Máquinas de soporte vectorial.

César Gómez

26 de octubre de 2020

Introducción

En este capítulo se presentará la máquina de soporte vectorial (SVM), se trata de un clasificador desarrollado en la comunidad de ciencias de la computación en la década de los 90s y que ha crecido en popularidad desde entonces. Las SVM han demostrado buen desempeño en una variedad de contextos. Son consideradas uno de los mejores clasificadores de la caja de herramientas del especialista en machine learnings. Se presentarán en este capítulo entre otras ideas:

- *Clasificador de margen máximal*.
- *Clasificador de soporte vectorial*.
- *Máquina de soporte vectorial*.

Clasificador de margen máximal

Producto interno de vectores en $\mathbb{R}^n$, **proyección ortógonal**.

Las ideas detrás de las máquinas de soporte vectorial son más fáciles de entender equipados con una buena comprensión del producto interno de vectores.

Definición 1

El producto interno de 2 vectores $a = [a_1, a_2, \ldots, a_n]$ *y* $b = [b_1, b_2, \ldots, b_n]$ *es el número real denotado por* $a \cdot b$ *definido por*

$$a \cdot b = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n. \tag{1}$$

Realmente nos interesa la idea de componente ortogonal.

- Considérense los vectores e_1, e_2 que constituyen la base canónica para vectores en $\mathbb{R}^2$.

$$e_1 = [1\ 0] \quad e_2 = [0\ 1],$$

- Obsérvese que en el plano $\mathbb{R}^2$, cualquier vector $a = [a_1\ a_2]$ se puede representar como

$$\begin{aligned} [a_1\ a_2] &= a_1 e_1 + a_2 e_2, \\ &= (a \cdot e_1)e_1 + (a \cdot e_2)e_2, \end{aligned} \tag{2}$$

- Observe también que dados 2 vectores en $\mathbb{R}^2$

$$a = [a_1 \ a_2] \quad b = [b_1 \ b_2],$$

entonces podemos representar uno de ellos, por ejemplo $\boldsymbol{a}$ como

$$a = \alpha b + \beta b^{\perp}, \tag{3}$$

Veamos cual es el valor de α

$$\begin{aligned} a \cdot b &= (\alpha b + \beta b^{\perp}) \cdot b, \\ &= \alpha b \cdot b + \beta b^{\perp} \cdot b, \\ &= \alpha b \cdot b, \end{aligned} \tag{4}$$

Por lo tanto

$$\alpha = \frac{a \cdot b}{b \cdot b}. \tag{5}$$

- A la cantidad

$$\alpha = \frac{a \cdot b}{b \cdot b}, \tag{6}$$

se le denomina la componente ortogonal de b en a.
Obsérvese que si b es un vector de norma 1,
$b \cdot b = b_1^2 + b_2^2 = |b|^2 = 1$ entonces

$$\alpha = a \cdot b. \tag{7}$$

Que es un hiperplano

- Comencemos por $\mathbb{R}^2$, en este caso un hiperplano es definido por la ecuación

$$\begin{aligned} \beta_0 + \beta_1 X_1 + \beta_2 X_2 &= 0, \\ &\iff \\ [\beta_1 \quad \beta_2] \cdot [X_1 \quad X_2] &= -\beta_0 \end{aligned} \tag{8}$$

(9)

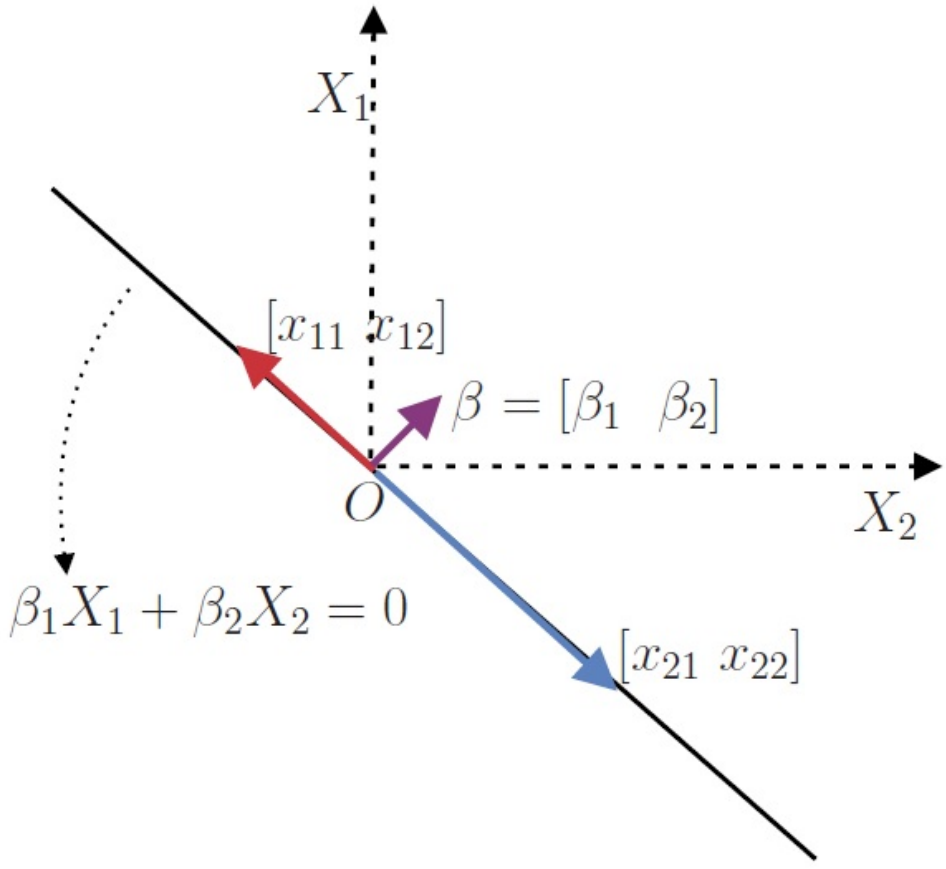


Figura 1: .

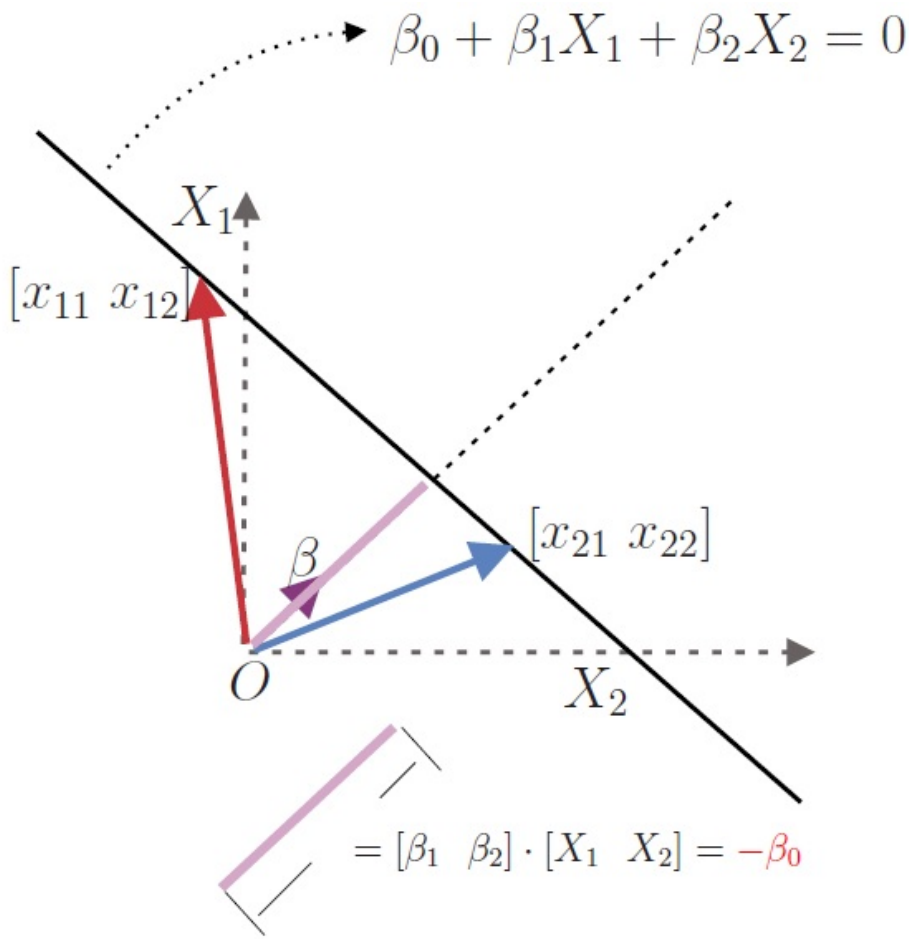

$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$
X_1
$[x_{11}\ x_{12}]$
β
$[x_{21}\ x_{22}]$
O
X_2
$= [\beta_1\ \ \beta_2] \cdot [X_1\ \ X_2] = -\beta_0$

Figura 2: .

- En $\mathbb{R}^3$ tenemos que la ecuación de un hipérplano viene dada por

$$\begin{aligned} \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 &= 0, \\ &\iff \\ [\beta_1 \quad \beta_2 \quad \beta_3] \cdot [X_1 \quad X_2 \quad X_3] &= -\beta_0 \end{aligned} \tag{10}$$

(11)

- Ahora en el caso p-dimensional, la ecuación que define un hiperplano es

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p = 0, \tag{12}$$

Esta ecuación define un hiperplano en el sentido de que si un punto $X = [X_1 \quad X_2 \cdots X_p]$ en el espacio p-dimensional, satisface (12) entonces el punto está en el hiperplano.

Suponga ahora que el punto $X = [X_1 \;\; X_2 \cdots X_p]$ no satisface exactamente (12) y en vez

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p > 0, \tag{13}$$

acá (13) indica que el punto X está en un lado del hiperplano. Y si en cambio

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p < 0, \tag{14}$$

entonces el punto X está al otro lado del hiperplano.

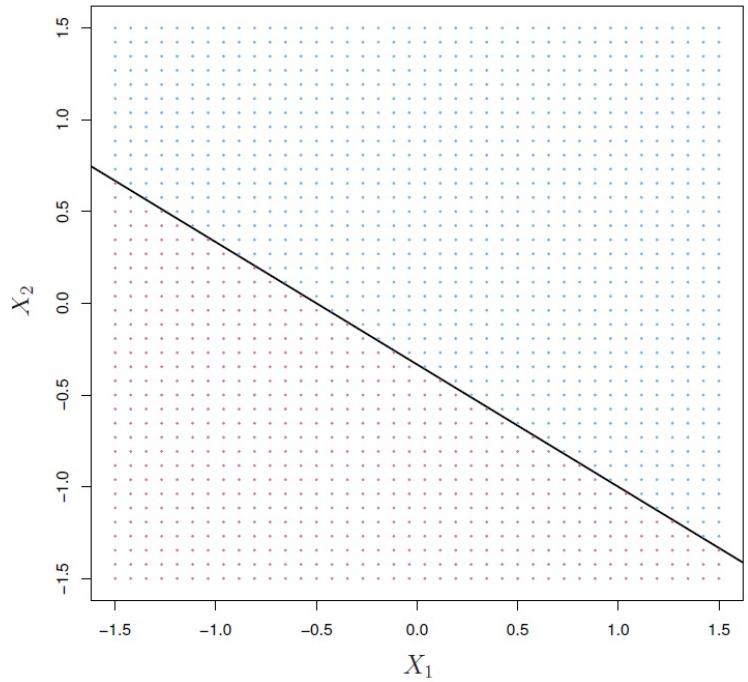


Figura 3: Se ilustra el hiperplano $1 + 2X_1 + 3X_2 = 0$. La región azul corresponde al conjunto de puntos para los cuales $1 + 2X_1 + 3X_2 > 0$ y la otra región al conjunto de puntos para los que $1 + 2X_1 + 3X_2 < 0$.

Clasificación utilizando un hiperplano separador

Supóngase que se tiene una matriz de datos $\boldsymbol{X}$ que consiste de n observaciones de entrenamiento en el espacio p-dimensional

$$x_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1p} \end{pmatrix}, \ldots, \quad x_n = \begin{pmatrix} x_{n1} \\ \vdots \\ x_{np} \end{pmatrix}. \tag{15}$$

y además que estas observaciones están "etiquetadas" en 2 clases, es decir $y_1, \ldots, y_n \in \{-1, 1\}$ donde -1 representa una de las clases y 1 la otra.

También se dispone de una observación de prueba o test correspondiente a un p-vector de atributos (features) observados

$$x^* = [x_1^* \ \cdots \ x_p^*]^T.$$

El objetivo consiste en establecer un clacificador que clasifique correctamente a x^* basado en el valor de sus atributos. Ya se han visto algunas metodologías como:

- Análisis de discriminante lineal.
- Regresión logística.
- Arboles de clasificación.

Ahora, desarrollaremos una metodología basada en un hiperplano separador.

Un hiperplano separador posee la propiedad de que

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} > 0 \quad \text{si} \quad y_i = 1, \tag{16}$$

y

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} < 0 \quad \text{si} \quad y_i = -1, \tag{17}$$

Obsérvese que tanto (16) como (17) se pueden representar por medio de

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) > 0, \tag{18}$$

- Considerando una observación test o de prueba $x^* = [x_1^* \ \cdots \ x_p^*]$, la clasificación de esta depende entonces del *signo* de

$$f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*, \tag{19}$$

- Considerando una observación test o de prueba $x^* = [x_1^* \ \cdots \ x_p^*]$, la clasificación de esta depende entonces del *signo* de

$$f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*, \tag{19}$$

- Pero también la **magnitud** de $f(x^*)$, es decir que tan lejos $f(x^*)$ está del cero 0 da una idea de la confianza o credibilidad de la clase asignada a x^*.

- Considerando una observación test o de prueba $x^* = [x_1^* \ \cdots \ x_p^*]$, la clasificación de esta depende entonces del *signo* de

$$f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*, \tag{19}$$

- Pero también la **magnitud** de $f(x^*)$, es decir que tan lejos $f(x^*)$ está del cero 0 da una idea de la confianza o credibilidad de la clase asignada a x^*.
- Un clasificador basado en un hiperplano separador conlleva a considerar una frontera de decisión lineal.

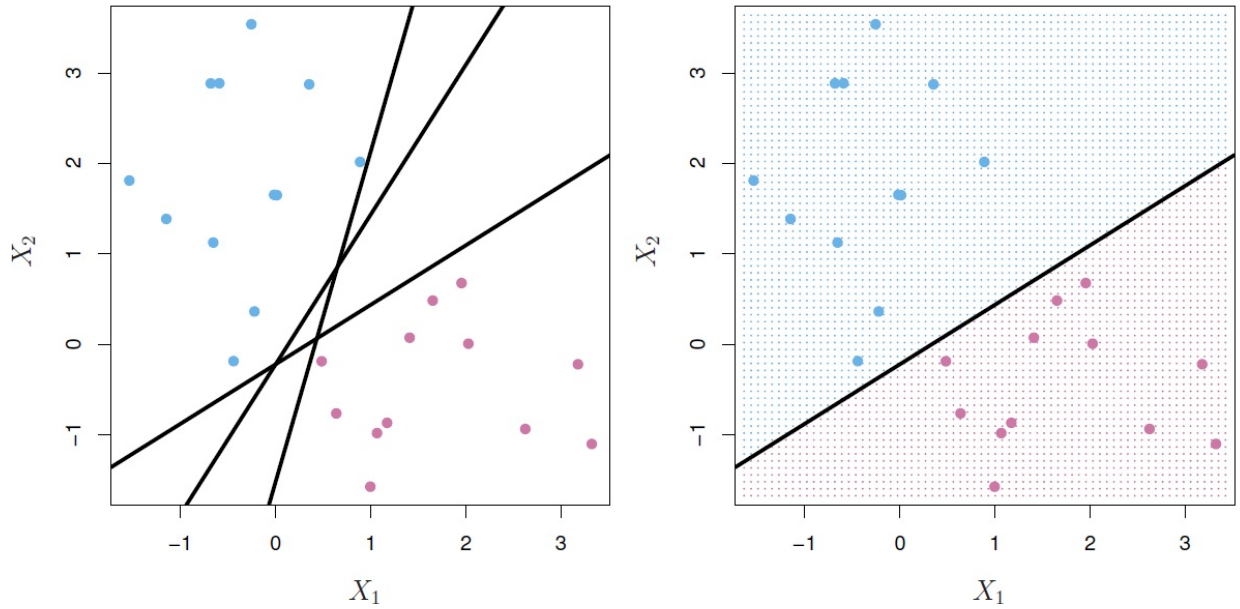
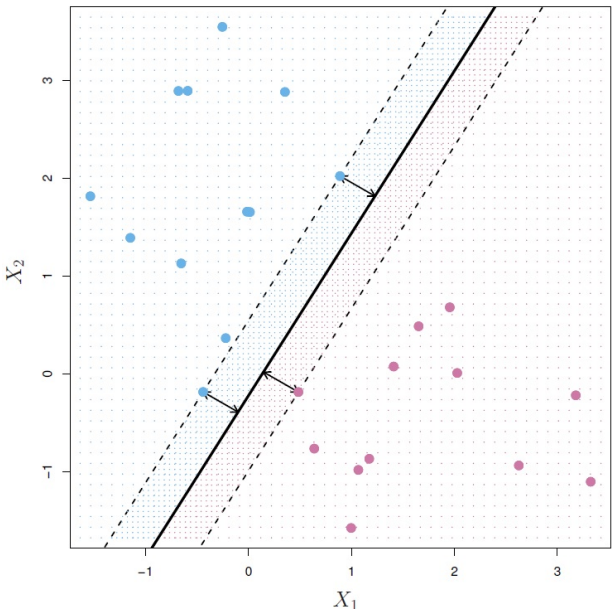


Figura 4: Izquierda. Hay dos clases de observaciones, azules y rojas, cada una con mediciones de 2 atributos o predictores. Se ilustran 3 hiperplanos separadores entre muchos otros posibles.
Derecha. Las mallas azul y roja indican la regla de decisión basada en el hiperplano separador.

Como se puede apreciar en la figura 4, puede existir no solamente uno pero varios hiperplanos separadores. ¿Cómo seleccionar uno, que sea el "más adecuado"?

El clasificador de margen máximal.



Obsérvese que el hiperplano de margen máximal solo depende de los vectores de soporte y no del resto de puntos.

- Recuerde una observación test o de prueba $x^* = [x_1^* \ \cdots \ x_p^*]$ es clasificada dependiendo del *signo* de

$$f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*, \qquad (20)$$

donde $\beta_0, \ldots, \beta_p$ resuelve el problema de optimización que se describe a continuación.

Construcción del clasificador de margen máximal

Considérese que la muestra de entrenamiento es $x_1, \ldots, x_n \in \mathbb{R}^p$ con "etiquetas" $y_1, \ldots, y_n \in \{-1, 1\}$. Brevemente el **hiperplano separador de margen maximal** es la solución del problema de optimización

$$\underset{\beta_0, \beta_1, \ldots, \beta_p, M}{\text{Maximize}} \; M \tag{21}$$

$$\text{Sujeto a} \quad \sum_{j=1}^{p} \beta_j^2 = 1 \tag{22}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M \quad \forall i = 1, 2, \ldots, n. \tag{23}$$

Que hacer en el caso no separable

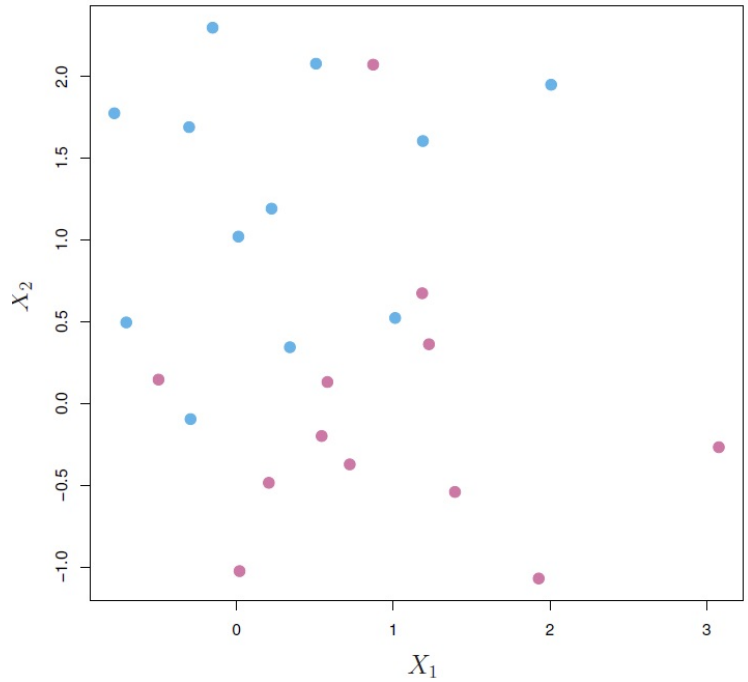


Figura 5: En este caso no existe un hiperplano separador

Acá no se puede utilizar el **separador de margen máximal.**

Clasificador de soporte vectorial

- Un clasificador basado en un hiperplano **separador de margen máximal** necesariamente clasifica de forma perfecta todas las observaciones de entrenamiento. (Claro si existe un hiperplano separador).

Clasificador de soporte vectorial

- Un clasificador basado en un hiperplano **separador de margen máximal** necesariamente clasifica de forma perfecta todas las observaciones de entrenamiento. (Claro si existe un hiperplano separador).
- Esto puede causar problemas de *estabilidad* en relación a observaciones individuales. O dicho de otra forma el hiperplano de margen máximal es extremadamente sensible al cambio de una sola observación, esto es una indicación de que puede estar teniendo lugar un *sobre-ajuste* de los datos.

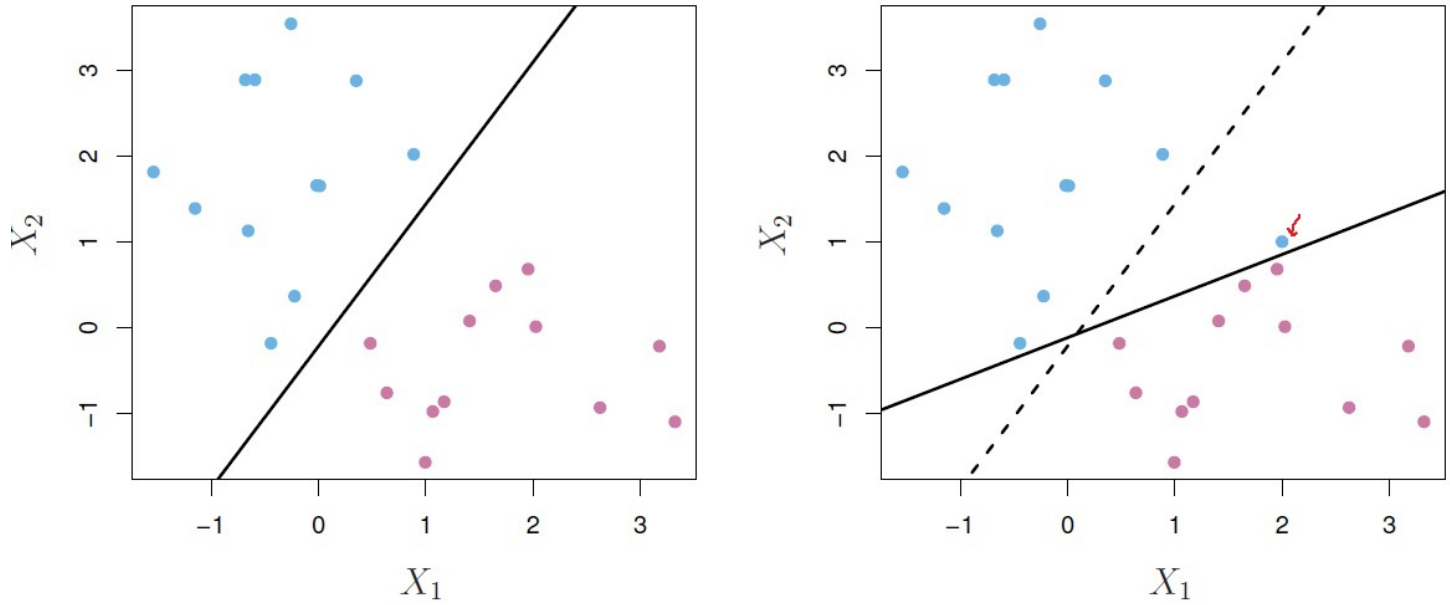


Figura 6: Izquierda. Se ilustra el hiperplano de margen máximal para un conjunto de datos.
Derecha. La inclusión del nuevo dato indicado, crea un cambio dramático en el hiperplano de margen máximal.

- El *clasificador de soporte vectorial* también es conocido como **clasificador de margen suave**.

- El *clasificador de soporte vectorial* también es conocido como **clasificador de margen suave**.
- Con este clasificador se obtiene mayor robustez a observaciones individuales.

- El *clasificador de soporte vectorial* también es conocido como **clasificador de margen suave**.
- Con este clasificador se obtiene mayor robustez a observaciones individuales.
- El énfasis se centra en clasificar correctamente la mayoría de las observaciones en vez de a todas.

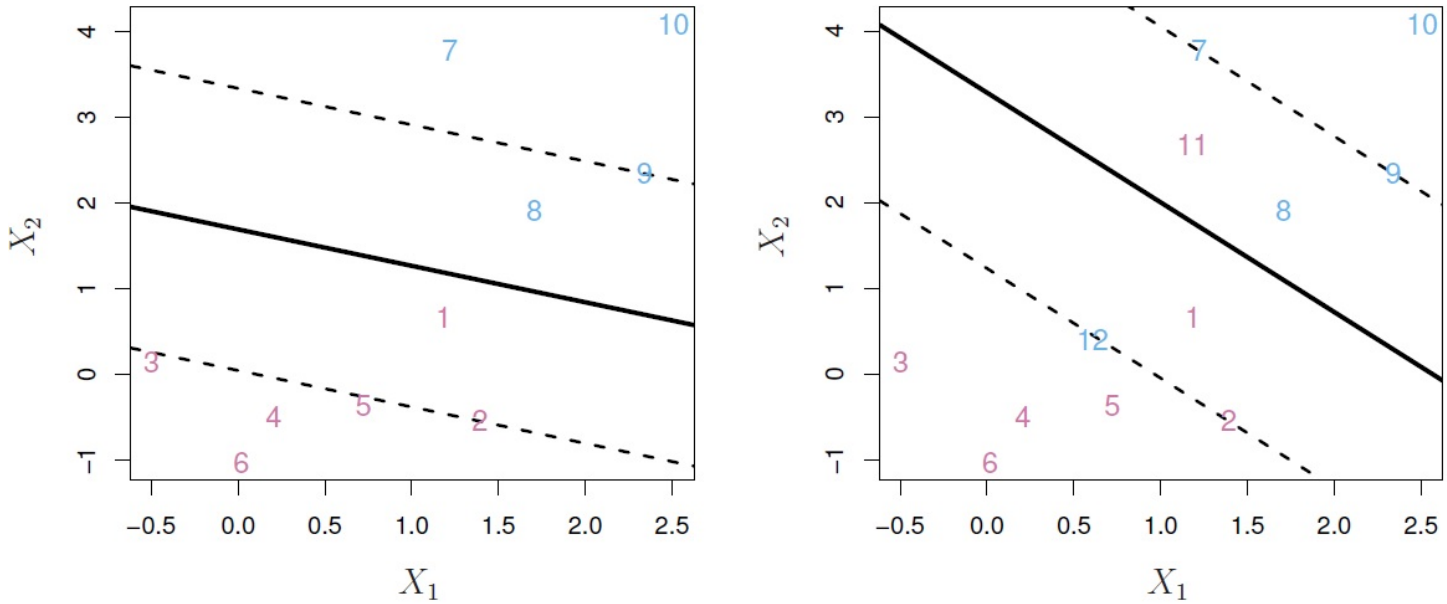


Figura 7: **Izquierda**: Un clasificador de soporte vectorial es ajustado a un conjunto pequeño de datos, observaciones 3, 4, 5 y 6 están en el lado correcto de la margen en relación a la clase roja. La observación 2 está sobre la margen y la observación 1 está en el lado incorrecto de la margen. **Derecha**: Lo mismo que en el lado izquierdo, pero con la inclusión de 2 nuevas observaciones 11 y 12 que están tanto en el lado errado del hiperplano como en el lado errado de la margen.

Construcción del clasificador de soporte vectorial

ó de margen suave.

$$\underset{\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots, \epsilon_n, M}{\text{Maximize}} \; M \tag{24}$$

$$\text{Sujeto a} \quad \sum_{j=1}^{p} \beta_j^2 = 1 \tag{25}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \epsilon_i) \quad \forall i = 1, 2, \ldots, n. \tag{26}$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C \tag{27}$$

- Una **observación de prueba** x^* es clasificada, según el signo de

$$f(x^*) = \beta_0 + \beta_1 x_1^* + \cdots + \beta_p x_p^*. \tag{28}$$

  donde los parámetros $\beta_0, \ldots, \beta_p$ resuelven el problema de optimización (24).

- C es un parámetro de ajuste, toma valores no negativos y controla la magnitud de las variables $\epsilon_1, \ldots, \epsilon_n$.
- $\epsilon_1, \ldots, \epsilon_n$ son variables "libres" o "sueltas" que permiten a las observaciones individuales estar en el lado errado de la margen o del hiperplano.

si $\epsilon_i > 0$ la observación x_i está en el lado errado de la **margen**.

si $\epsilon_i > 1$ la observación x_i está en el lado errado del **semiplano**.

- Si $C = 0$ entonces $\epsilon_1 = 0, \ldots, \epsilon_n = 0$ y ninguna observación viola la margen.

- Si $C = 0$ entonces $\epsilon_1 = 0, \ldots, \epsilon_n = 0$ y ninguna observación viola la margen.
- Si $C > 0$ no más que C observaciones pueden violar la margen.

- Si $C = 0$ entonces $\epsilon_1 = 0, \ldots, \epsilon_n = 0$ y ninguna observación viola la margen.
- Si $C > 0$ no más que C observaciones pueden violar la margen.
- C en realidad controla el balance entre sesgo y varianza del clasificador.

- Si $C = 0$ entonces $\epsilon_1 = 0, \ldots, \epsilon_n = 0$ y ninguna observación viola la margen.
- Si $C > 0$ no más que C observaciones pueden violar la margen.
- C en realidad controla el balance entre sesgo y varianza del clasificador.
- De hecho si C aumenta entonces la margen se amplía en este caso se permiten más violaciones de la margen, el método posee menos varianza a costas de un aumento en el sesgo y si C disminuye la margen se angosta, se disminuye el sesgo, los datos son ajustados de forma más fuerte y hay más varianza.

- Si $C = 0$ entonces $\epsilon_1 = 0, \ldots, \epsilon_n = 0$ y ninguna observación viola la margen.
- Si $C > 0$ no más que C observaciones pueden violar la margen.
- C en realidad controla el balance entre sesgo y varianza del clasificador.
- De hecho si C aumenta entonces la margen se amplía en este caso se permiten más violaciones de la margen, el método posee menos varianza a costas de un aumento en el sesgo y si C disminuye la margen se angosta, se disminuye el sesgo, los datos son ajustados de forma más fuerte y hay más varianza.
- El hiperplano de margen suave **solo** depende de las observaciones que están o bien sobre la margen o que violan la margen. A estas observaciones se les denomina **vectores de soporte**.

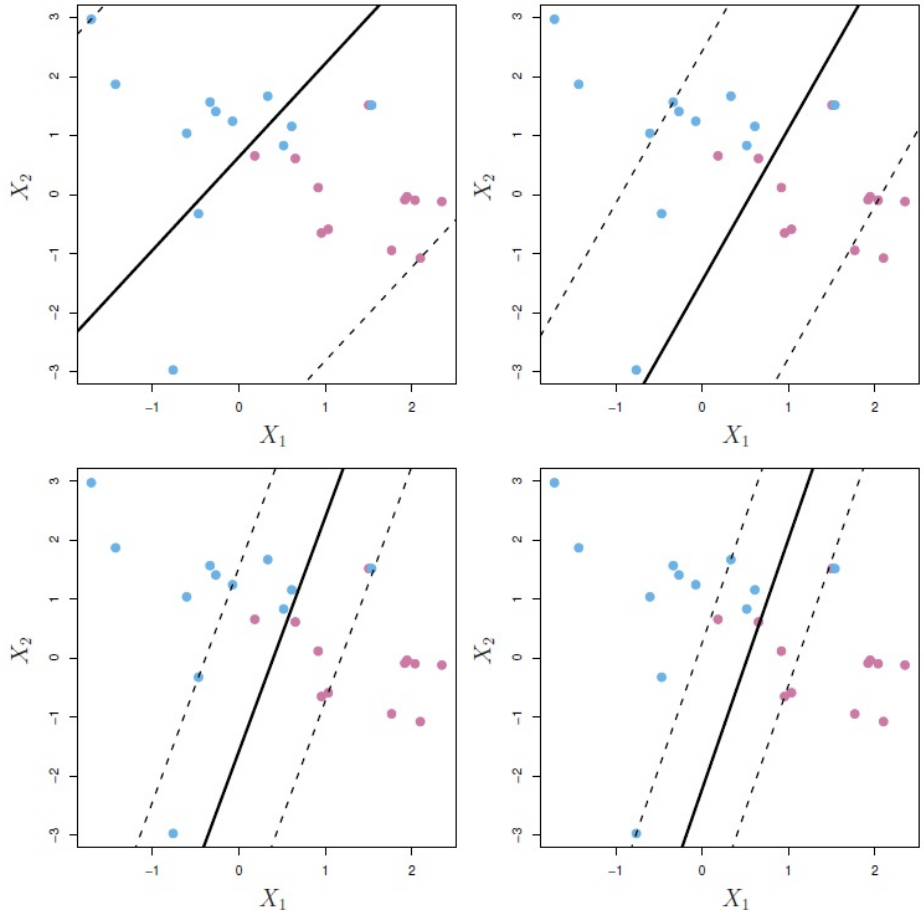


Figura 8: Ajustes de clasificador de soporte vectorial con diferentes valores del parámetro C.

- El hecho de que la regla de decisión basada en el clasificador de soporte vectorial depende de una cantidad potencialmente pequeña de observaciones de entrenamiento (vectores de soporte), implica que el clasificador sea bastante **robusto** en relación a observaciones que se encuentran apartadas de la margen.

- El hecho de que la regla de decisión basada en el clasificador de soporte vectorial depende de una cantidad potencialmente pequeña de observaciones de entrenamiento (vectores de soporte), implica que el clasificador sea bastante **robusto** en relación a observaciones que se encuentran apartadas de la margen.
- La propiedad descrita arriba está en contraste con la forma en que se comportan otros métodos de clasificación como es el caso del discriminante lineal.

- El hecho de que la regla de decisión basada en el clasificador de soporte vectorial depende de una cantidad potencialmente pequeña de observaciones de entrenamiento (vectores de soporte), implica que el clasificador sea bastante **robusto** en relación a observaciones que se encuentran apartadas de la margen.
- La propiedad descrita arriba está en contraste con la forma en que se comportan otros métodos de clasificación como es el caso del discriminante lineal.
- Recuérdese que la regla de clasificación basada en el discriminante lineal depende tanto de la media de todas las observaciones dentro de cada clase, como de la matriz de covarianza dentro de cada clase calculada utilizando todas las observaciones.