concepto de **Núcleo (Kernel).**

Máquina basada en vectores de soporte

El clasificador del margen maximal solamente depende de los vectores soporte, entonces, $\exists$, α_1, . . . , α_k tal que:

Dado una observación x', si queremos saber a qué clase pertenece, calculamos f (x') como:

$$\mathbf{f(x') = b_o + \Sigma^k_{i=1} \alpha_i \langle x', x_i \rangle}$$

donde los x_i son los vectores de soporte.

Máquina basada en vectores de soporte

Podemos escribir a f (x) como:

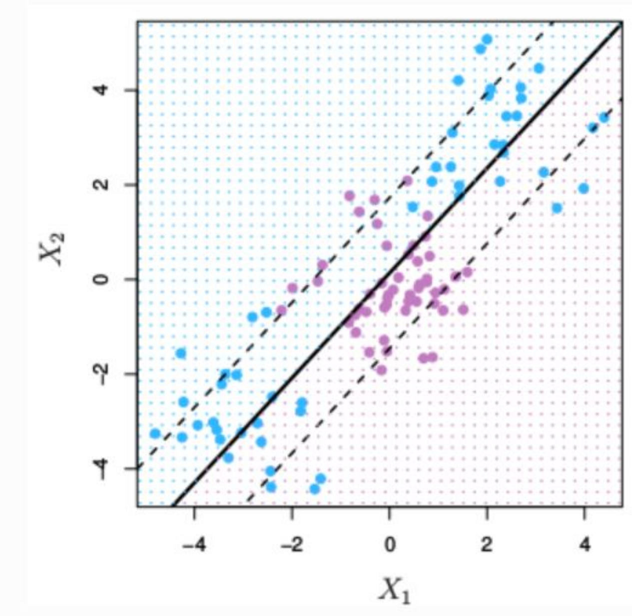
$$f(x) = b_0 + \sum_{i=1}^{k} \alpha_i K(x, x_i)$$

donde $K(x, x_i) = \langle x, x_i \rangle$ y a **K** se lo denomina **Núcleo (Kernel)**.

Máquina basada en vectores de soporte

Núcleo lineal

$$K(\mathbf{x}, \mathbf{x'}) = \mathbf{x} \cdot \mathbf{x'}$$



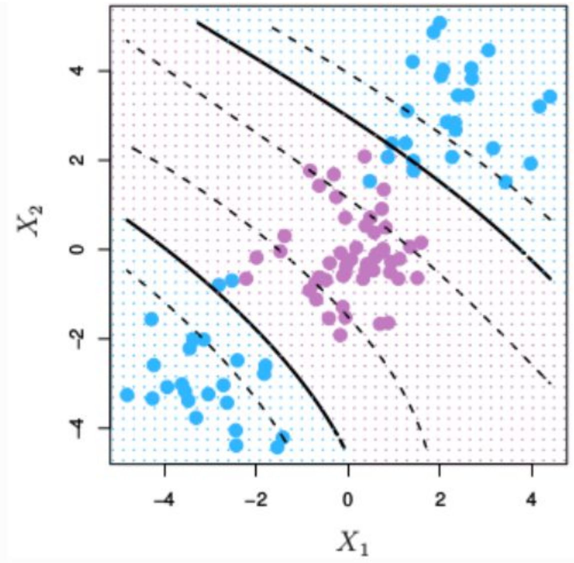
Máquina basada en vectores de soporte

Núcleo polinómico

$$K(x', x_i) = (1 + \sum_{j=1}^{p} x_{ij} x'_j)^d$$

donde d es el grado del polinomio.

A medida que d se incrementa habrá mayor flexibilidad para encontrar una separación lineal en el nuevo espacio de los ejemplos (el espacio ampliado).



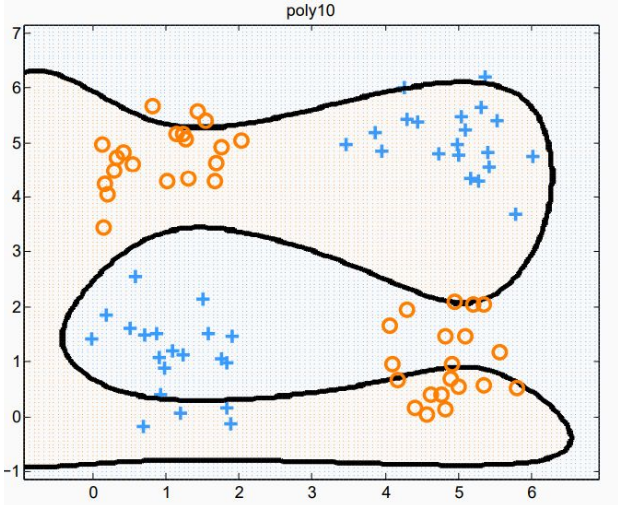
Máquina basada en vectores de soporte

Núcleo polinómico

$$K(x', x_i) = (1 + \sum_{j=1}^{p} x_{ij} x_j')^d$$

donde d es el grado del polinomio.

A medida que d se incrementa habrá mayor flexibilidad para encontrar una separación lineal en el nuevo espacio de los ejemplos (el espacio ampliado).

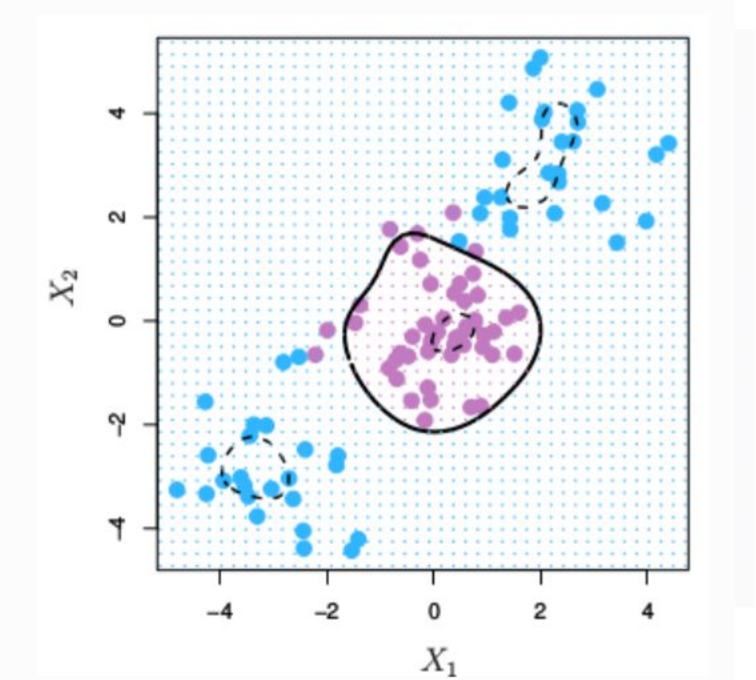


Máquina basada en vectores de soporte

Núcleo radial

$$K(x', x_i) = e^{-\gamma \sum_{j=1}^{p} (x_{ij} - x'_j)^2}$$

donde γ es una constante positiva.

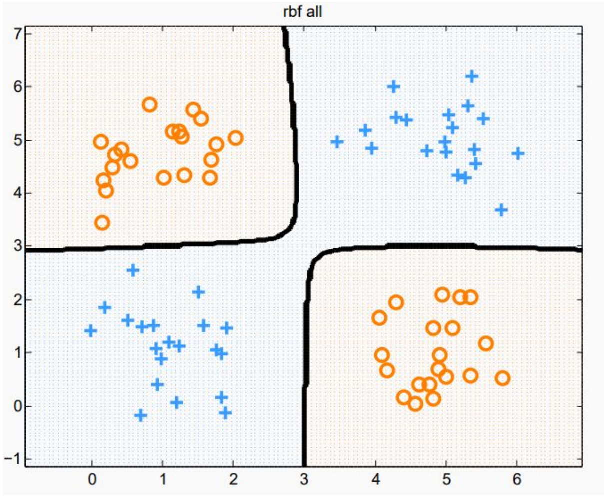


Máquina basada en vectores de soporte

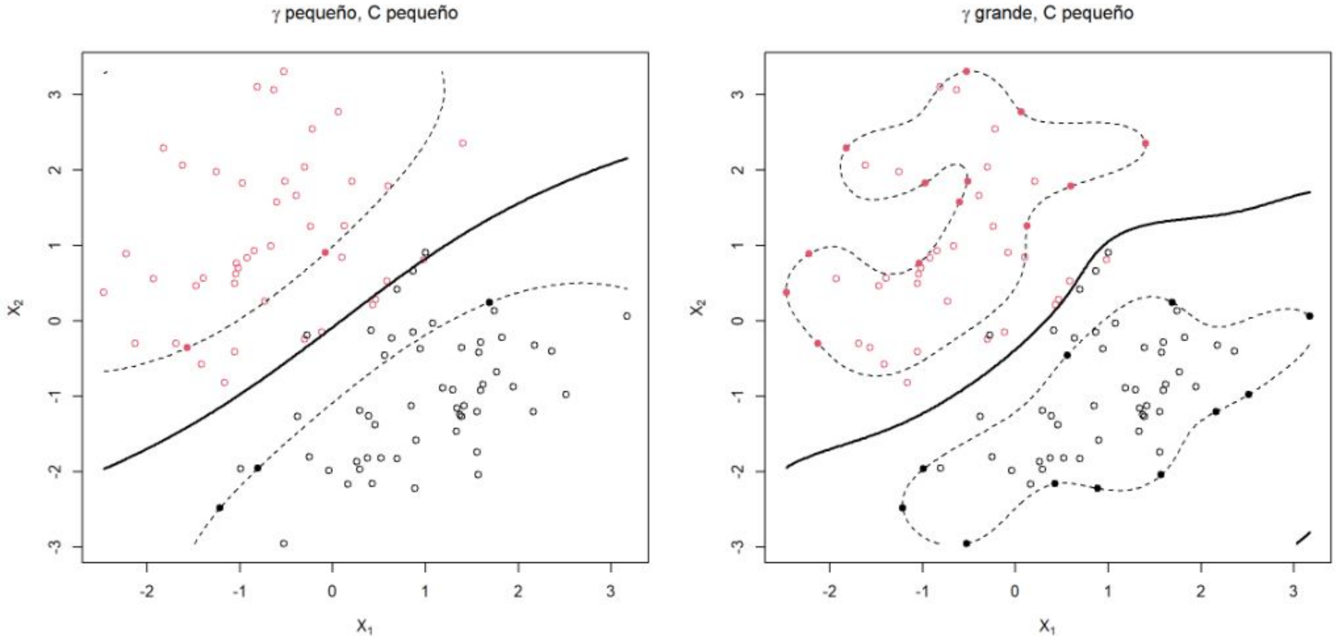
Núcleo radial

$$K(x', x_i) = e^{-\gamma \sum_{j=1}^{p}(x_{ij}-x'_j)^2}$$

donde γ es una constante positiva.



Máquina basada en vectores de soporte



Máquina basada en vectores de soporte

Aunque cambiemos el Núcleo, la formulación para el problema no cambia.

El cálculo de f(x) seguirá siendo:

$$f(x) = b_0 + \sum_{i=1}^{n} \alpha_i K(x, x_i)$$

Support vector machine multiclase

¿Qué ocurre cuando nuestro conjunto de entrenamiento tiene más de 2 clases?

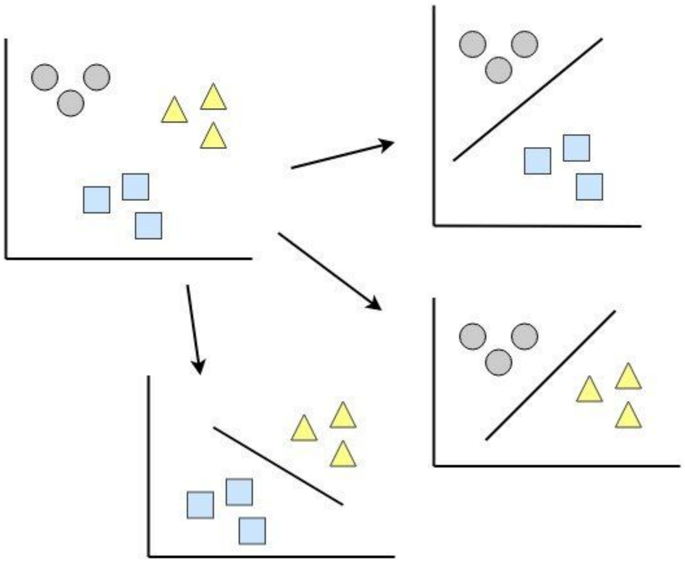
Hay dos abordajes:

- Uno contra uno
- Uno contra el resto

Support vector machine multiclase

Uno contra uno

Se construye un SVM para cada par distinto de clases (una clase se le asignar á +1 y a la otra -1). El resto de los ejemplos de las clases restantes se ignorará.



Cuando se presente una nueva observación de cada SMV construido se obtiene una respuesta y se elige la clase más votada.

SVM - Support vector machine multiclase

Uno contra todos

Se construye un SVM para cada clase (a los ejemplos de clase se le asignar á +1 y al resto de los ejemplos del conjunto de entrenamiento se le asignar á -1).

Cuando se presente una nueva observación de cada SMV construido se obtiene una respuesta y la clase resultante será la que corresponda al SVM cuyo f (x') sea mayor.