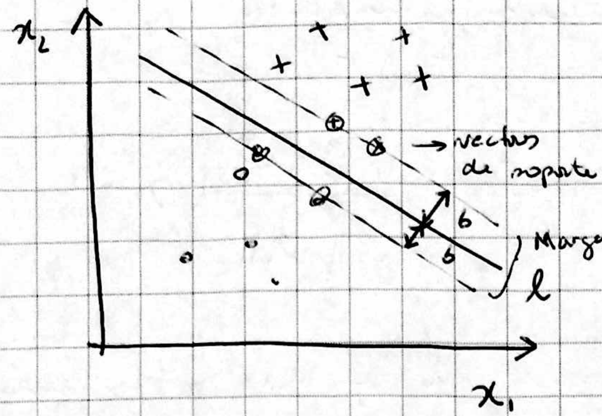
Support Vector Machines (SVM : 1996)



¿Cómo escoger l tal que se maximice b?

l es definido como una recta $g(x) = 0$

donde $g(x) = a_1 x_1 + a_2 x_2 + a_0$

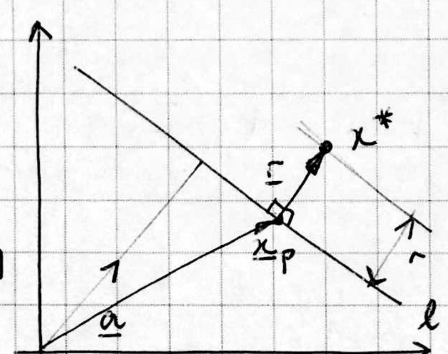
clasificación : $g(\underline{x})$ $> 0 \rightarrow$ clase $+1$
$\leq 0 \rightarrow$ clase -1

¿Cómo se calcula esta línea?

- Distancia de un punto x^* a la recta l

$\underline{r} = [r_1 \; r_2]$

$\underline{x}_p = [x_{p1} \; x_{p2}]$



$\underline{a} = [a_1 \;\; a_2]^T$
vector perpendicular a la recta ⊕

$\underline{x}^* = \underline{x}_p + \underline{r}$

⊕ Sean 2 puntos $\underline{x}_A$ y $\underline{x}_B$ que pertenecen a l

$a_1 x_{A1} + a_2 x_{A2} + a_0 = 0$

$- \; a_1 x_{B1} + a_2 x_{B2} + a_0 = 0$

$[a_1 \;\; a_2](\underline{x}_A - \underline{x}_B) = 0$

$\Rightarrow \underline{a} = [a_1 \;\; a_2]^T \perp l$

(**)

$g(x^*) = a_1(x_{p1} + r_1) + a_2(x_{p2} + r_2) + a_0$

$= g(x_p) + \underline{a}^T \underline{r}$

$= 0 + r \dfrac{\underline{a}^T \underline{a}}{\|a\|}$

$\underline{r} = r \dfrac{\underline{a}}{\|a\|} = r \hat{a}$

(**)

$g(\underline{x}^*) = \cancel{g(\underline{x}_p)}^{\,0} + \underline{a}^T \underline{r}$

$= r \|a\| \rightarrow \boxed{r = \dfrac{|g(x^*)|}{\|a\|}}$

Notas / Notes:

Fecha / Date:

Para encontrar l tal que las distancias de las muestras ~~de~~ de entrenamiento a l sean máximas.

(1): Se puede definir $\frac{|g(x_i^*)|}{\|a\|} > b$ para toda muestra de entrenamiento x_i^*.

Se persigue encontrar $\underline{a}$ y a_0 tal que se cumpla (1) $\forall$ $i = 1 \dots N$.

Se representa (1) como:

(2): $\frac{z_i \, g(x_i)}{\|a\|} > b \qquad i = 1 \dots N$

donde $z_i = \pm 1$ según la clase a que pertenece x_i (± 1)
(se elimina el valor absoluto)

Es necesario imponer una restricción adicional a $\underline{a}$ y a_0 sino escala al infinito.

sujeto a $b\,\|a\| = 1$

(2) entonces es:

$$z_i \, g(x_i) \geq 1$$

(3) $$z_i \, g(x_i) - 1 \geq 0$$

Se persigue entonces minimizar $\|a\|$

maximizar b, entonces equivale a

Notas / Notes:

Fecha / Date:

IDEA:

Hiperplano de separación: $g(\underline{x}) = \underline{a}^T x + a_0 = 0$

Clarificación: $\text{sign}\left(\underline{a}^T \underline{x} + a_0\right)$

Distancia de una muestra $\underline{x}_i$ al hiperplano $r = \dfrac{|g(\underline{x}_i)|}{\|a\|}$ $\quad \|a\| = \sqrt{a_1^2 + \dots a_m^2}$

Se define $z_i \pm 1$ como la clasificación de $\underline{x}_i$

$$\Rightarrow \quad z_i\, g(x_i) > 0$$

La idea es entonces maximizar $\underset{r}{\overset{b}{\updownarrow}}$, i.e minimizar $\|a\|$.

<u>Optimización:</u>

$$\text{Minimizar } \frac{1}{2}\|a\|$$

Sujeto a $\quad z_i\left(\underline{a}^T x_i + a_0\right) - 1 \geq 0 \quad i = 1 \dots N$

Esto se traduce a:

$$(4) \quad J(\underline{a}, \lambda) = \underbrace{\frac{1}{2}\|a\|^2}_{A} \underbrace{-}_{B} \underbrace{\sum_{i=1}^{N} \lambda_i \left[z_i\left(\underline{a}^T x_i + a_0\right) - 1\right]}_{C} \rightarrow \text{Min}$$

A) Se persigue minimizar $\|a\|$

B) Se persigue Maximizar segundo termino

C) Con $\lambda_i \geq 0$. Se fuerza a clarificar correctamente.

Notas / Notes:

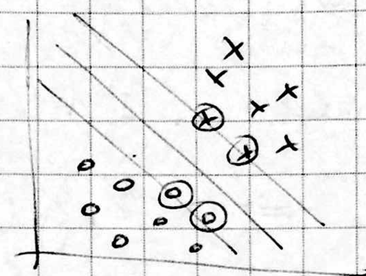
Fecha / Date:

(4) Se puede transformar en (según Kuhn-Tucker)

(5) $$J(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j}^{N} \lambda_i \lambda_j z_i z_j \underbrace{\underline{x}_i^T \underline{x}_j}_{\langle \underline{x}_i, \underline{x}_j \rangle} \to \text{Max}$$

Sujeto a $\lambda_i \geq 0$

Al resolver (5) sólo los vectores de soporte, i.e, los puntos que tienen contacto con los bordes del margen tienen $\lambda_i \neq 0$



El resto son muestras irrelevantes

(5) es un problema de optimización $\Rightarrow \boxed{\text{resultado } \lambda_i}$

<u>Solución:</u>

El hiperplano queda definido como:

$$\hat{g}(\underline{x}) = \underline{\hat{a}}^T \underline{x} + \hat{a}_0$$

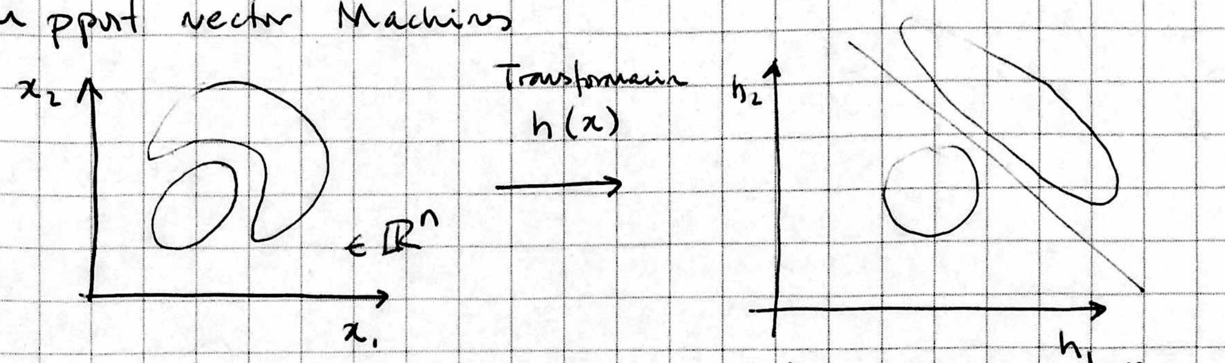
donde $$\underline{\hat{a}} = \sum_{i=1}^{N} \lambda_i z_i \underline{x}_i$$ (derivando (4) e igualando a 0)

$$\hat{g}(\underline{x}) = \sum_{i=1}^{N} \lambda_i z_i \underbrace{\underline{x}_i^T \underline{x}}_{\langle x_i, x \rangle} + \hat{a}_0$$

$$z_i (\underline{\hat{a}}^T x_i + \hat{a}_0) = 1, \to \hat{a}^T x_i + \hat{a}_0 = 1/z_i = z_i$$

Notas / Notes: $$\Rightarrow \hat{a}_0 = z_i - \hat{a}^T x_i \to \frac{1}{N_{SV}} \sum_{i=1}^{N_{SV}} (z_i - \hat{a}^T x_i)$$

Support vector Machines



(puede ser de $\mathbb{R}^m$ con $m>n$)

La idea es hacer una transformación tal que en el espacio transformado se pueda encontrar un hiperplano.

La linea de separación sería

$$\hat{g}(h(\underline{x})) = \hat{a}^T h(x) + \hat{a}_0$$

$$\Rightarrow \quad \hat{g}(x) = \sum_{i=1}^{N} \lambda_i z_i \langle h(x_i), h(x) \rangle + \dots + \hat{a}_0$$

No interesa $h(x)$ sino el producto punto

$$K(\underline{x}_i, \underline{x}) = \langle h(x_i), h(\underline{x}) \rangle$$

↳ Funciones de Kernel.

Muy utilizados son:

Polinomio de grado n: $K(x, x') = (1 + \langle x, x' \rangle)^n$

Radial Basis : $= \exp^{-\|x-x'\|^2/c}$

Neural network : $= \tanh(K_1 \langle x, x' \rangle + K_2)$