Elementos de máquinas de vectores de soporte

Clasificación binaria y funciones *kernel*

Julio Waissman Vilanova

Departamento de Matemáticas
Universidad de Sonora

Seminario de Control y Sistemas Estocásticos 2010

Plan de la presentación

Aprendizaje supervisado

Se tiene un conjunto de aprendizaje $X_t = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ donde

- $x_i \in \mathrm{X} = A_1 \times A_2 \times \cdots \times A_n$ son los datos (objetos, patrones) de entrada
- A_j es el j–ésimo *atributo* de x_i.
- $y_i \in \mathrm{Y} = \{0, 1\}$ es la clase a la que pertenece el dato x_i

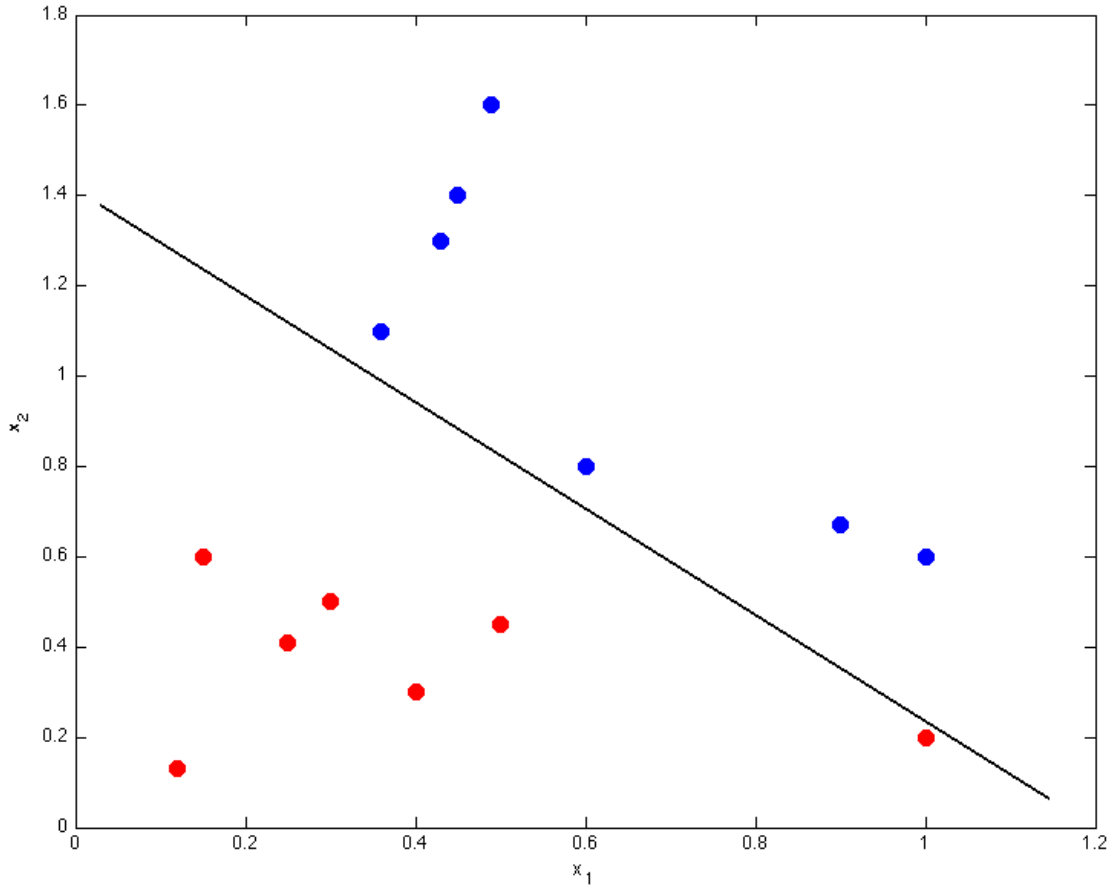
El aprendizaje supervisado consiste en estimar una *regla de decisión* $f : \mathrm{X} \times \Theta \to \mathrm{Y}$, donde Θ es el espacio de parámetros, tal que para un dato desconocido x, $f(x, \theta) = \hat{y}$ sea una estimación *aceptable* de y.

- ¿Que es una estimación aceptable?
- ¿Cual es la estructura de la regla de decisión?

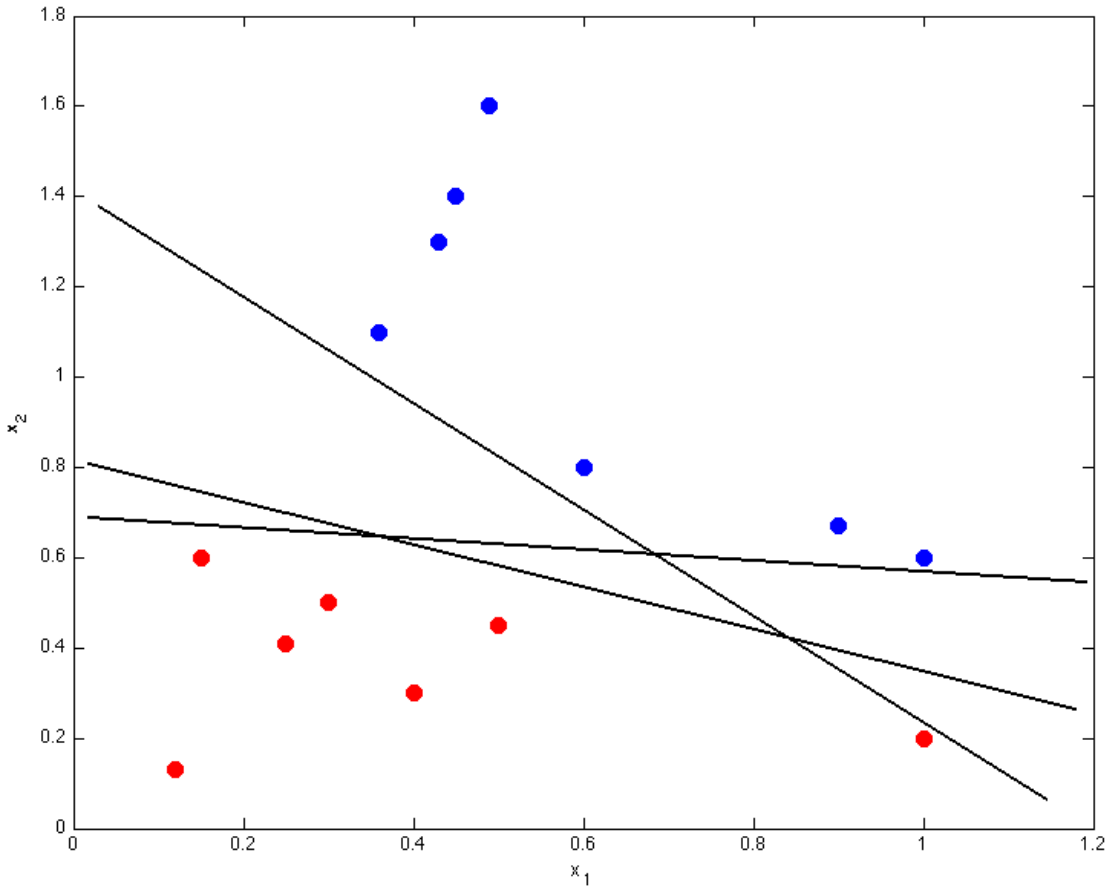
Clasificador lineal binario

Si se asume que $\mathrm{X} \subset \mathbb{R}^n$, un clasificador lineal es

$$\hat{y} = \operatorname{sign}\left(\omega^T x + b\right) = \operatorname{sign}\left(\langle \omega, x \rangle + b\right)$$



¿Cual es la mejor separación?



¿Cual es la mejor separación?

- Perceptron:

$$J(\omega_e) = \min_{\omega_e \in \mathbb{R}^{n+1}} \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

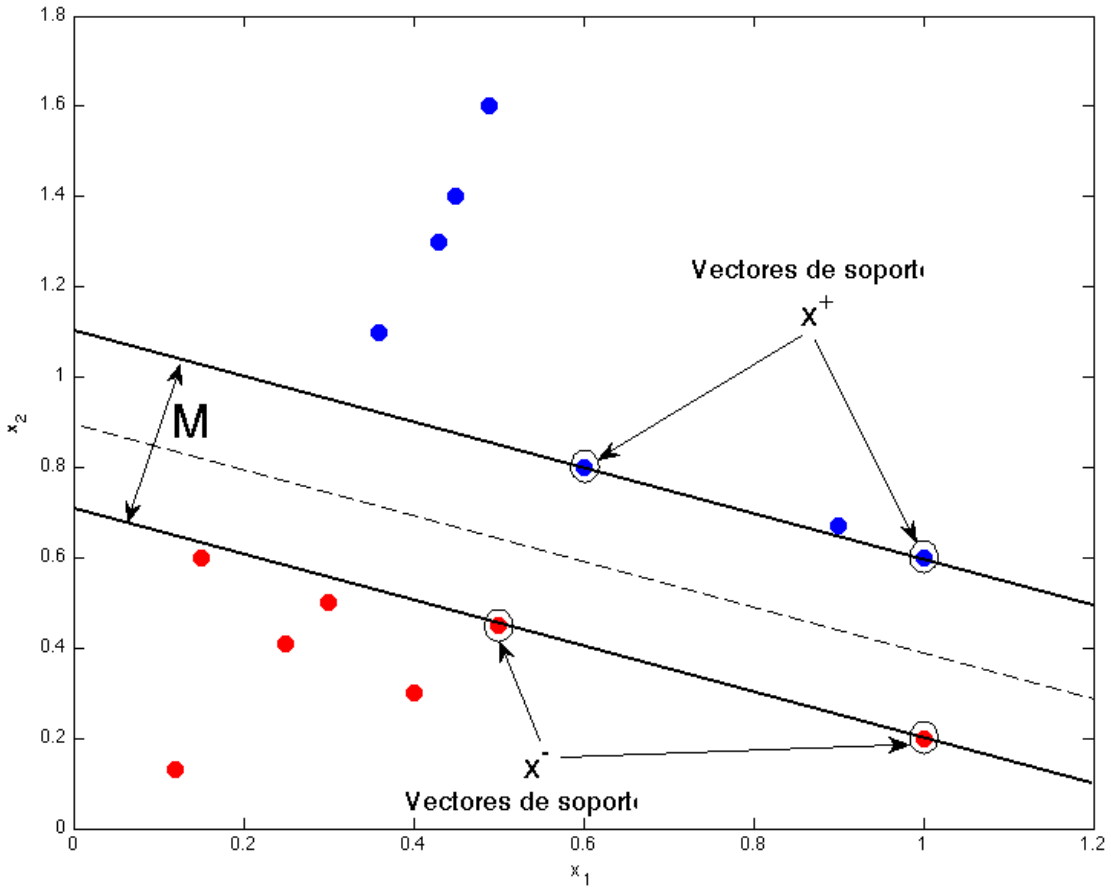
- Discriminante lineal

$$J(\omega_e) = \max_{\omega \in \mathbb{R}^{n+1}} \frac{\omega_e^T S_B \omega_e}{\omega_e^T S_W \omega_e}$$

- Maquinas de Vectores de Soporte (SVM):

Maximizar el margen de separación

Maximización del margen de separación



Maximización del margen de separación

- $\omega^T x^+ + b = 1$ para todo x^+
- $\omega^T x^- + b = -1$ para todo x^-
- Para cada x^- corresponde un x^+ tal que $x^+ = x^- + M\frac{\omega}{\|\omega\|}$

$$\omega^T x^+ + b = 1$$
$$\omega^T (x^- + M\frac{\omega}{\|\omega\|}) + b = 1$$
$$\omega^T x^- + b + \frac{M}{\|\omega\|}\omega^T\omega = 1$$
$$M = \frac{2}{\sqrt{\omega^T\omega}}$$

Criterio de optimización para las SVM

Criterio de optimización

$$J(w, b) = \min_{\omega \in \mathbb{R}^n} \frac{1}{2} \omega^T \omega$$

bajo las restricciones:

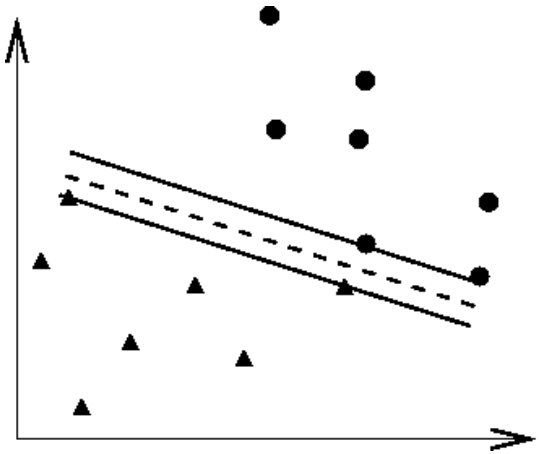
$$(\omega^T x_i + b) y_i \geq 1, \qquad \text{para todo } i = 1, 2, \ldots, m$$

- El error de clasificación se incorpora al criterio como restricciones
- La constante b se calcula después de la optimización

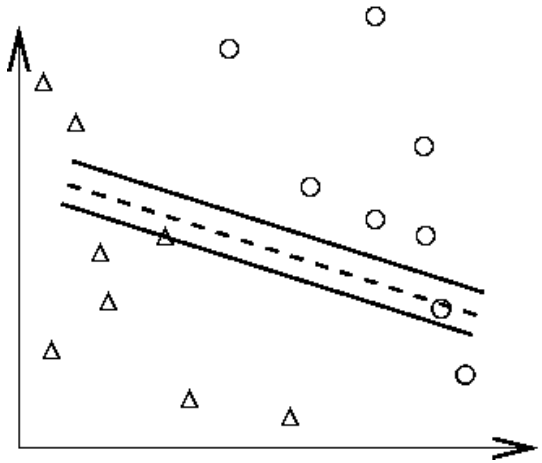
¿Y si X_T no es linealmente separable?

- Permitir errores en clasificación
- Compromiso entre el máximo margen de separación y los posibles errores de predicción
- Mejora la capacidad de generalización de las SVM
- Evita el sobreaprendizaje
- Uso de *variables de holgura*

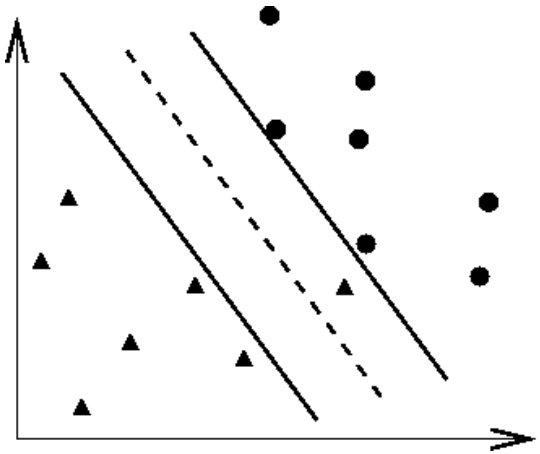
Ejemplo de sobreaprendizaje



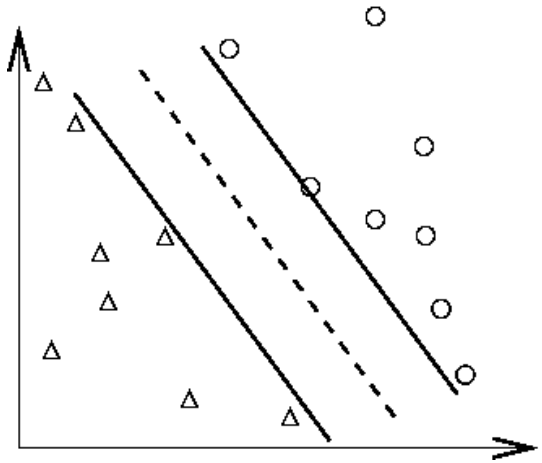
(a) Training data and an overfitting classifier



(b) Applying an overfitting classifier on testing data



(c) Training data and a better classifier



(d) Applying a better classifier on testing data

Planteamiento del problema de SVM

Criterio de optimización

$$J(w, b, \xi) = \min_{\omega \in \mathbb{R}^n} \frac{1}{2}\omega^T\omega + \frac{C}{m}\sum_{i=1}^{m}\xi_i \tag{1}$$

bajo las restricciones:

$$(\omega^T x_i + b)y_i \geq 1 - \xi_i,$$
$$\xi_i \geq 0,$$

para todo $i = 1, 2, \ldots, m$.

- Variables de holgura $\xi = [\xi_1, \ldots, \xi_m]$
- Es necesario establecer C
- Problema de optimización cuadrática con restricciones

Problema primal

Resolver el problema (1) equivale a encontrar para

$$L_P = \frac{1}{2}\omega^T\omega + \frac{C}{m}\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\Big(\alpha_i(\omega^T x_i + b)y_i - 1 + \xi_i) + \beta_i\xi_i\Big)$$

los valores de $\bar{\omega}$, $\bar{b}$, $\bar{\xi}$, $\bar{\alpha}$ y $\bar{\beta}$ tal que para toda ω, b, ξ, α y β

$$L_P(\bar{\omega}, \bar{b}, \bar{\xi}, \alpha, \beta) \leq L_P(\bar{\omega}, \bar{b}, \bar{\xi}, \bar{\alpha}, \bar{\beta}) \leq L_P(\omega, b, \xi, \bar{\alpha}, \bar{\beta})$$

donde $\alpha_i \geq 0$, $\beta_i \geq 0$ son los multiplicadores de Lagrange.

Condiciones necesarias Karush–Kuhn–Tucker

Para que $\bar{\omega}$, $\bar{b}$, $\bar{\xi}$, $\bar{\alpha}$ y $\bar{\beta}$ sea una solución, es necesario que:

$$\begin{aligned}
\bar{\alpha}_i(\bar{\omega}^T x_i + b)y_i - 1 + \xi_i) &= 0, \quad \forall i = 1, \ldots, m, \\
\bar{\beta}_i \bar{\xi}_i &= 0, \quad \forall i = 1, \ldots, m, \\
\partial_\omega L_P(\bar{\omega}, \bar{b}, \bar{\xi}, \bar{\alpha}, \bar{\beta}) &= 0 \\
&= \bar{\omega} - \sum_{i=1}^{m} \bar{\alpha}_i y_i x_i \\
\partial_b L_P(\bar{\omega}, \bar{b}, \bar{\xi}, \bar{\alpha}, \bar{\beta}) &= 0 \\
&= -\sum_{i=1}^{m} \bar{\alpha}_i y_i \\
\partial_\xi L_P(\bar{\omega}, \bar{b}, \bar{\xi}, \bar{\alpha}, \bar{\beta}) &= 0 \\
&= \frac{C}{m} - \sum_{i=1}^{m} \left(\bar{\alpha}_i + \bar{\beta}_i\right)
\end{aligned}$$

Problema dual

De las condiciones KKT

$$\omega = \sum_{i=1}^{m} \alpha_i y_i x_i,$$

$$\frac{C}{m} = \alpha_i + \beta_i, \quad \text{de donde } \alpha_i \leq \frac{C}{m},$$

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

se sustituyen en L_P para encontar el problema dual

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

el cual es únicamente un problema de maximización en α.

Reconocimiento

La regla de decisión del clasificador lineal es

$$\begin{aligned}\hat{y} &= \text{sign}\left(\omega^T x + b\right) \\ &= \text{sign}\left(\sum_{i \in SV} \alpha_i y_i x_i^T x + b\right),\end{aligned}$$

donde SV es el conjunto de índices tales que $\alpha_i > 0$.

Para calcular b *numéricamente estable*, se considera un promedio del valor de b los *vectores soporte* en los cuales $\xi_i = 0$, tal que,

$$\left(\omega^T x_i + b\right) y_i = 1.$$

Si $\xi_i > 0$, entonces $\beta_i = 0$ y por lo tanto $\alpha_i = \frac{C}{m}$. Por lo tanto:

$$b = \frac{1}{|SV^*|}\left(y_i - \sum_{i \in SV^*} y_i - \omega^T x_i\right),$$

donde $SV*$ es el conjunto de índices tales que $0 < \alpha_i \leq C/m$.

Máquina de vectores de soporte

Entrenamiento

$$\max_{\alpha \in \mathbb{R}^m} L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle,$$

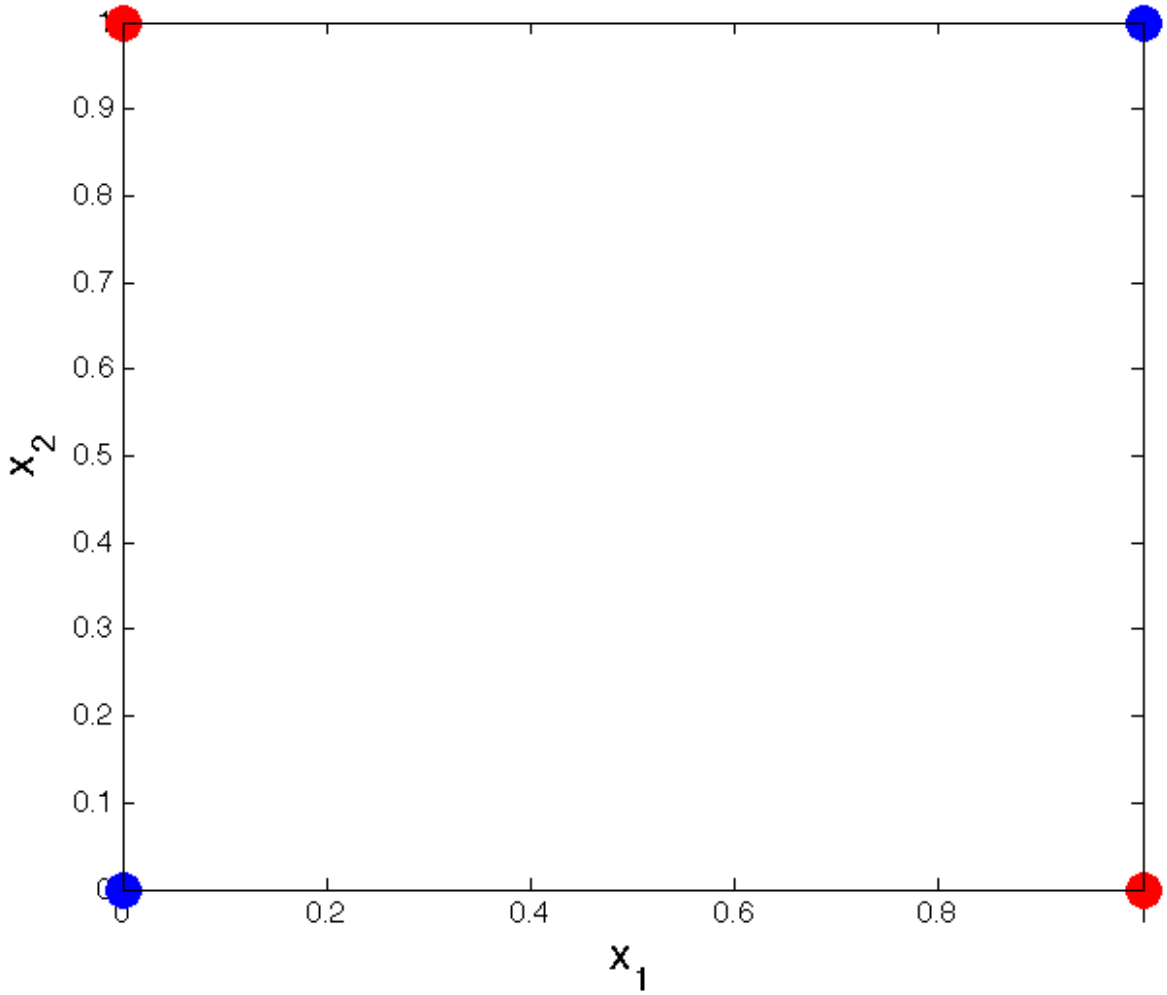
bajo $0 \leq \alpha_i \leq \frac{C}{m}$, para $i = 1, \ldots, m$.

Reconocimiento

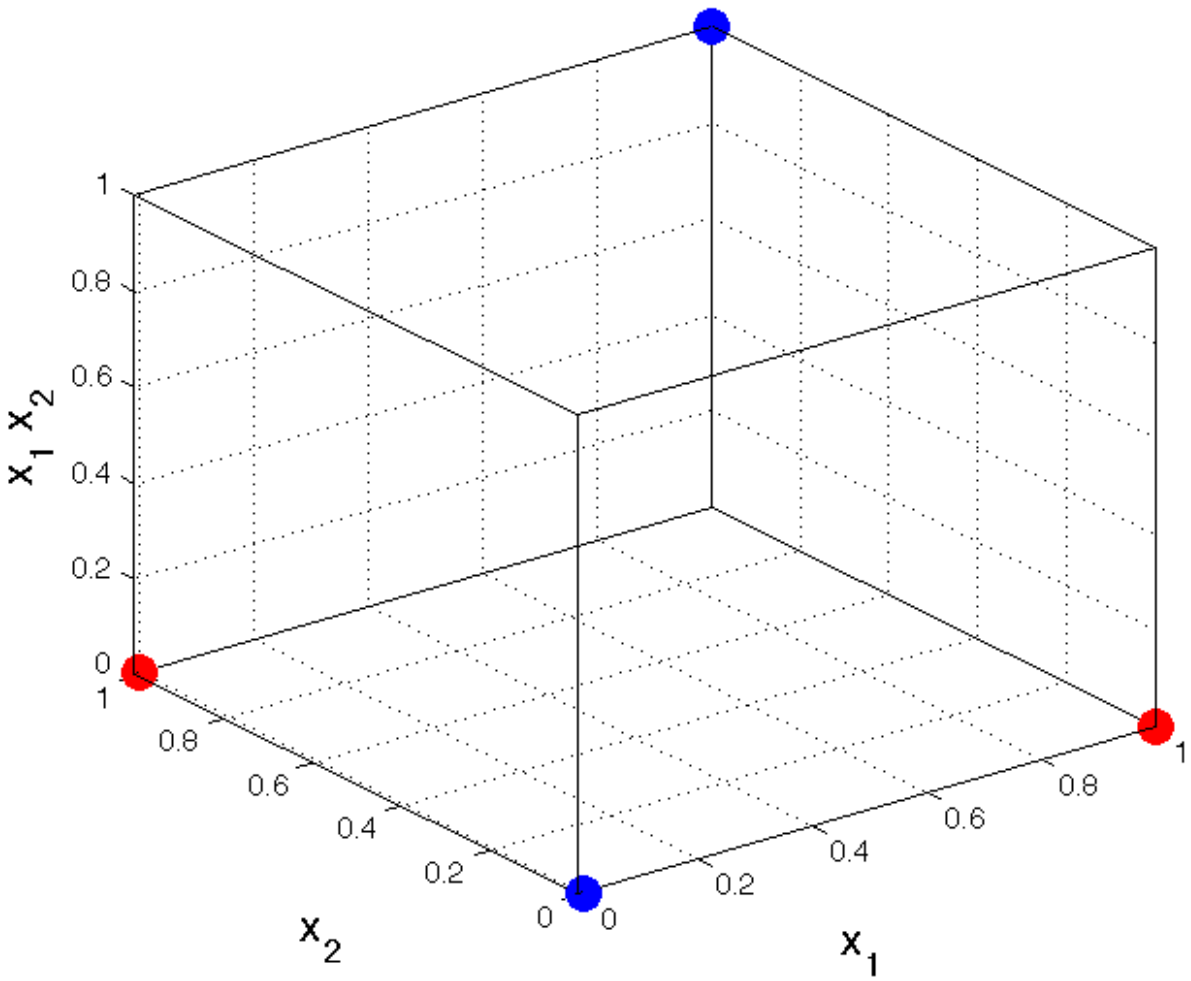
$$\hat{y} = \text{sign}\Big(\sum_{i \in SV} \alpha_i y_i \langle x_i, x \rangle + b \Big)$$

¡Tanto el aprendizaje como el reconocimiento dependen solamente en el producto interno entre vectores!

El truco del kernel: el problema de la *Xor*



El truco del kernel: el problema de la *Xor*



Clasificadores polinomiales

- Sea $C_d : \mathbb{R}^n \to \mathbb{R}^{N_C}$ donde las entrada de $C_d(x)$ son todos los posibles productos ordenados de grado d de x.
- Sea $\theta : \mathbb{R}^n \to \mathbb{R}^{N_\theta}$ los productos *no ordenados* de grado d, tal que $\langle C_d(x), C_d(x') \rangle = \langle \theta(x), \theta(x') \rangle$. Por ejemplo

$$C_2([x_1, x_2]^T) = [x_1^2, x_2^2, x_1 x_2, x_2 x_1]^T,$$
$$\theta([x_1, x_2]^T) = [x_1^2, x_2^2, \sqrt{2} x_1 x_2]^T.$$

- Un clasificador polinomial realiza una transformación del espacio de entrada a un *espacio de características* de mayor dimensión, esperando encontrar una separación lineal en el nuevo espacio.
- Se pueden aplicar otro tipo de transformaciones y utilizar un clasificador lineal en el espacio de características.

La maldición de la dimensionalidad

¿Cual es la dimensión del nuevo espacio de características?

$$N_\theta = \left(\begin{array}{c} d+n-1 \\ d \end{array} \right) = \frac{d+n-1}{d!(n-1)!)}$$

Ejemplo:

- Reconocimiento de caracteres escritos a mano, en forma de mapa de bits de 16×16 (256 atributos).
- Si se considera una transformación θ_5, $N_\theta \approx 10^{10}$.
- ¡El problema no es tratable!

El truco de las funciones *kernel*

Para entrenar y utilizar un clasificado con SVM, solamente se requiere poder calcular el producto interno $\langle \theta_d(x), \theta_d x' \rangle$. Si se define $k : \mathbb{R}^n \to \mathbb{R}$ tal que

$$
\begin{aligned}
k(x, x') &= \langle \theta_d(x), \theta_d x' \rangle = \langle C_d(x'), C_d x' \rangle \\
&= \sum_{j_1=1}^{n} \cdots \sum_{j_d=1}^{n} x_{j_1} \ldots x_{j_d} x'_{j_1} \ldots x'_{j_d} \\
&= \sum_{j_1=1}^{n} x_{j_1} x'_{j_1} \cdots \sum_{j_1=1}^{n} x_{j_d} x'_{j_d} \\
&= \left(\sum_{j=1}^{n} x_j x'_j \right)^d = \langle x, x' \rangle^d
\end{aligned}
$$

Funciones *kernel*

- Dada una función $k : \mathrm{X}^2 \to \mathbb{R}$, y un conjunto $X_T = \{x_1, \ldots, x_m\}$, $x_i \in \mathrm{X}$, la matriz K de dimensión $m \times m$ con elementos:

$$K_{ij} = k(x_i, x_j)$$

  se conoce como Matriz de Gram (o matriz de kernel) generada por k respecto a X_T.
- Sea X un conjunto no vacío, la función $k : \mathrm{X}^2 \to \mathbb{R}$ la cual, para todo $m \in \mathbb{N}$ y para todo conjunto $X_T = \{x_1, \ldots, x_m\}$ genera una matriz de Gram K simétrica definida positiva es llamado un *kernel real definido positivo*, o simplemente *kernel*.

Propiedades de las funciones *kernel*

- $k(x, x) \geq 0$ para todo $x \in \mathrm{X}$
- $k(x, x') = k(x', x)$
- $|k(x, x')| \leq k(x, x')k(x', x)$

Resultado interesante

Una función kernel $k : \mathrm{X}^2 \to \mathbb{R}$ es un producto interno en al menos un espacio de características.

- Sea $\theta : \mathrm{X} \to \{f : \mathrm{X} \to \mathbb{R}\}$ tal que $\theta(x)(\cdot) = k(x, \cdot)$.
- Se genera un espacio vectorial con la imagen de θ:

$$f(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i), \quad m \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathrm{X},$$

$$g(\cdot) = \sum_{j=1}^{m'} \beta_j k(\cdot, x'_j), \quad m' \in \mathbb{N}, \beta_j \in \mathbb{R}, x'_j \in \mathrm{X},$$

Resultado interesante

- Se define un producto interno:

$$\langle f, g \rangle := \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x_j)$$

- Para demostrar que $\langle f, g \rangle$ es un producto interno:

$$\langle f, g \rangle = \sum_{j=1}^{m'} \beta_j f(x'_j) = \sum_{i=1}^{m} \alpha_i g(x_i),$$

$$\langle f, g \rangle = \sum_{i,j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \geq 0,$$

$$f(x) = \langle f, k(\cdot, x) \rangle,$$

$$|f(x)|^2 = |\langle f, k(\cdot, x) \rangle|^2 \leq \langle f, f \rangle k(x, x)$$

por lo que el producto interno es bilineal, simétrico, definido positivo y $\langle f, f \rangle$ implica $f = 0$.

Función kernel como producto interno

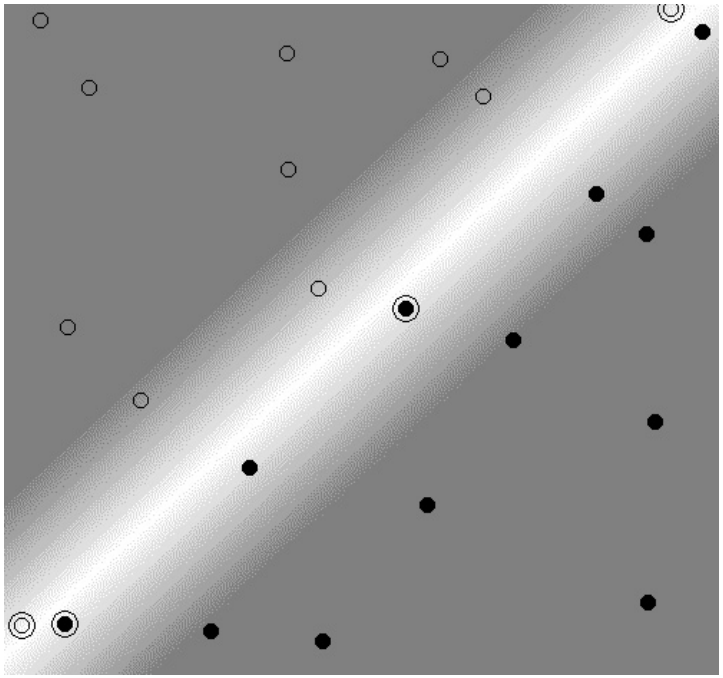
Por lo tanto:

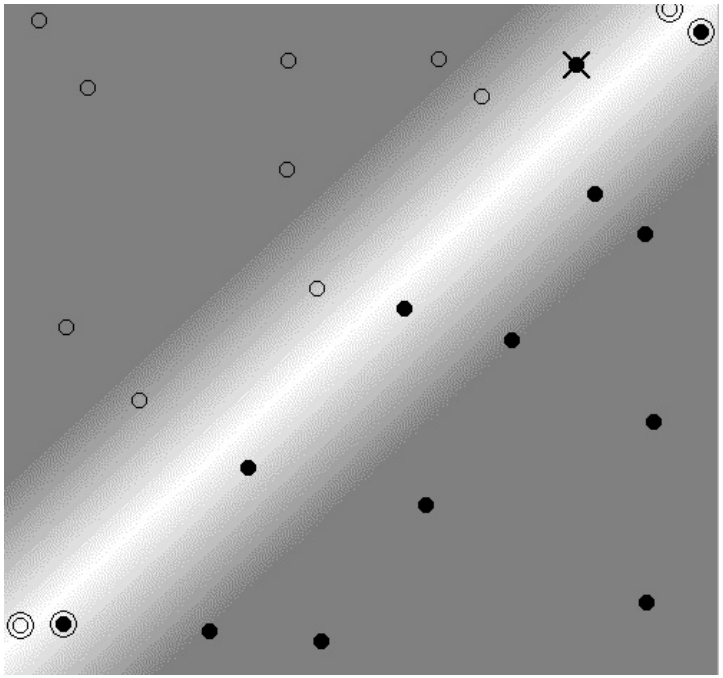
$$k(x, x') = \langle k(\cdot, x), k(\cdot, x') \rangle = \langle \theta(x), \theta(x') \rangle$$

Principales funciones kernel utilizadas:

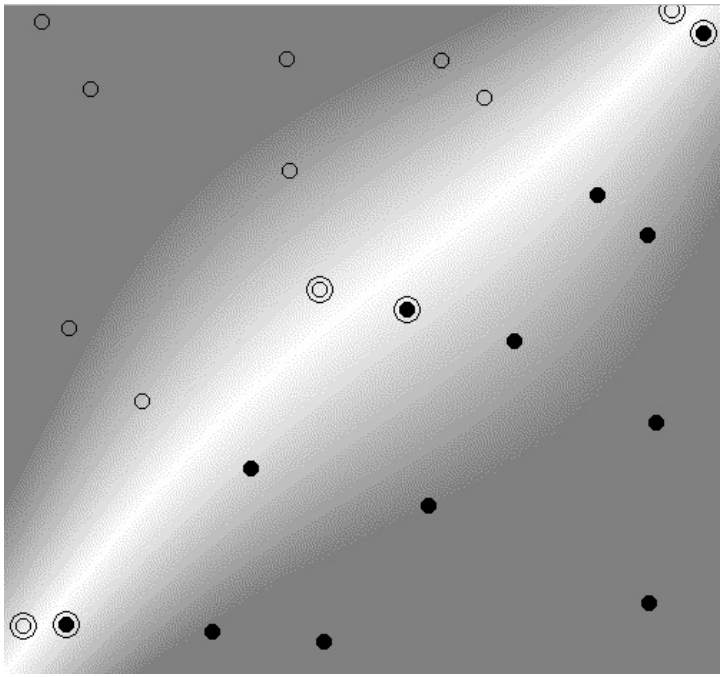
- Polinomial: $k(x, x') = \langle x, x' \rangle^d$,
- Polinomial no homogeneo: $k(x, x') = (\langle x, x' \rangle + c)^d$,
- Gaussiano: $k(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$,
- Sigmoide: $k(x, x') = \tanh(\kappa \langle x, x' \rangle + \nu)$.

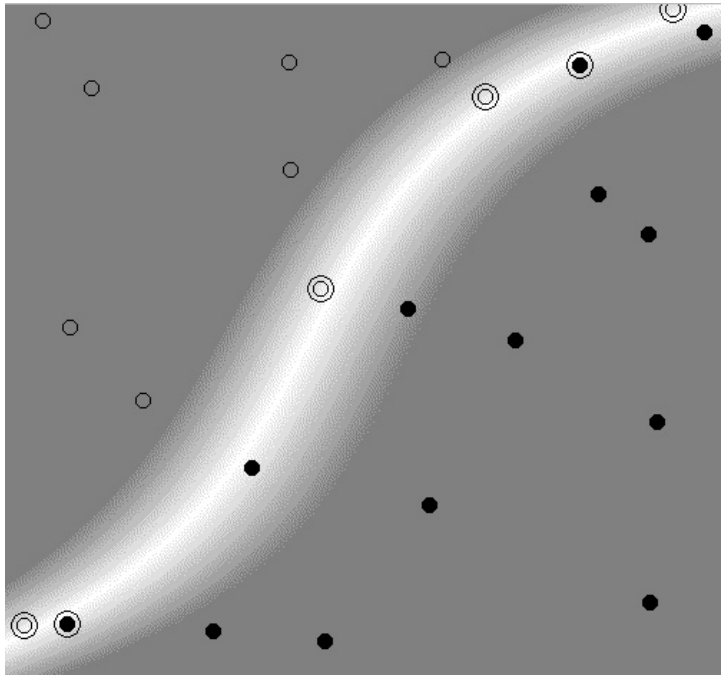
Ejemplo





Ejemplo





... y esto fue todo por hoy!

Muchas gracias por su atención