Máquinas de Soporte Vectorial (SVM)

1EST17 - Aprendizaje Estadístico I

Mg. Enver Gerald Tarazona Vargas

enver.tarazona@pucp.edu.pe

Maestría en Estadística

Escuela de Posgrado

Resumen I

1. Métodos de Caja Negra
   - Máquinas de Soporte Vectorial (SVM)
     - Maximal Margin Classifier
     - Suppor Vector Classifiers
     - Suppor Vector Machines

¿Qué es un Hiperplano? I

- En el espacio p-dimensional, un hiperplano es un subespacio plano afín (no necesita pasar por el origen) de dimensión $p-1$.
- Por ejemplo, en dos dimensiones, un hiperplano es el subespacio plano de una dimensión, en otras palabras, una línea recta. En tres dimensiones, un hiperplano es el subespacio plano de dos dimensiones, el cual corresponde a un plano.
- Cuando $p > 3$ dimensiones, es difícil visualizar un hiperplano, pero la noción de un subespacio plano $(p-1)$-dimensional todavía es aplicable.
- La definición matemática de un hiperplano es simple. En dos dimensiones el hiperplano es definido por la ecuación:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0 \tag{1}$$

con parámetro β_0, β_1 y β_2.

¿Qué es un Hiperplano? II

- Cuando decimos ques (1) deifine el hiperplano, queremos decir que cualquier $X = (X_1, X_2)^T$ para el cual (1) es válido, es un punto en el hiperplano.
- Notar que (1) puede ser extendida fácilmente para p dimensiones:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0 \tag{2}$$

  define un hiperplano p-dimensional.
- Si un punto $X = (X_1, X_2, \ldots, X_p)^T$ en el espacio p-dimensional (vector de longitud p) satisface (2), entonces X cae sobre el hiperplano.
- Supongamos que X no satisface (2) sino:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p > 0 \tag{3}$$

  Esto indica que X cae a un lado del hiperplano.

¿Qué es un Hiperplano? III

- Por otro lado, si:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p < 0 \tag{4}$$

  entonces X cae en el otro lado del hiperplano.
- Por lo anterior, podemos pensar en el hiperplano como una división del espacio p-dimensional en dos mitades.
- Uno puede determinar fácilmente en que lado del hiperplano se encuentra un punto hallando el signo del lado izquierdo de (2).
- En la Figura 1 puede verse un hiperplano en el espacio de dos dimensiones.

Hiperplano en dos dimensiones

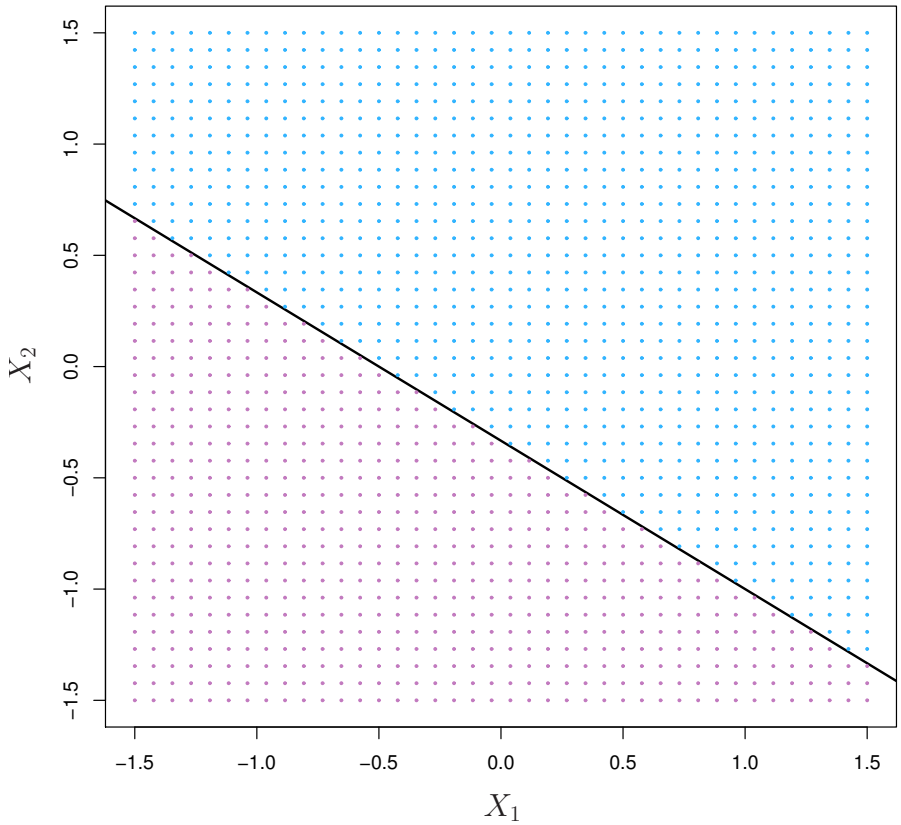

Figura: Se muestra al hiperplano $1+2X_1+3X_2 = 0$. La región azul es el conjunto de puntos para el cual $1+2X_1+3X_2 > 0$, y la región púrpura es el conjunto de puntos para el cual $1 + 2X_1 + 3X_2 < 0$

Clasificación usando Hiperplanos Separadores I

- Ahora supongamos que tenemos una matriz $\mathbf{X}$ que consiste de n observaciones de entrenamiento en el espacio p-dimensional

$$x_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1p} \end{pmatrix}, \ldots, x_n = \begin{pmatrix} x_{n1} \\ \vdots \\ x_{np} \end{pmatrix} \tag{5}$$

- Estas observaciones caen en dos clases (categorías), esto es $y_1 \ldots, y_n \in \{-1, 1\}$, donde -1 representa a una clase y 1 a la otra.
- Tenemos también a la observación de prueba $x^* = \left(x_1^*, x_2^*, \ldots, x_p^*\right)^T$
- El objetivo es desarrollar un clasificador basado en los datos de entrenamiento que clasifique correctamente a la observación de prueba usando las mediciones de sus características (variables).

Clasificación usando Hiperplanos Separadores II

- Supongamos que es posible construir un hiperplano que separe perfectamente a las observaciones de entremiento de acuerdo a sus etiquetas de clase.
- Ejemplos de tres hiperplanos separadores son mostrados el lado izquierdo del panel en la Figura 2. Podemos etiquetar a las observaciones de la clase azul como $y_i = 1$ y a las que son púrpura como $y_i = -1$.

Clasificación usando Hiperplanos Separadores III

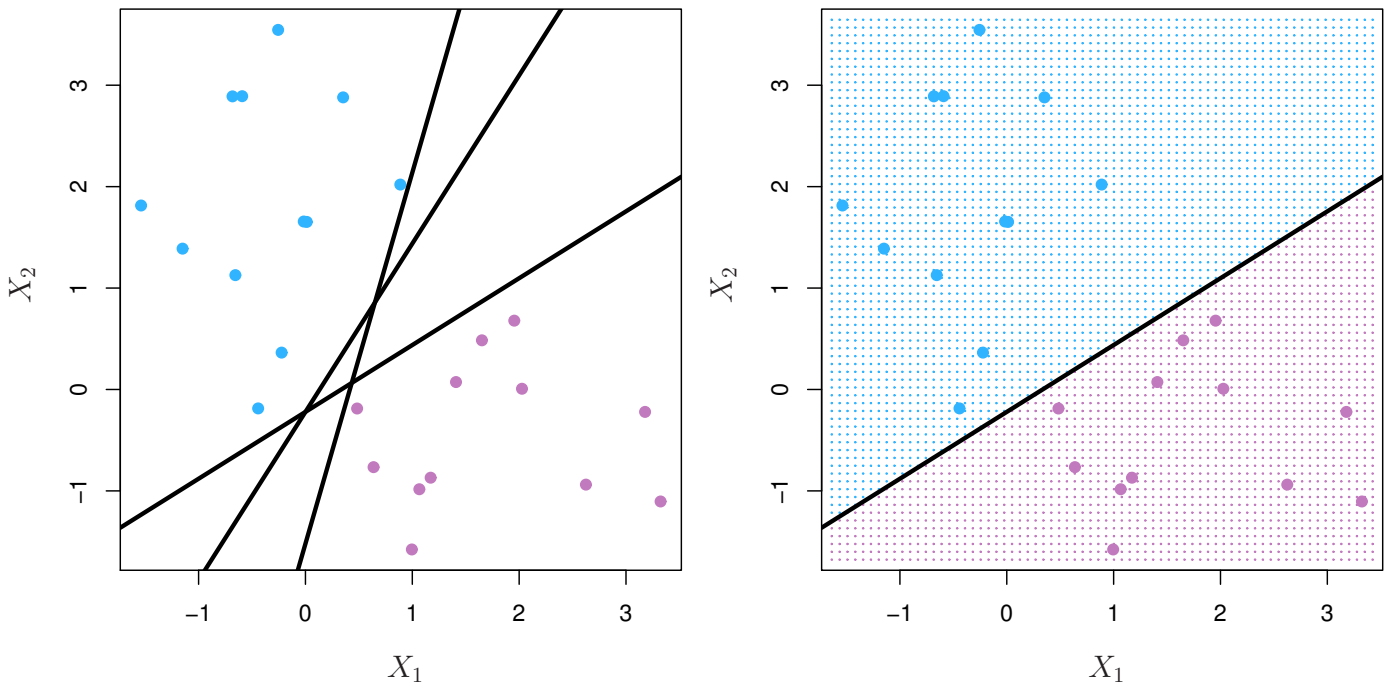

Figura: Izquierda: Hay dos clases de observaciones, mostradas en azul y púrpura, cada una con valores de dos variables. Son mostrados en negro tres hiperplanos separadores de los muchos posibles. Derecha: Un hiperplano separador es mostrado en negro. Las cuadrículas en azul y púrpura indican la regla de decisión dada por el clasificador basada en hiperplanos separadores: una observación de prueba que cae en la porción azul de la grilla será asignada a la clase azul, y una observación de prueba que cae en la porción púrpura de la grilla será asignada a la clase púrpura

Clasificación usando Hiperplanos Separadores IV

- Entonces un hiperplano separador tendrá la propiedad de

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} > 0 \text{ si } y_i = 1, \tag{6}$$

y

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} < 0 \text{ si } y_i = -1, \tag{7}$$

- De manera equivalente, un hiperplano separador tendrá la propiedad

$$y_i \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right) > 0 \tag{8}$$

para todo $i = 1, \ldots, n$

- x^* es clasificado basado en el signo de

$$f(x^*) = \left(\beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \ldots + \beta_p x_p^*\right)$$

  - Si $f(x^*)$ es positivo, la observación es asignada a la clase 1.

Clasificación usando Hiperplanos Separadores V

- Si $f(x^*)$ es negativo, la observación es asignada a la clase -1.

- También es posible usar la magnitud:
  - Si $f(x^*)$ está lejos de cero, entonces x^* cae lejos del hiperplano, por lo que podemos estar confiados acerca de la clasificación.
  - Si $f(x^*)$ es cercano a cero, entonces x^* está localizado cerca del hiperplano, por lo que estamos menos confiados acerca de la clasificación

Maximal Margin Classifier I

- Si los datos pueden ser perfectamente separados por los hiperplanos, existirán un infinito número de ellos.
- Se debe de establecer una regla para determinar cual de ellos usar.
- Una opción natural es el hiperplano de frontera máxima (maximal margin hyperplane), también conocido como el hiperplano separador óptimo, el cual es el hiperplano separador que se encuentra más lejos de las observaciones de entrenamiento.
- Podemos calcular la distancia (perpendicular) de cada observación de entrenamiento hacia un hiperplano separador.
- La menor de esas distancias es la mínima distancia de las observaciones hacia el hiperplano, también conocida como la frontera (margin).
- El hiperplano de de frontera máxima es el hiperplano separador que tiene una frontera mayor, es decir, el hiperplano que tiene la mínima distancia hacia las observaciones de entrenamiento más lejana.

Maximal Margin Classifier II

- Clasificador de frontera máxima (maximal margin classifier): Permite clasificar una observación de prueba basada en el lado en que se encuentre del hiperplano de máxima frontera.
- Se espera que el clasificador que tenga una mayor frontera para los datos de entrenamiento también lo tenga para los datos de preuba.
- Por más que este clasificador sea por lo general satisfactorio, puede llevar a casos de sobreajuste.
- Si $\beta_0, \beta_1, \ldots, \beta_p$ son los coeficientes del hiperplano de frontera máxima, entonces el clasificador de frontera máxima clasificará a la observación x^* basada en el signo de $f(x^*) = \left(\beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \ldots + \beta_p x_p^*\right)$.
- La Figura 3 muestra el hiperplano de frontera máxima para el conjunto de datos de la Figura 2.

Maximal Margin Classifier III

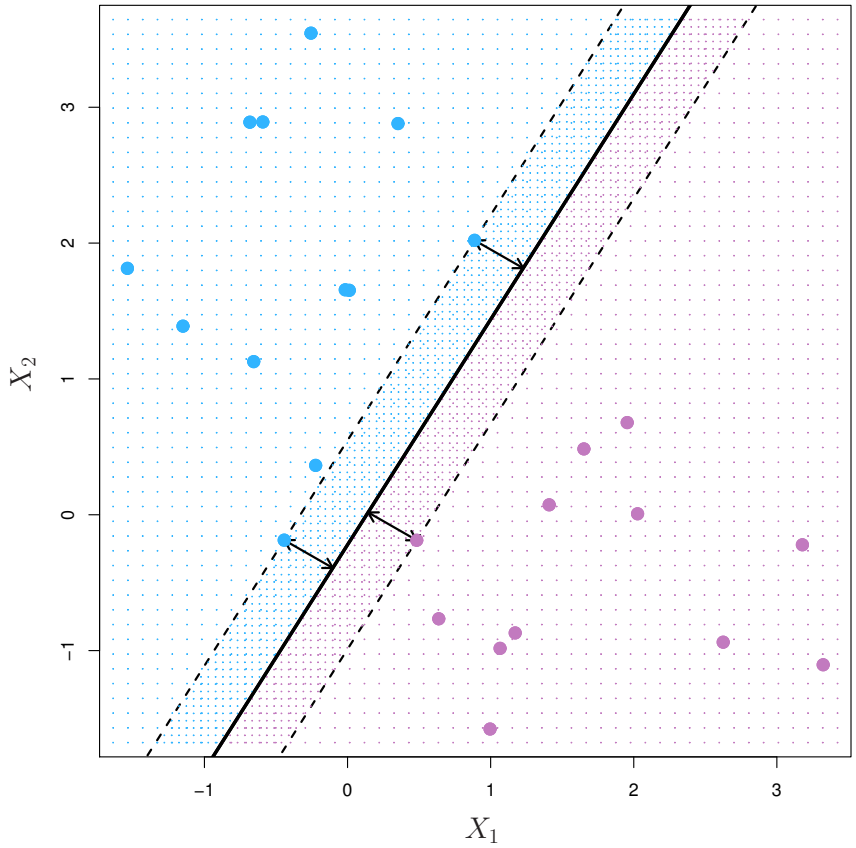

Figura: El hiperplano de frontera máxima es mostrado como una linea sólida. La frontera es la distancia de la linea sólida hacia cada una de las lineas punteadas.

Maximal Margin Classifier IV

- Examinando la Figura 3 puede observarse que hay tres puntos equidistantes del hiperplano de frontera máxima y que caen a lo largo de las lineas punteadas indicando el ancho de la frontera.
- Estas observaciones son conocidas como los vectores de soporte dado que son vectores en el espacio p-dimensional.
- Una propiedad muy importante es que el hiperplano de frontera máxima depende directamente de estos vectores de soporte, pero no de las otras observaciones.

Construción del Clasificador de Frontera Máxima I

Sean n observaciones de entrenamiento $x_1, \ldots, x_n \in \mathbb{R}^p$ con etiquetas de clase $y_1 \ldots, y_n \in \{-1, 1\}$, el hiperplano de frontera máxima es la solución al problema de optimización:

$$\underset{\beta_0, \beta_1, \ldots, \beta_p}{\text{maximizar}} M \tag{9}$$

$$\text{sujeto a } \sum_{j=1}^{p} \beta_j^2 = 1 \tag{10}$$

$$y_i \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right) \geq M \; \forall i = 1, \ldots, n \tag{11}$$

Las restricciones dadas en (10) y (10) aseguran que cada observación se encuentre en el lado correcto del hiperplano y que la distancia del hiperplano sea de al menos M (frontera del hiperplano).

Caso de no separación I

- En algunos casos no existe un hiperplano separador, por lo que no habrá un clasificador de frontera máxima.
- En esos casos el problema de optimización no tendrá solución para $M > 0$.
- Un ejemplo de esa situación es mostrado en la Figura 4 .
- Es posible extender el concepto de un hiperplano separador para desarrollar un hiperplano que separe la mayoría de las clases, usando una frontera suave.
- La generalización del clasificador de frontera máxima para el caso no separable es conocido como el clasificador de soporte vectorial (support vector classifier)

Caso de no separación II

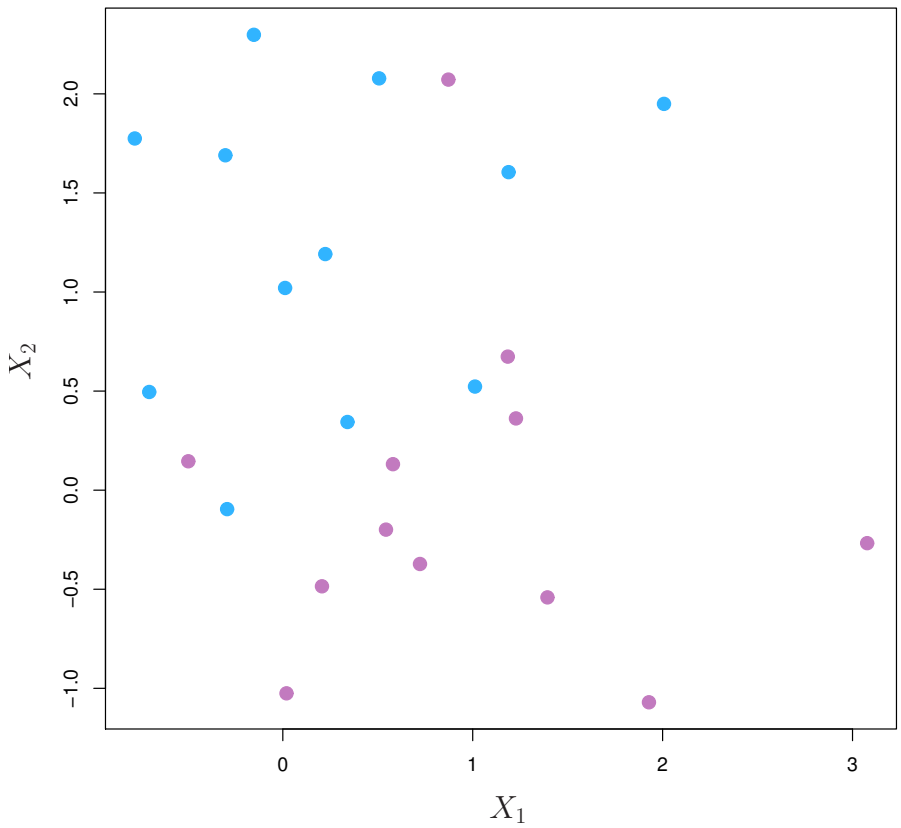

Figura: Las clases no son separables por el hiperplano

Comentarios Generales I

- Cómo se observó en la figura anterior, no todos los problemas con dos clases son perfectamente separables por un hiperplano.
- Incluso en los casos en que puedan ser separables, no siempre esto será deseable.
- Un clasificador basado en un hiperplano puede ser muy sensible al efecto de observaciones individuales (Figura 19).
- Esto es problemático dado que la distancia de una observación hacia el hiperplano es una medida de la confianza de que una observación sea correctamente clasificada.
- Esto sugiere un problema de sobreajuste.
- Por ello se puede considerar un clasificador basado en un hiperplano que no separe perfectamente las dos clases con la finalidad de
  - Mayor robustez ante el efecto de observaciones individuales.

Comentarios Generales II

- Mejor clasificación para la mayoría de las observaciones de entrenamiento.

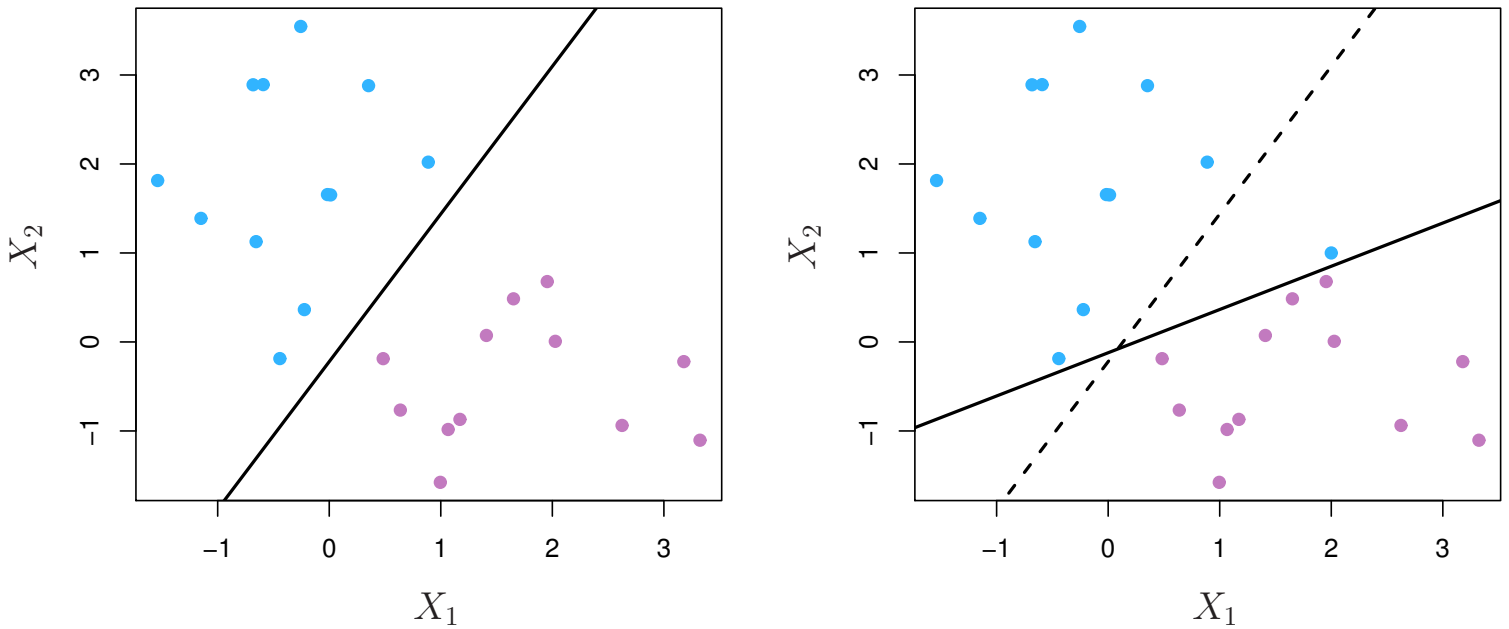

Figura: Efecto de las observaciones individuales

Clasificadores de soporte vectorial I

- Los clasificadores de soporte vectorial o clasificadores con fronteras suaves realizan esta tarea.
- En vez de buscar la mayor frontera, permite que algunas observaciones se encuentre en el lado incorrecto de la frontera e incluso del hiperplano.
- Un ejemplo es mostrado en la Figura 6.

Clasificadores de soporte vectorial II

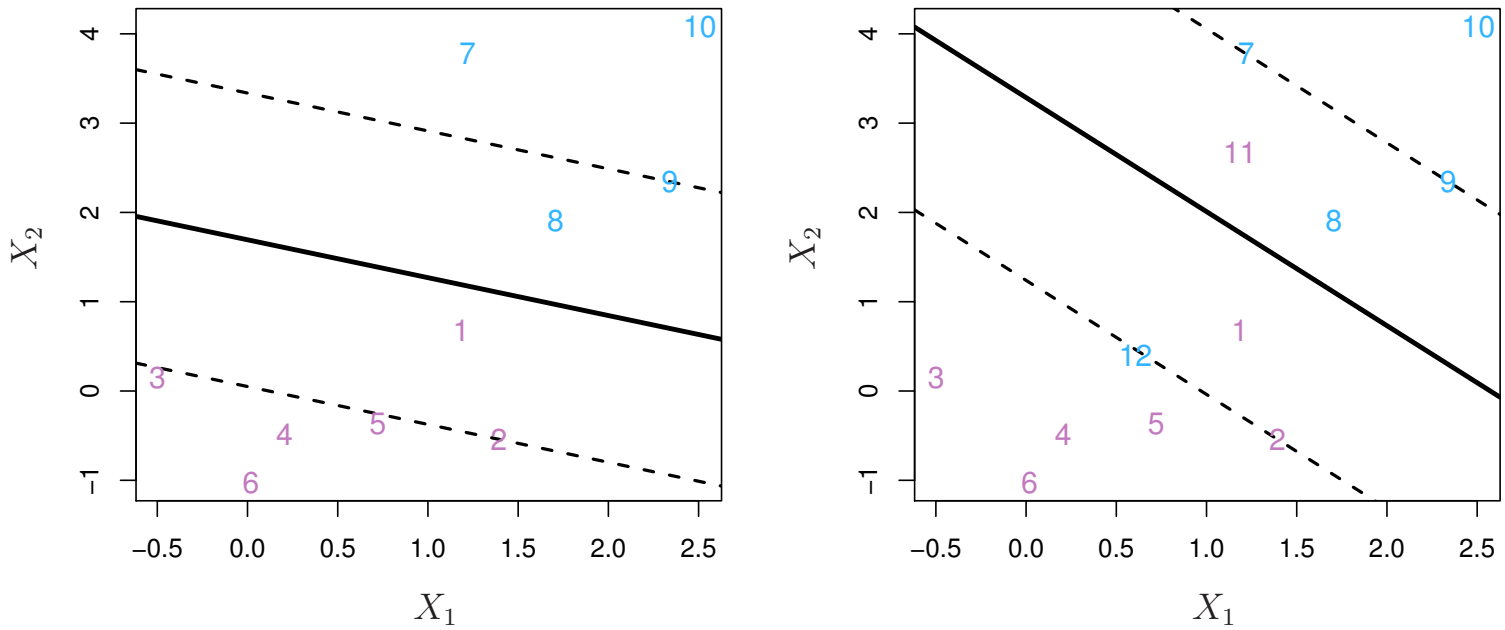

Figura: Clasificadores de Soporte Vectorial

Construción del Clasificador de Soporte Vectorial I

Sean n observaciones de entrenamiento $x_1, \ldots, x_n \in \mathbb{R}^p$ con etiquetas de clase $y_1 \ldots, y_n \in \{-1, 1\}$, el clasificador de soporte vectorial es la solución al problema de optimización:

$$\underset{\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots, \epsilon_n}{\text{maximizar}} \quad M \tag{12}$$

$$\text{sujeto a} \sum_{j=1}^{p} \beta_j^2 = 1 \tag{13}$$

$$y_i \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right) \geq M \left(1 - \epsilon_i\right) \; \forall i = 1, \ldots, n \tag{14}$$

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

Construción del Clasificador de Soporte Vectorial II

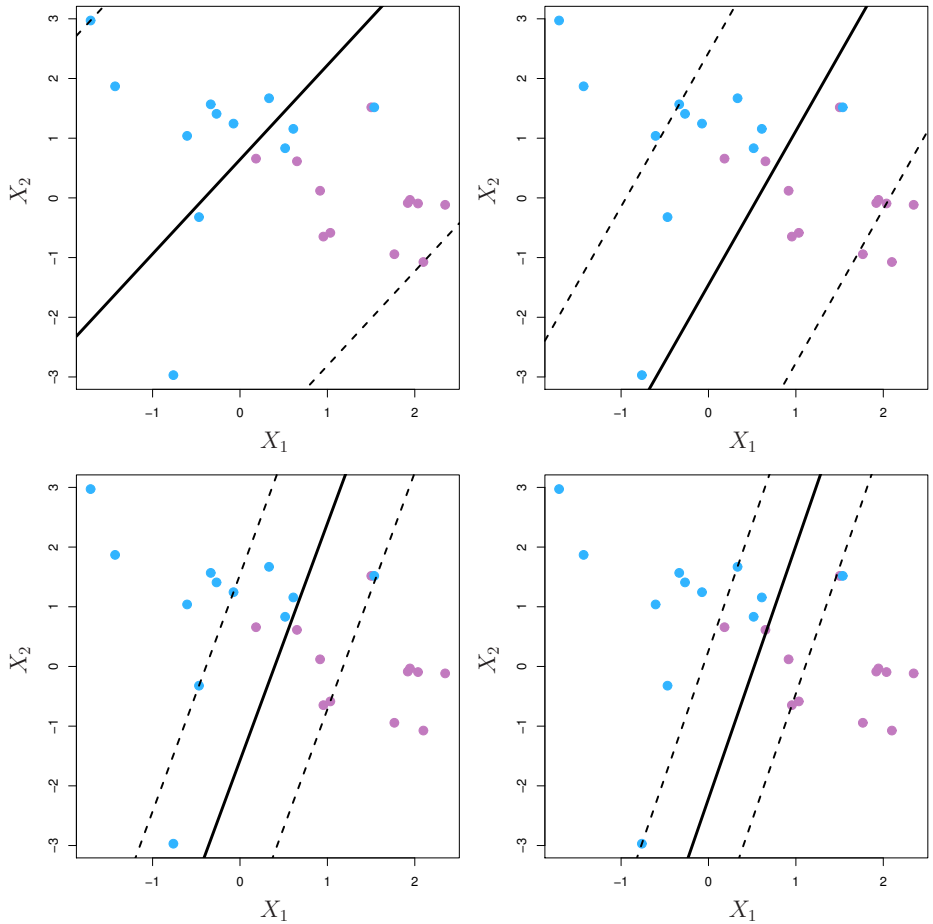

Figura: Efecto del parámetro C. A mayor valor (figura superior izquierda) mayor tolerancia de observaciones en el lado equivocado de la frontera

Comentarios Generales I

- Las Máquinas de Soporte Vectorial (SVM) consideran los casos en que los límites entre las clases no son lineales.
- En la Figura se muestra un caso en el cual los clasificadores de soporte vectorial (o cualquier otro clasificador lineal) no tendría una buena performance.

Comentarios Generales II

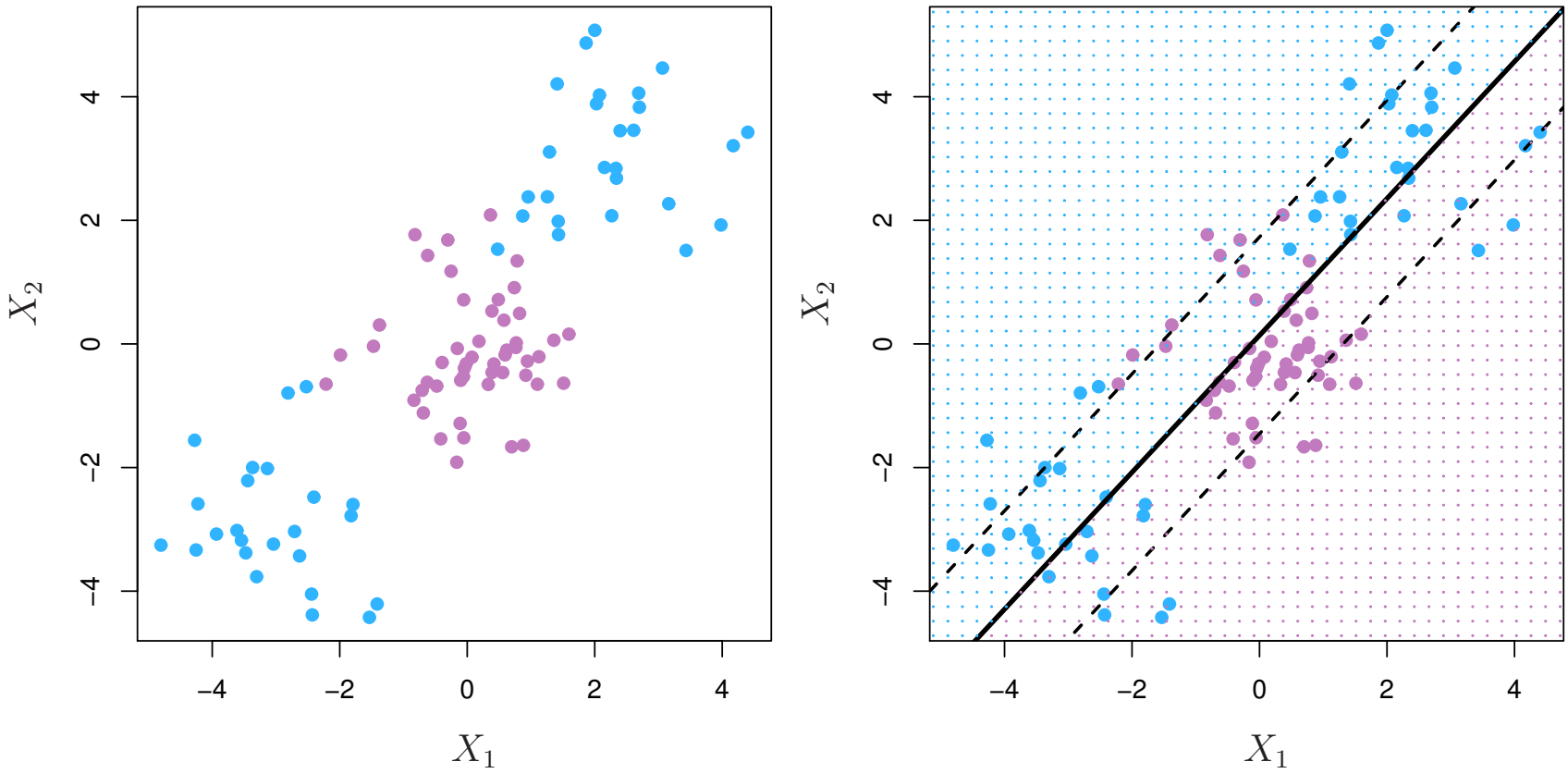

Figura: Observaciones con fronteras no lineales

Kernels I

- Los SVM son una extensión de los clasificadores de soporte vectorial que resultan de expandir el espacio de las variables de una manera específica usando kernels.
- el clasificador estará dado por

$$f(x) = \beta_0 + \sum_{i \epsilon S} \alpha_i K\left(x, x_i\right) \tag{15}$$

donde $K\left(x, x_i\right)$ es el kernel y S la colección de indicadores para los vectores de soporte.

- Kernel Lineal

$$K\left(x, x_{i'}\right) = \sum_{j=1}^{p} x_{ij} x_{i'j} \tag{16}$$

Kernels II

- Kernel Polinomial

$$K\left(x, x_{i'}\right) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d \tag{17}$$

- Kernel Radial

$$K\left(x, x_{i'}\right) = \exp\left(-\gamma \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2\right) \tag{18}$$

Kernels III

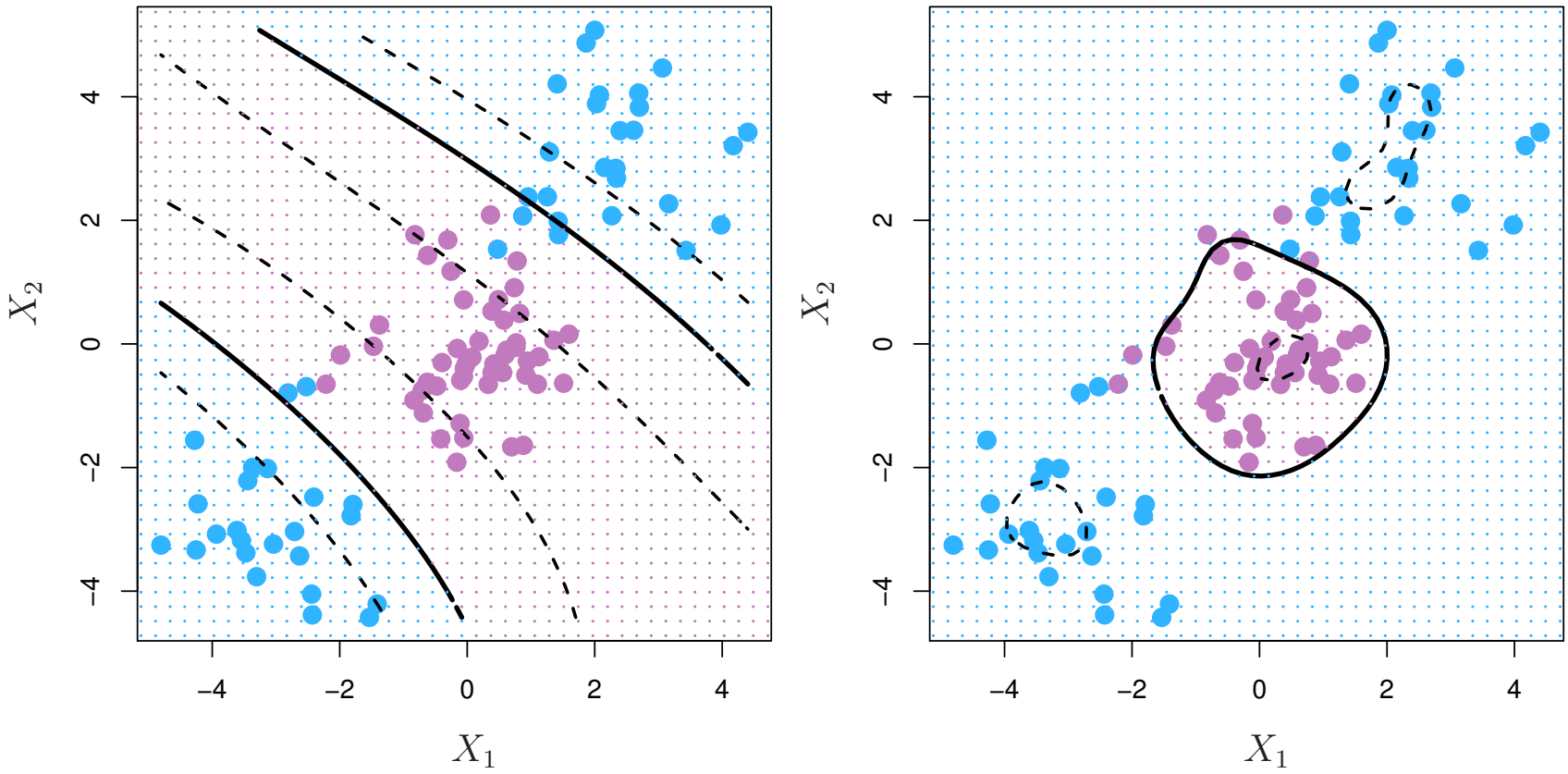

Figura: SVM con kernel polinomial de grado 3 (izquierda) y radial (derecha)

SVM para más de dos categorías I

Cuando $K > 2$ se suelen usar dos alternativas:

- **Clasificación Uno contra Uno**: Se construyen $\begin{pmatrix} K \\ 2 \end{pmatrix}$ SVMs, cada uno de ellos comparando un par de clases. Por ejemplo, un SVM podría comparar la k-ésima clase codificada como $+1$ con la j-ésima cláse codificada como -1. Se clasifica un observación usando cada uno de los clasificadores y se cuenta el número de veces que la observación es asignada a cada una de las K clases. La clasificación final es realizada asignando a la observación de prueba la clase a la que fue asignada con mayor frecuencia.
- **Clasificación Uno contra Todos**: Se ajustan K SVMs, en cada caso comparando la clase K con las restantes $K - 1$ clases. Se asigna la clase al clasificador que tenga mayor valor.