Máquinas de Vectores de Soporte

Alvaro J. Riascos Villegas

Junio 2018

Contenido

Introducción

- SVM es una técnica que generaliza las técnicas de hiperplanos separadores.
- Construye hiperplanos separadores en un espacio de variables transformado por funciones no lineales. Estas funciones llevan los features a espacios potencialmente mucho más grandes.
- SVM es una de las técnicas más estudiadas actualmente por ser un problema convexo y se conoce la dimensión VC del clasificador.

Introducción

- SVM es una técnica que generaliza las técnicas de hiperplanos separadores.
- Construye hiperplanos separadores en un espacio de variables transformado por funciones no lineales. Estas funciones llevan los features a espacios potencialmente mucho más grandes.
- SVM es una de las técnicas más estudiadas actualmente por ser un problema convexo y se conoce la dimensión VC del clasificador.

Introducción

- SVM es una técnica que generaliza las técnicas de hiperplanos separadores.
- Construye hiperplanos separadores en un espacio de variables transformado por funciones no lineales. Estas funciones llevan los features a espacios potencialmente mucho más grandes.
- SVM es una de las técnicas más estudiadas actualmente por ser un problema convexo y se conoce la dimensión VC del clasificador.

Contenido

Hiperplano separador óptimo

- Los datos de entrenamiento son $\{(x_1, y_1), ..., (x_N, y_N)\}, x_i \in R^p, y_i \in \{-1, 1\}$.
- Un hiperplano separador es un conjunto de la forma: $\{x : f(x) = x^T \beta + \beta_0 = 0\}$. Donde β se puede normalizar a que tenga norma 1.
- La regla de clasificación que induce el hiperplano es: $sign[x^T \beta + \beta_0]$.
- En la figura los datos de entrenamiento son perfectamente separables por el hiperplano y se puede garantizar que para todo i, $f(x_i)y_i > 0$.

Hiperplano separador óptimo

- Los datos de entrenamiento son $\{(x_1, y_1), ..., (x_N, y_N)\}, x_i \in R^p, y_i \in \{-1, 1\}$.
- Un hiperplano separador es un conjunto de la forma: $\{x : f(x) = x^T\beta + \beta_0 = 0\}$. Donde β se puede normalizar a que tenga norma 1.
- La regla de clasificación que induce el hiperplano es: $sign[x^T\beta + \beta_0]$.
- En la figura los datos de entrenamiento son perfectamente separables por el hiperplano y se puede garantizar que para todo i, $f(x_i)y_i > 0$.

Hiperplano separador óptimo

- Los datos de entrenamiento son $\{(x_1, y_1), ..., (x_N, y_N)\}, x_i \in R^p, y_i \in \{-1, 1\}$.
- Un hiperplano separador es un conjunto de la forma: $\{x : f(x) = x^T\beta + \beta_0 = 0\}$. Donde β se puede normalizar a que tenga norma 1.
- La regla de clasificación que induce el hiperplano es: $sign[x^T\beta + \beta_0]$.
- En la figura los datos de entrenamiento son perfectamente separables por el hiperplano y se puede garantizar que para todo i, $f(x_i)y_i > 0$.

Hiperplano separador óptimo

- Los datos de entrenamiento son $\{(x_1, y_1), ..., (x_N, y_N)\}, x_i \in R^p, y_i \in \{-1, 1\}$.
- Un hiperplano separador es un conjunto de la forma: $\{x : f(x) = x^T\beta + \beta_0 = 0\}$. Donde β se puede normalizar a que tenga norma 1.
- La regla de clasificación que induce el hiperplano es: $sign[x^T\beta + \beta_0]$.
- En la figura los datos de entrenamiento son perfectamente separables por el hiperplano y se puede garantizar que para todo i, $f(x_i)y_i > 0$.

Hiperplano separador óptimo

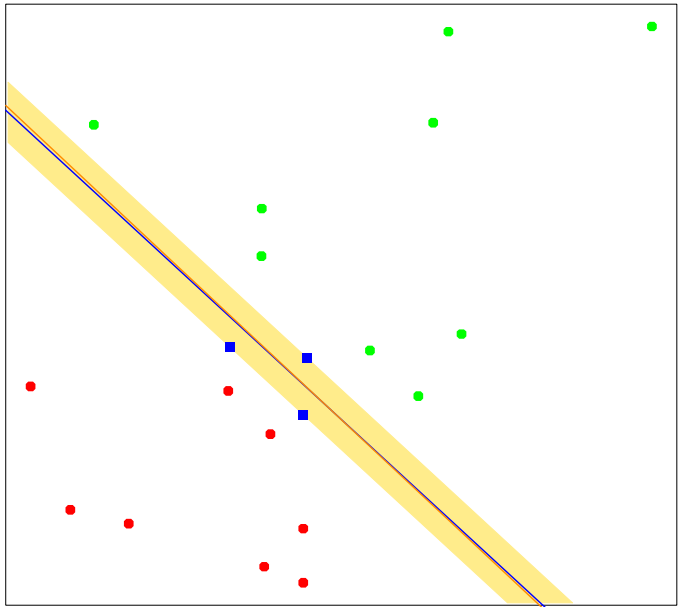


FIGURE 4.16. *The same data as in Figure 4.14. The shaded region delineates the maximum margin separating the two classes. There are three support points indicated, which lie on the boundary of the margin, and the optimal separating hyperplane (blue line) bisects the slab. Included in the figure is the boundary found using logistic regression (red line), which is very close to the optimal separating hyperplane (see Section 12.3.3).*

Hiperplano separador óptimo

- Si $\mid \beta \mid = 1$, la función $f(x) = x^T\beta + \beta_0$ nos da la distancia (con signo) al hiperplano.
- Para ver esto sea x_0 la proyección perpendicular al hiperplano. Entonces: $(x - x_0).\beta = \mid x - x_0 \mid$ excepto por el signo y:

$$(x - x_0).\beta = x^T\beta + \beta_0 - x^T\beta - x_0 = x^T\beta + \beta_0$$

- Luego el siguiente problema refleja la restricción de clasificar perfectamente y que el márgen M, sea máximo:

$$\max_{\beta,\beta_0,M} M$$
$$\mid \beta \mid = 1$$
$$y_i(x_i^T\beta + \beta_0) \geq M$$

Hiperplano separador óptimo

- Si $\mid \beta \mid = 1$, la función $f(x) = x^T\beta + \beta_0$ nos da la distancia (con signo) al hiperplano.
- Para ver esto sea x_0 la proyección perpendicular al hiperplano. Entonces: $(x - x_0).\beta = \mid x - x_0 \mid$ excepto por el signo y:

$$(x - x_0).\beta = x^T\beta + \beta_0 - x^T\beta - x_0 = x^T\beta + \beta_0$$

- Luego el siguiente problema refleja la restricción de clasificar perfectamente y que el márgen M, sea máximo:

$$\max_{\beta,\beta_0,M} M$$
$$\mid \beta \mid = 1$$
$$y_i(x_i^T\beta + \beta_0) \geq M$$

Hiperplano separador óptimo

- Si $\mid \beta \mid = 1$, la función $f(x) = x^T\beta + \beta_0$ nos da la distancia (con signo) al hiperplano.
- Para ver esto sea x_0 la proyección perpendicular al hiperplano. Entonces: $(x - x_0).\beta = \mid x - x_0 \mid$ excepto por el signo y:

$$(x - x_0).\beta = x^T\beta + \beta_0 - x^T\beta - x_0 = x^T\beta + \beta_0$$

- Luego el siguiente problema refleja la restricción de clasificar perfectamente y que el márgen M, sea máximo:

$$\begin{aligned} &\max_{\beta,\beta_0,M} M \\ &\mid \beta \mid = 1 \\ &y_i(x_i^T\beta + \beta_0) \geq M \end{aligned}$$

Hiperplano separador óptimo

- Este problema se puede escribir como:

$$\max_{\beta,\beta_0,M} M$$

$$y_i(x_i^T\beta + \beta_0 \mid \beta \mid) \geq M \mid \beta \mid$$

- Que es equivalente a:

$$\max_{\beta,\beta_0,M} M$$

$$y_i(x_i^T\beta + \beta_0) \geq M \mid \beta \mid$$

Hiperplano separador óptimo

- Este problema se puede escribir como:

$$\max_{\beta,\beta_0,M} M$$

$$y_i(x_i^T\beta + \beta_0 \mid \beta \mid) \geq M \mid \beta \mid$$

- Que es equivalente a:

$$\max_{\beta,\beta_0,M} M$$

$$y_i(x_i^T\beta + \beta_0) \geq M \mid \beta \mid$$

Hiperplano separador óptimo

- Que es equivalente a:

$$\max_{\beta,\beta_0,M,M'} M$$
$$y_i(x_i^T\beta + \beta_0) \geq M'$$
$$M' = M \mid \beta \mid$$

Hiperplano separador óptimo

- Que es equivalente a:

$$\max_{\beta,\beta_0,M'} \frac{M'}{\mid \beta \mid}$$
$$y_i(x_i^T \beta + \beta_0) \geq M'$$

- Equivalente a:

$$\max_{\beta,\beta_0,M'} \frac{1}{\frac{|\beta|}{M'}}$$
$$y_i(x_i^T \frac{\beta}{M'} + \beta_0) \geq 1$$

- Obsérvese que podemos elegir $M' = 1$

Hiperplano separador óptimo

- Que es equivalente a:

$$\max_{\beta,\beta_0,M'} \frac{M'}{\mid \beta \mid}$$
$$y_i(x_i^T \beta + \beta_0) \geq M'$$

- Equivalente a:

$$\max_{\beta,\beta_0,M'} \frac{1}{\frac{|\beta|}{M'}}$$
$$y_i(x_i^T \frac{\beta}{M'} + \beta_0) \geq 1$$

- Obsérvese que podemos elegir $M' = 1$

Hiperplano separador óptimo

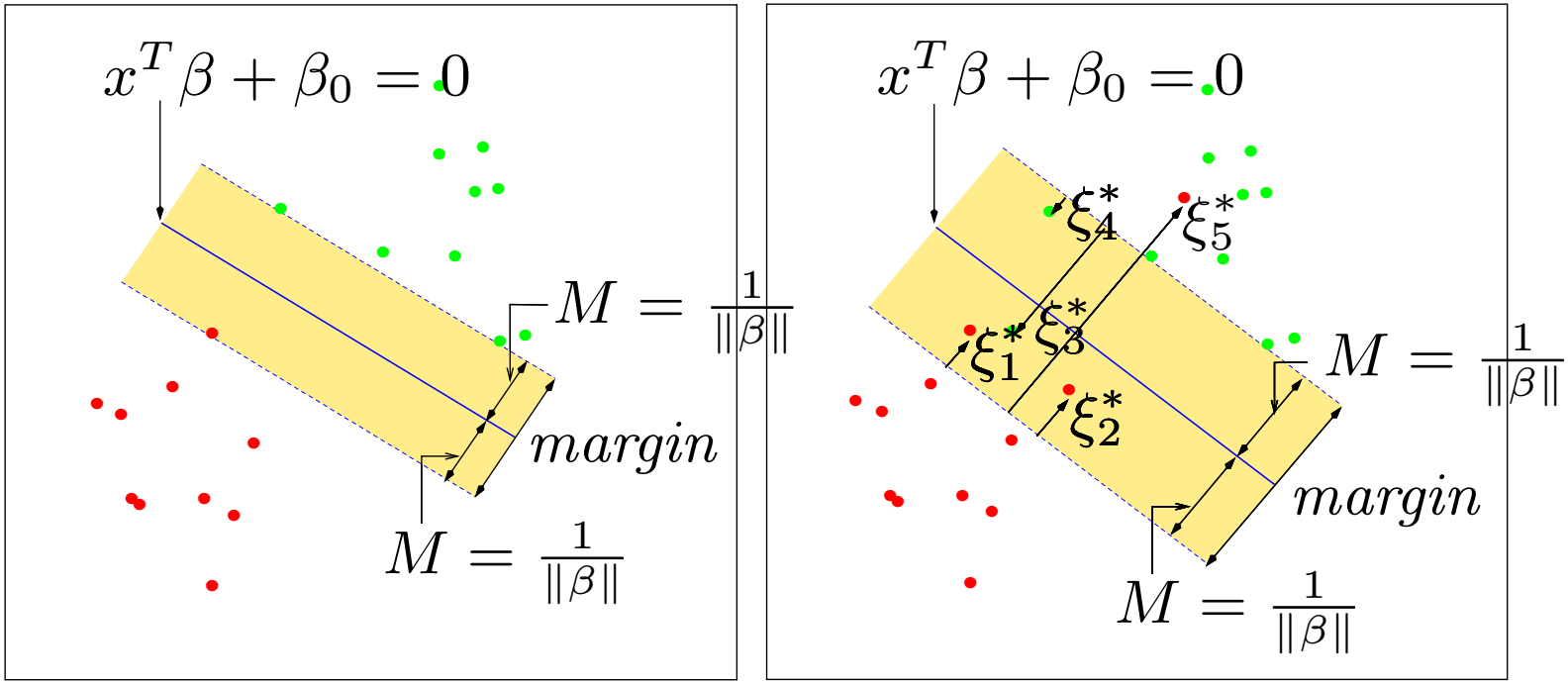
- Finalmente:

$$\min_{\beta,\beta_0} \mid \beta \mid$$

$$y_i(x_i^T \beta + \beta_0) \geq 1$$

  donde $M = \frac{1}{|\beta|}$.
- Panel izquierdo.

Hiperplano separador óptimo



Hiperplano separador óptimo: CPO

- El Lagrangiano del problema es:

$$L = \frac{1}{2}(\mid \beta \mid)^2 + \sum_{i=1}^{n}(\alpha_i(x_i^T \beta + \beta_0 - 1)$$

- Las CPO:

$$\beta = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$0 = \sum_{i=1}^{n} \alpha_i y_i$$

$$\alpha_i(x_i^T \beta + \beta_0 - 1) = 0$$

Hiperplano separador óptimo: CPO

- El Lagrangiano del problema es:

$$L = \frac{1}{2}(\mid \beta \mid)^2 + \sum_{i=1}^{n}(\alpha_i(x_i^T \beta + \beta_0 - 1)$$

- Las CPO:

$$\beta = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$0 = \sum_{i=1}^{n} \alpha_i y_i$$

$$\alpha_i(x_i^T \beta + \beta_0 - 1) = 0$$

Hiperplano separador óptimo: CPO

- Obsérvese que: $\alpha_i > 0$ si y solo si (x_i, y_i) está en el borde de la región de separción.
- β está definido únicamente por los puntos que están en el borde de la región. Estos puntos se llaman vectores de soporte.
- Panel izquierdo figura.

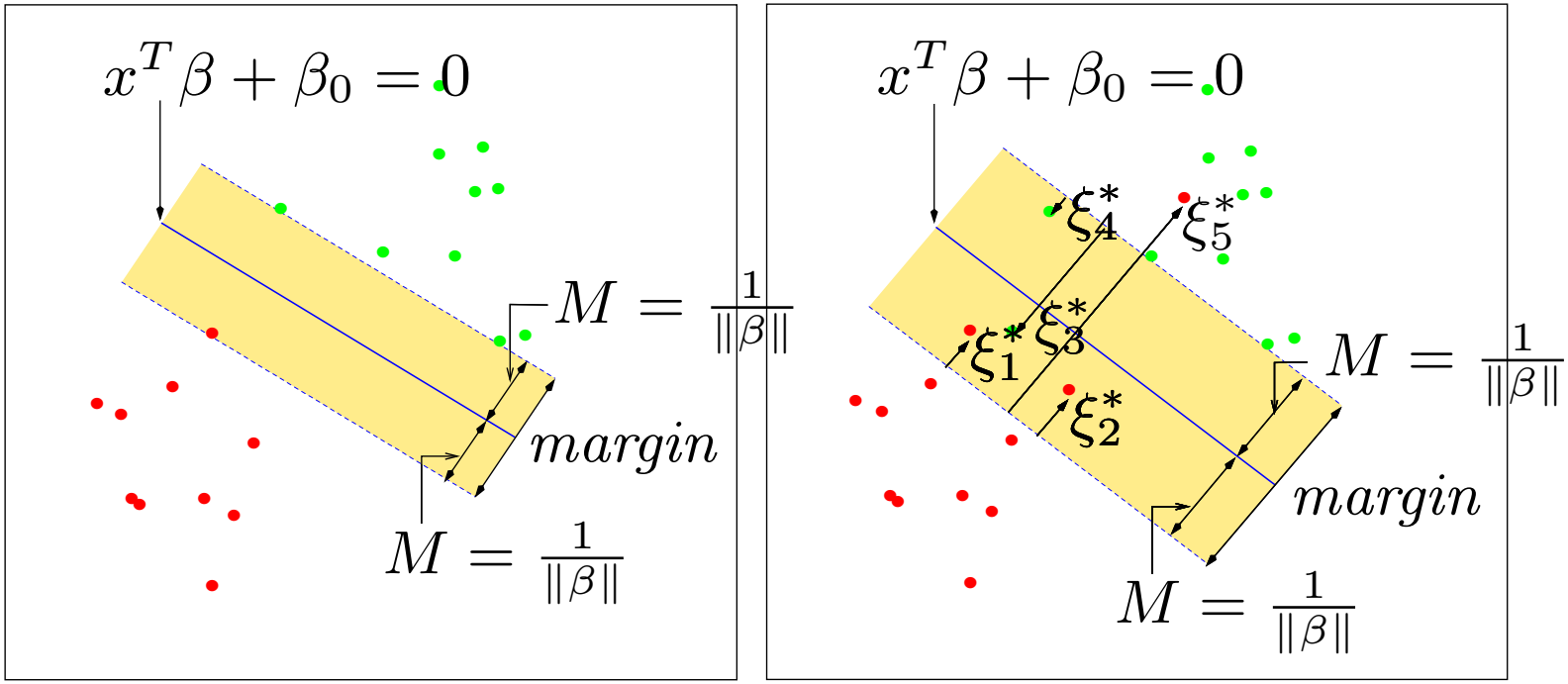
Hiperplano separador óptimo: CPO

- Obsérvese que: $\alpha_i > 0$ si y solo si (x_i, y_i) está en el borde de la región de separción.
- β está definido únicamente por los puntos que están en el borde de la región. Estos puntos se llaman vectores de soporte.
- Panel izquierdo figura.

Hiperplano separador óptimo: CPO

- Obsérvese que: $\alpha_i > 0$ si y solo si (x_i, y_i) está en el borde de la región de separción.
- β está definido únicamente por los puntos que están en el borde de la región. Estos puntos se llaman vectores de soporte.
- Panel izquierdo figura.

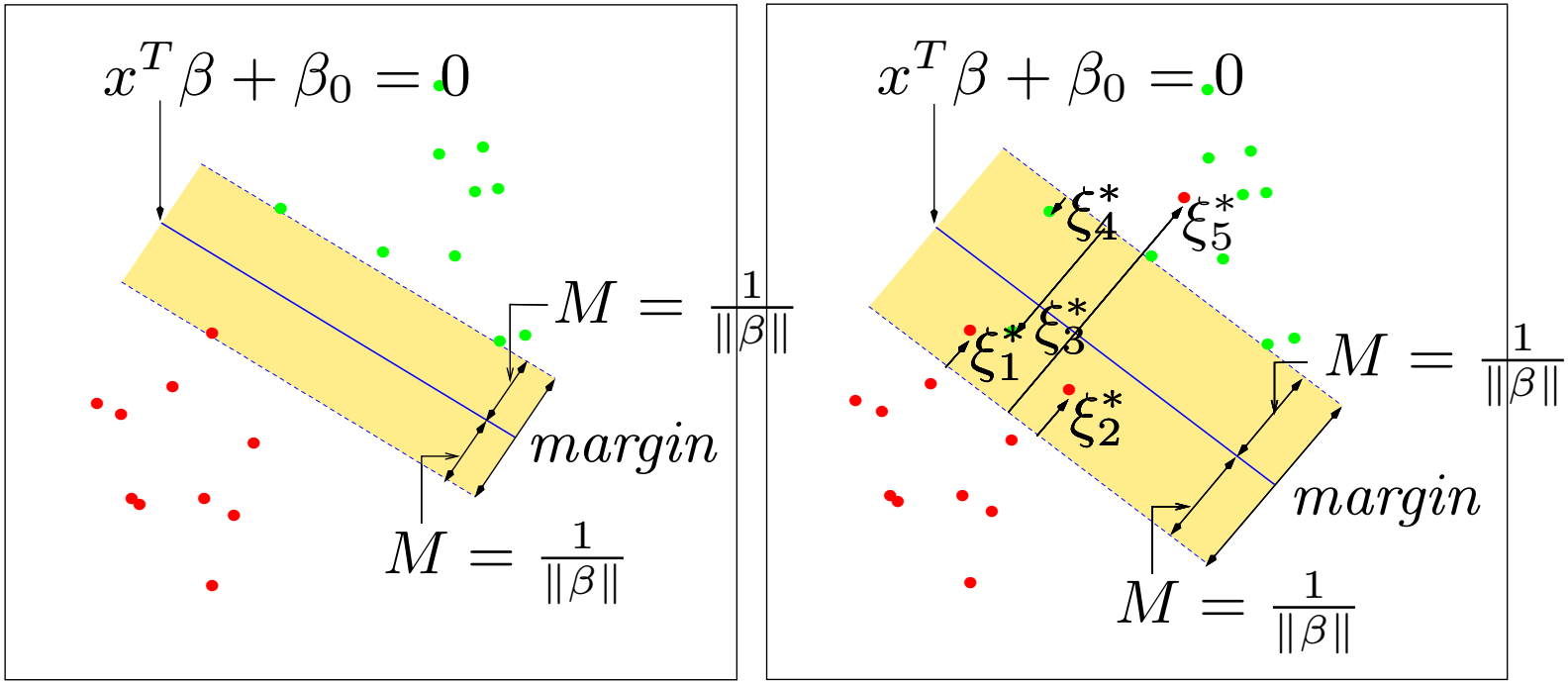
Hiperplano separador óptimo



Hiperplano separador óptimo: clasificación imperfecta

- Aun cuando se pueda separar perfectamente puede no ser deseable. El hiperplano puede ser muy sensible a nuevos datos (i.e., varianza alta).
- Panel derecho figura.

Hiperplano separador óptimo



Contenido

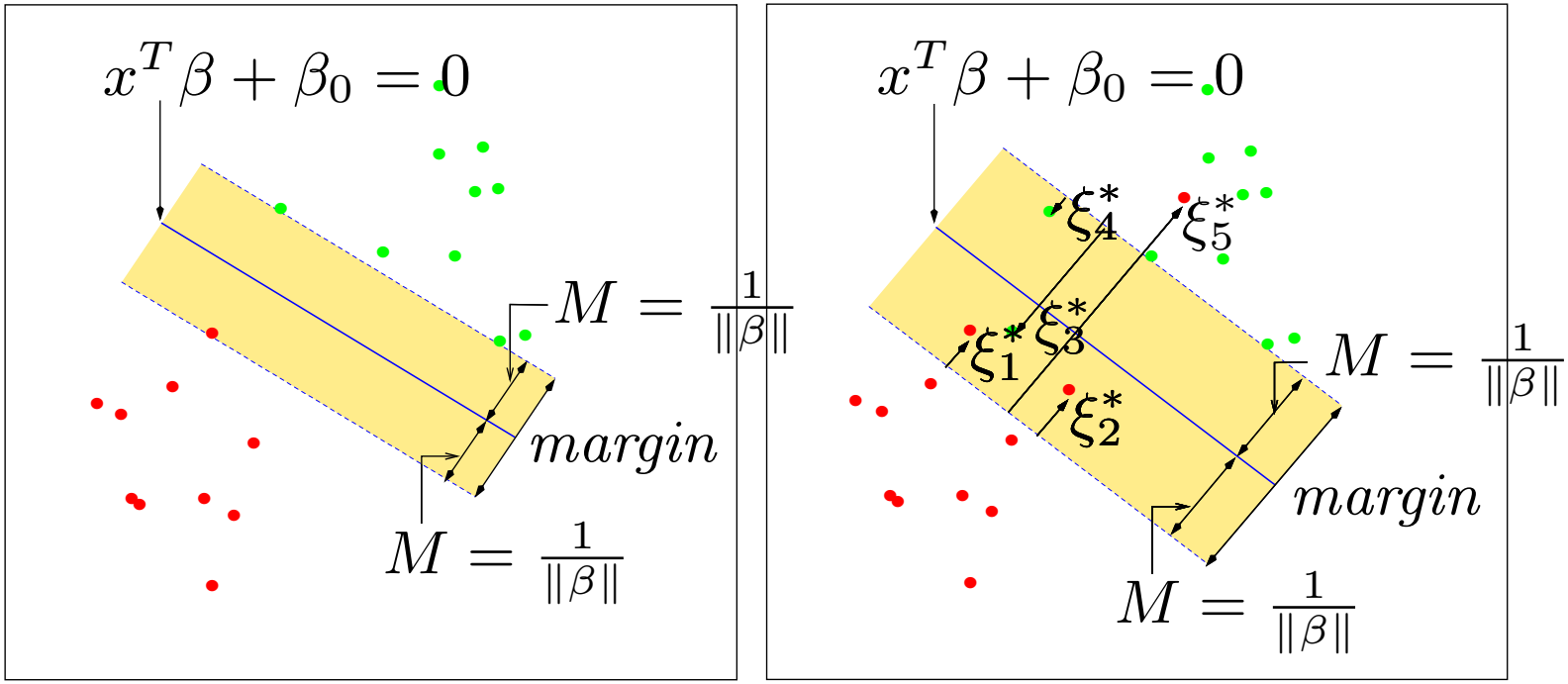
Hiperplano separador óptimo: clasificación imperfecta

- Cuando los datos de entrenamiento no son perfectamente separables, o no es deseable la separación perfecta por un hiperplano, se puede formular un problema similar:

$$\begin{aligned} &\max_{\beta,\beta_0,\xi,M} M \\ &| \beta | = 1 \\ &y_i(x_i^T \beta + \beta_0) \geq M - \xi_i \\ &\xi_i \geq 0 \\ &\sum_{i=1}^{n} \xi_i \leq Const \end{aligned}$$

- Panel derecho figura.

Hiperplano separador óptimo



Hiperplano separador óptimo: clasificación imperfecta

- O aún otra forma de hacerlo es:

$$\max_{\beta,\beta_0,\xi,M} M$$
$$| \beta | = 1$$
$$y_i(x_i^T \beta + \beta_0) \geq M(1 - \xi_i)$$

  donde $\xi_i \geq 0$ y $\sum_i \xi_i \leq Const$, $Const$ es una constante.

- El primero es más natural pero no resulta ser un problema convexo. El segundo sí lo es.
- Vamos a ver que, en la versión convexa del problema:
  1. Cuando $\xi > 1$ hay un error de clasificación.
  2. Cuando $0 \leq \xi \leq 1$ el punto está a una distancia del hiperplano menor o igual al márgen.

Hiperplano separador óptimo: clasificación imperfecta

- O aún otra forma de hacerlo es:

$$\max_{\beta,\beta_0,\xi,M} M$$
$$|\beta| = 1$$
$$y_i(x_i^T\beta + \beta_0) \geq M(1-\xi_i)$$

  donde $\xi_i \geq 0$ y $\sum_i \xi_i \leq Const$, $Const$ es una constante.
- El primero es más natural pero no resulta ser un problema convexo. El segundo sí lo es.
- Vamos a ver que, en la versión convexa del problema:
  1. Cuando $\xi > 1$ hay un error de clasificación.
  2. Cuando $0 \leq \xi \leq 1$ el punto está a una distancia del hiperplano menor o igual al márgen.

Hiperplano separador óptimo: clasificación imperfecta

- O aún otra forma de hacerlo es:

$$\max_{\beta,\beta_0,\xi,M} M$$
$$| \beta | = 1$$
$$y_i(x_i^T \beta + \beta_0) \geq M(1 - \xi_i)$$

  donde $\xi_i \geq 0$ y $\sum_i \xi_i \leq Const$, $Const$ es una constante.
- El primero es más natural pero no resulta ser un problema convexo. El segundo sí lo es.
- Vamos a ver que, en la versión convexa del problema:
  1. Cuando $\xi > 1$ hay un error de clasificación.
  2. Cuando $0 \leq \xi \leq 1$ el punto está a una distancia del hiperplano menor o igual al márgen.

Hiperplano separador óptimo

- Un argumento similar al presentado anteriormente para el caso de separción perfecta muestra que el problema es equivalente a:

$$\begin{aligned} \min_{\beta,\beta_0} \mid \beta \mid \\ y_i(x_i^T \beta + \beta_0) \geq 1 - \xi \\ \xi \geq 0 \\ \sum_i \xi_i \leq \mathit{Const} \end{aligned}$$

donde $M = \frac{1}{|\beta|}$ y Const es una constante.

Hiperplano separador óptimo: CPO

- Planteamos el problema como:

$$\min_{\beta,\beta_o} \frac{1}{2}(\mid \beta \mid)^2 + C\sum_{i=1}^{n} \xi_i$$
$$y_i(x_i^T \beta + \beta_0) \geq 1 - \xi$$
$$\xi \geq 0$$

donde C es una constante de penalización.

Hiperplano separador óptimo: CPO

- Al hacer las condiciones de primer orden con respecto a β, β_0, ξ_i se puede demostrar que:

$$\beta_i = \sum_{i=1}^{n} \alpha_i y_i x_i$$

y α es distinto de cero únicamente cuando:

$$y_i(x_i^T \beta + \beta_0) - (1 - \xi_i) = 0$$

es decir, cuando el ejemplo i esta justo en el **márgen o en lado equivocado dependiendo de** ξ_i. Estos son los vectores de soporte en este caso.

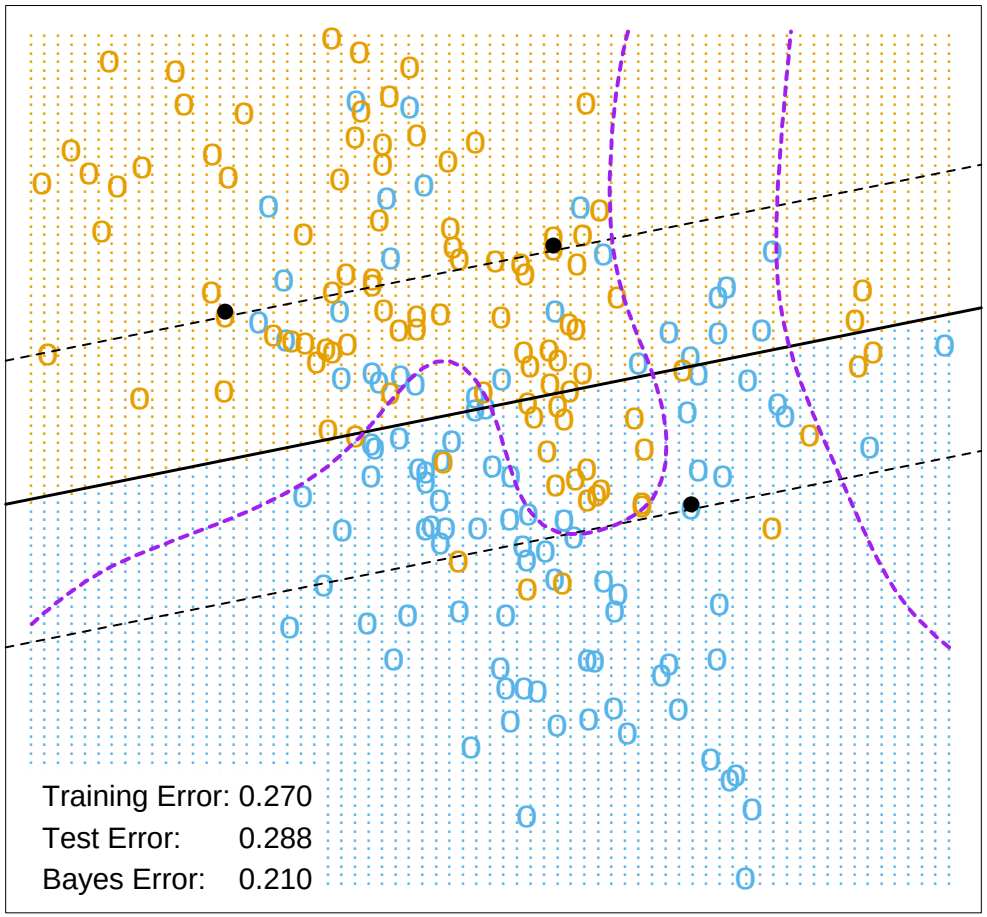
Hiperplano separador óptimo

- La función de aprendizaje es:

$$G(x) = sign[x^T \beta + \beta_0]$$

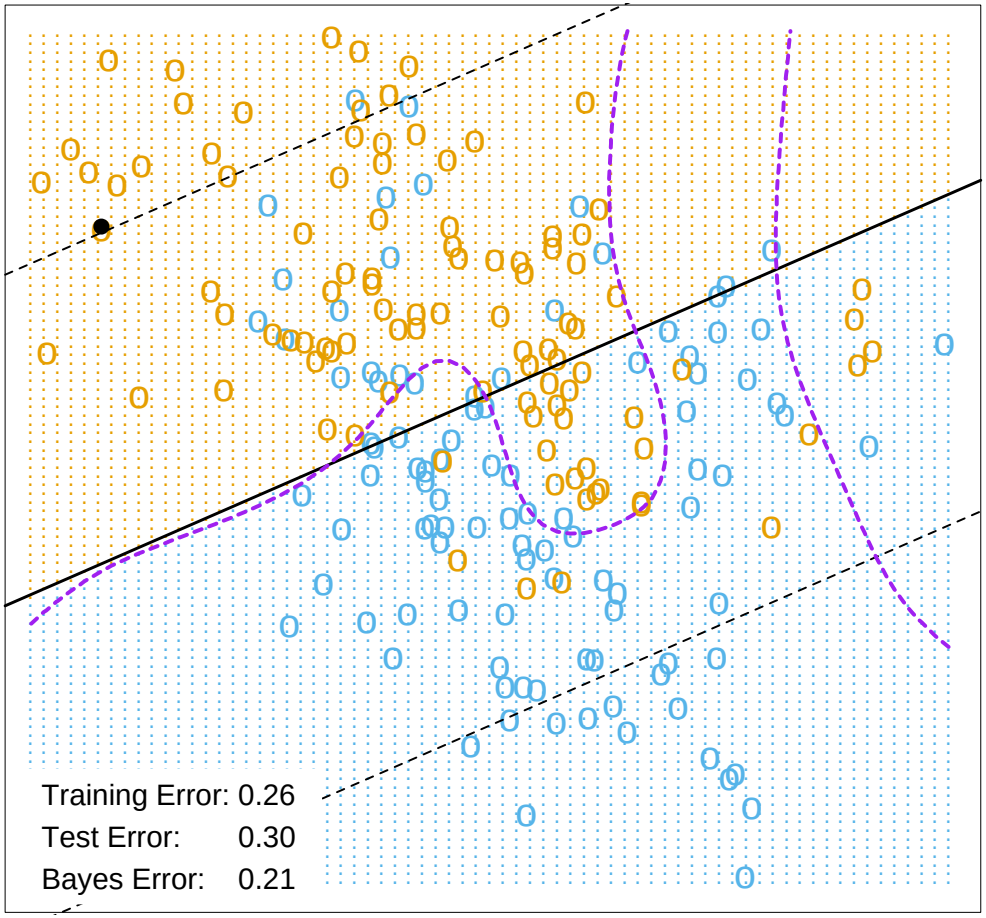
donde C es un meta parámetro.

SVM, C=10.000



$$C = 10000$$

SVM, C=0.01



$$C = 0.01$$

Contenido

Lagrangiano dual de Wolfe

- Una forma equivalente al problema de Lagrange introducido anteriormente (para el problema de separción imperfecta o SVM) es usar el Lagrangiano dual de Wolfe del problema (váse The Elements of Statitical Learning, página 420):

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

- En este caso la funció $G(x)$ se puede escribir como:

$$G(x) = sign[x^T \sum_{i}^{N} \alpha_i y_i x_i^T + \beta_0]$$

- Esta representación sugiere una generalización adicional.

Lagrangiano dual de Wolfe

- Una forma equivalente al problema de Lagrange introducido anteriormente (para el problema de separción imperfecta o SVM) es usar el Lagrangiano dual de Wolfe del problema (váse The Elements of Statitical Learning, página 420):

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

- En este caso la funció $G(x)$ se puede escribir como:

$$G(x) = sign[x^T \sum_{i}^{N} \alpha_i y_i x_i^T + \beta_0]$$

- Esta representación sugiere una generalización adicional.

Lagrangiano dual de Wolfe

- Una forma equivalente al problema de Lagrange introducido anteriormente (para el problema de separción imperfecta o SVM) es usar el Lagrangiano dual de Wolfe del problema (váse The Elements of Statitical Learning, página 420):

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

- En este caso la funció $G(x)$ se puede escribir como:

$$G(x) = sign[x^T \sum_{i}^{N} \alpha_i y_i x_i^T + \beta_0]$$

- Esta representación sugiere una generalización adicional.

Kernels

- Considere ahora el Lagrangiano

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle h(x_i), h(x_j) \rangle)$$

donde h son funciones de R^p en R^d donde d puede ser ∞.

- En este caso la funció$f(x)$ se puede escribir como:

$$G(x) = sign[\sum_{i=1}^{N} \alpha_i y_i \langle h(x), h(x_i) \rangle + \beta_0]$$

donde $\langle , \rangle$ denotan un producto punto.

Kernels

- Considere ahora el Lagrangiano

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle h(x_i), h(x_j) \rangle)$$

donde h son funciones de R^p en R^d donde d puede ser ∞.

- En este caso la funció$f(x)$ se puede escribir como:

$$G(x) = sign[\sum_{i=1}^{N} \alpha_i y_i \langle h(x), h(x_i) \rangle + \beta_0]$$

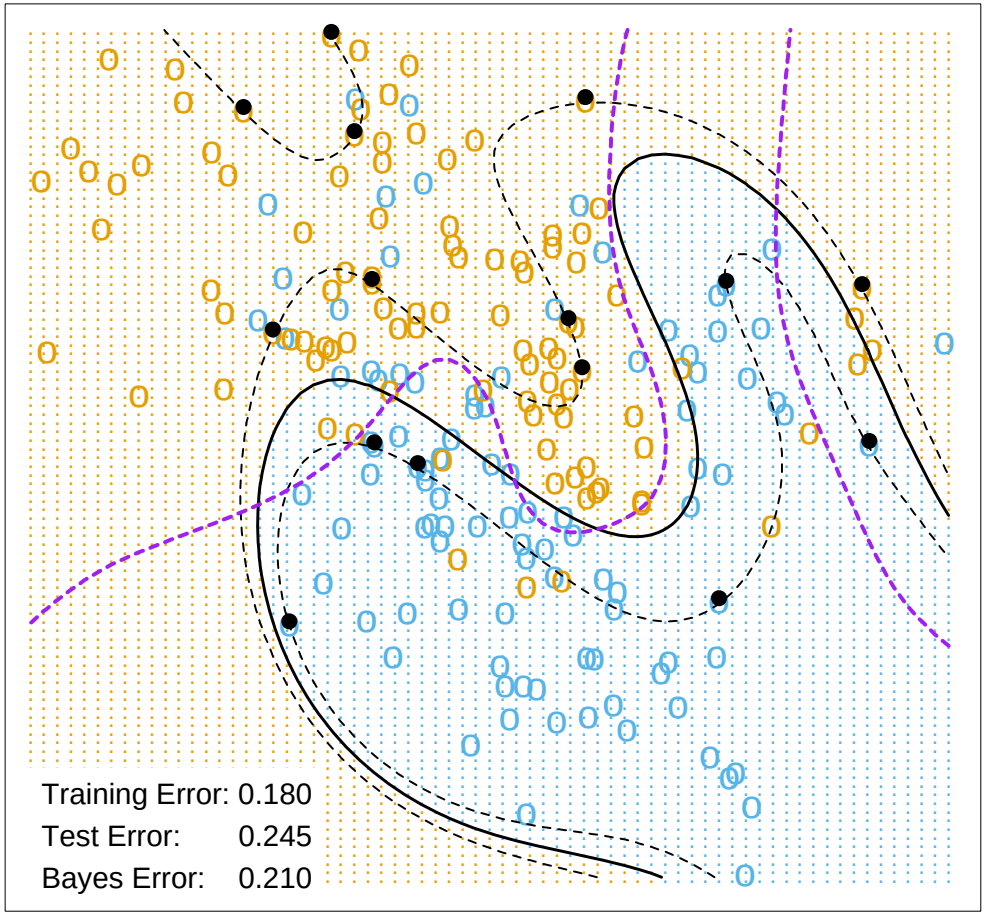
donde $\langle , \rangle$ denotan un producto punto.

Kernels

- Luego, tanto el Lagrangianos como la funció G solo dependen de el producto interno.
- Defina la función de Kernels: $K(x, y) = \langle h(x), h(y) \rangle$.
- Los Kernels más utilizados son:
  1. Polinomio: $K(x, y) = (1 + \langle x, y \rangle)^d$
  2. Radial basis: $K(x, y) = \exp(-\gamma(x - y)^2)$
  3. Red neuronal $K(x, y) = \tanh(-\kappa_1 \langle x, y \rangle) + \kappa_2$

Kernels

SVM - Degree-4 Polynomial in Feature Space



Máquinas de vectores de soporte

LR - Radial Kernel in Feature Space