Support Vector Machine

Juan Carlos Caicedo
Juan Carlos Mendivelso

Maestria en Ingenieria de Sistemas y Computacion
Universidad Nacional de Colombia

20 de marzo de 2007

Agenda

Outline

Introduccion

Clasificador lineal que utiliza la siguiente metodología

- Mapear puntos de entrenamiento a un espacio dimensional mayor
- Construir un hiperplano que separa los puntos en las clases respectivas
- Clasificar un punto nuevo de acuerdo a su ubicación con respecto al hiperplano de separación

Outline

Preliminares

Para el caso de 2 clases:

- Los x_k son los patrones de entrenamiento en $\Re^j$, $k = 1, ..., n$
- Los patrones x_k tienen un atributo z_k que determina la clase
- $z_k \in \{-1, 1\}$
- Los patrones x_k son transformados a $y_k = \varphi(x_k)$
- Los y_k están en $\Re^d$, con $d > j$

Discriminador lineal

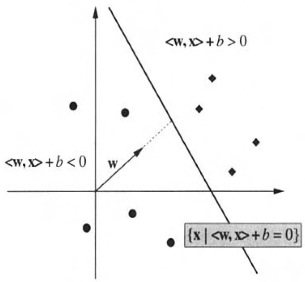
Se construye un discriminador lineal en el espacio aumentado de la forma

$$g(y) = \langle w, y \rangle = w^t y$$

Hiperplano de Separación

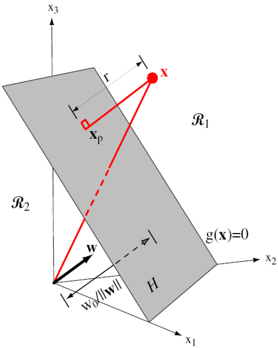
Este discriminador es una familia de hiperplanos y el hiperplano de separación es:

$$w^t y = 0$$



Escala del vector w

El vector w puede tener cualquier escala y sigue generando el mismo hiperplano. Aunque el plano y los patrones permanezcan estáticos la distancia entre ellos depende de la norma de w:



Escala del vector w

Un patrón y puede expresarse como:

$$y = y_p + r \frac{w}{\| w \|}$$

Teniendo en cuenta que $g(y_p) = 0$:

$$g(y) = w^t y = w^t \left(y_p + r \frac{w}{\| w \|} \right)$$

$$g(y) = r \frac{\| w \|^2}{\| w \|} = r \| w \|$$

Entonces, la distancia del patrón y_k al plano es:

$$r = \frac{g(y_k)}{\| w \|}$$

Definición

Como hay varios vectores w que generan el mismo plano, seleccionaremos uno con el siguiente criterio:
El vector de pesos w es llamado **forma canonica** del hiperplano $w^t y = 0$ con respecto a los patrones $y_1, y_2, ..., y_n$, si se escala de manera que:

$$min_{i=1,...,n} |< w, y_i >| = 1$$

Este plano asegura que:

$$z_k g(y_k) \geq 1$$

Hiperplano

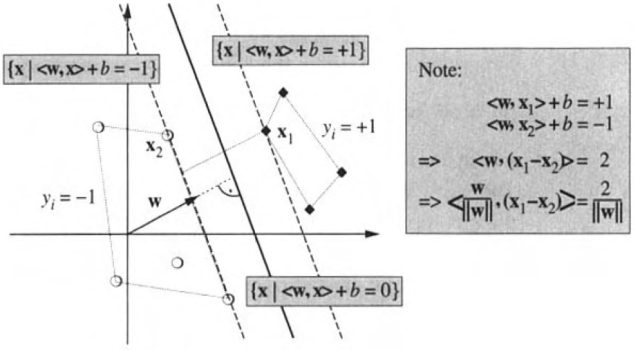


Figura: Forma canónica del hiperplano. La margen medida perpendicularmente al hiperplano es $\frac{1}{\| w \|}$

Outline

Funcion φ

- El conocimiento del diseñador en dominio de aplicación
- Funciones polinomiales o gausianas
- Otras funciones base (*kernel trick*)

Propósito

El propósito del entrenamiento de un SVM es maximizar la distancia entre los patrones y_k y el hiperplano de separación.

$$max : r = \frac{z_k g(y_k)}{\| w \|}$$

Optimización

Esto se logra minimzando $\parallel w \parallel$ dado que es inversamente proporcional a r:

$$minimizar : \tau(w) = \frac{1}{2} \parallel w \parallel^2$$

sujeto a la restricción:

$$\mid g(y_k) \mid = z_k w^t y_k \geq 1$$

Optimización

Se utilizan los multiplicadores de Lagrange para minimizar w:

$$L(w, \alpha) = \frac{1}{2} \parallel w \parallel^2 - \sum_{k=1}^{n} \alpha_k (z_k w^t y_k - 1)$$

Encontrando el mínimo

Si existe un mínimo local, entonces:

$$\frac{\delta}{\delta w} L(w, \alpha) = 0$$

$$\frac{\delta}{\delta w}\left(\frac{1}{2} \parallel w \parallel^2 - \sum_{k=1}^{n} \alpha_k (z_k w^t y_k - 1)\right) = 0$$

$$\frac{\delta}{\delta w}(\frac{1}{2} \parallel w \parallel^2) - \sum_{k=1}^{n} \alpha_k \frac{\delta}{\delta w}(z_k w^t y_k - 1) = 0$$

$$w - \sum_{k=1}^{n} \alpha_k z_k y_k = 0$$

Condiciones KKT

Según las condiciones de KKT, $\alpha_k \neq 0$:

$$\alpha_k[z_k w^t y_k - 1] = 0, \forall k = 0, ..., n$$

Esto significa que los vectores de soporte están en la margen. Los demás vectores de entrenamiento son irrelevantes, porque $z_k g(t_k) \geq 1$ y no satisfacen la condición

Problema Dual

Al reemplazar la solución de w en la fórmula de Lagrange se obtiene la forma dual del problema de optimización, que es el que se resuelve en la práctica:

$$max : W(\alpha) = \sum_{k=1}^{m} \alpha_k - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j z_i z_j < y_i, y_j >$$

Sujeto a:

$$\alpha_k \geq 0, \forall k = 1, ..., m$$

$$\sum_{i=0}^{m} \alpha_i z_i = 0$$

Funcion de Decisión

Utilizando la solución para w tenemos que:

$$f(x) = sng\left(\sum_{k=1}^{m} z_k \alpha_k < x, x_k >\right)$$

Outline

Problema XOR

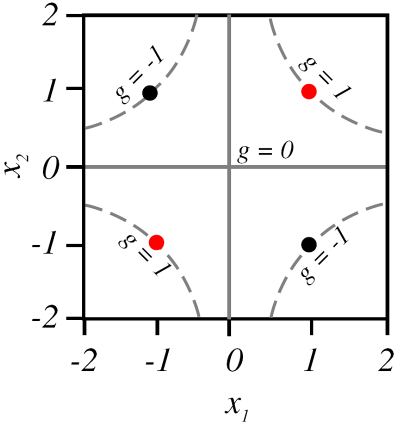


Figura: Este problema no puede resolverse con un clasificador lineal

Formulacion

Los vectores de entrenamiento son los siguientes:

k	x_1	x_2	z_k
1	1	1	ω_1
2	1	-1	ω_2
3	-1	-1	ω_1
4	-1	1	ω_2

En primer lugar se mapean a otro espacio.

Funcion φ

Existen varias funciones que pueden aplicarse, se eligió la siguiente expansión de segundo orden:

$$\varphi : \Re^2 \longrightarrow \Re^6$$

$$(x_1, x_2) \longrightarrow (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, x_1^2, x_2^2)$$

Optimizacion

Se requiere maximizar:

$$\sum_{k=1}^{4} \alpha_k - \frac{1}{2} \sum_{i}^{n} \sum_{j}^{n} \alpha_i \alpha_j z_i z_j y_i^t y_j$$

Sujeto a:

$$\alpha_1 - \alpha_2 + \alpha_3 - \alpha_4 = 0$$

$$\alpha_k \geq 0, k = 1, 2, 3, 4$$

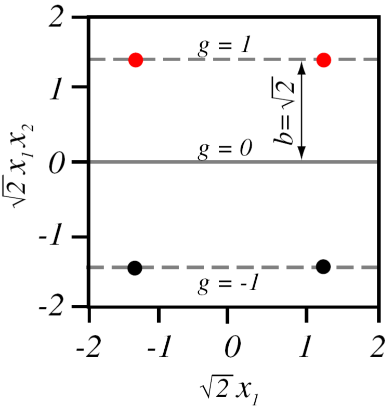
Optimizacion

- La solución puede encontrarse mediante algún procedimiento de optimización como el descenso del gradiente
- En este problema pequeño se puede encontrar analíticamente
- La solución óptima es $w^* = (1/8, 1/8, 1/8, 1/8)$
- Todos los patrones se utilizan como vectores de soporte, debido a la simetría del problema, algo que es inusual

Discriminante Lineal

- La función lineal es $g(x_1, x_2) = x_1 x_2$ y el hiperplano de separación es $g = 0$
- La longitud de la margen es $r = \dfrac{1}{\| w \|} = \sqrt{2}$
- La solución se puede visualizar en un sub-espacio 2d proyectado

Solucion



Software

El método más utilizado para entrenar SVM solucionando el problema dual es SMO (Sequential Minimal Optimization)
Está implementado en:

- SVM-JAVA: http://idis.hwanjoyu.org/svm-java/
- Weka: http://www.cs.waikato.ac.nz/ml/weka/

Fin

Gracias.