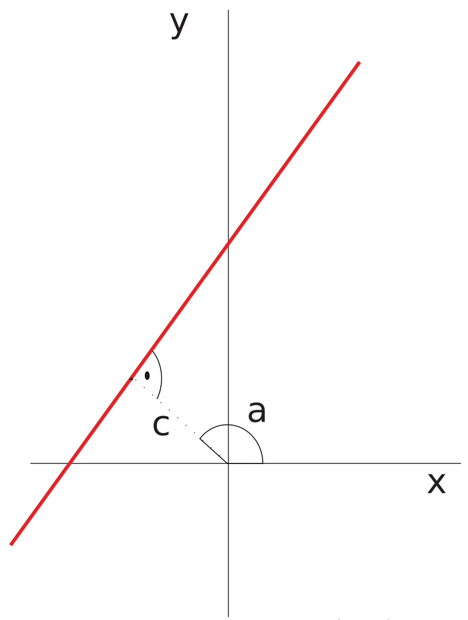
Máquinas de Vectores Soporte

Ref.: A Tutorial on Support Vector Machines for Pattern Recognition

CHRISTOPHER J.C. BURGES

Saturnino Maldonado 2013

Ecuación Normal de la recta

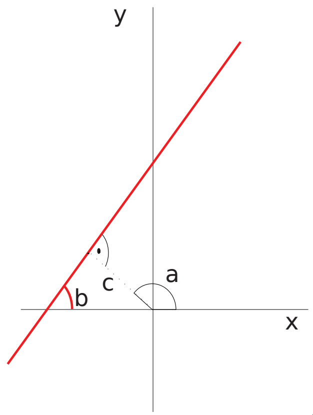


$$\cos\ (a)\ x + \operatorname{sen}\ (a)\ y + c = 0$$

a: ángulo del eje x a la normal a la recta

c: distancia del origen a la recta

Ecuación Explícita de la recta



$$a = b + \pi/2$$
$$cos(a)\ x + sen(a)\ y + c = 0$$
$$cos(b + \pi/2)\ x + sen(b + \pi/2)\ y + c = 0$$
$$-sen(b)\ x + cos(b)\ y + c = 0$$
$$y = tag(b)\ x - \frac{c}{cos(b)}$$

Representación vectorial de la recta

$$w = (cos(a), sen(a))$$

$$||w|| = \sqrt{cos^2(a) + sen^2(a)} = 1 \text{ Vector normal a la recta}$$

$$\mathbf{wx^T} + c = 0$$

Representación alternativa de norma k

$$w = (k\ cos(a), k\ sen(a))$$

$$||w|| = \sqrt{k^2\ cos^2(a) + k^2\ sen^2(a)} = k$$

$$\mathbf{wx^T} + k\ c = 0$$

Información de la recta

$$A\ x + B\ y + C = 0$$

$$a = arctan(\frac{B}{A})$$

a: ángulo del eje x a la normal de la recta

$$c = \frac{C}{\sqrt{A^2+B^2}}$$ c: distancia de la recta al origen

$$\mathbf{w} = (A, B)$$ Vector normal a la recta

$$c = \frac{C}{||\mathbf{w}||}$$ distancia del origen a la recta

$$(\mathbf{w}, C)$$ Parámetros que definen la recta

Información de la recta

El plano se divide con la aplicación de:

$$\mathbf{wx^T} + k\ c = 0$$

$\mathbf{wx^T} + k\ c > 0$ A un lado de la recta frontera

$\mathbf{wx^T} + k\ c < 0$ A otro lado de la recta frontera

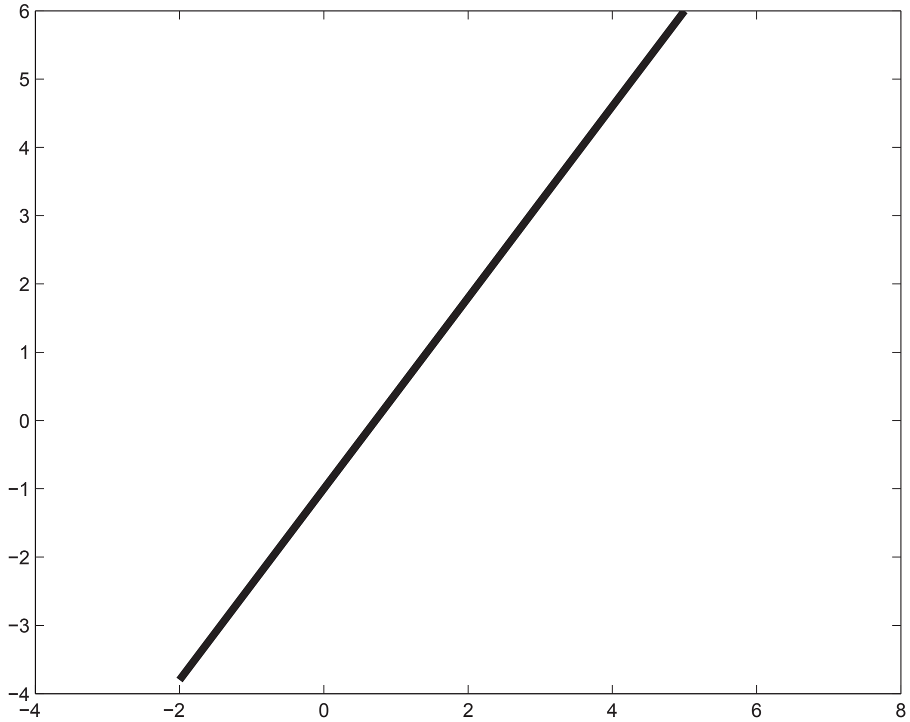
Por tanto para cualquier valor fuera de la frontera podemos optar por un valor cualquiera de la recta

aplicada a ese punto pero del mismo signo.

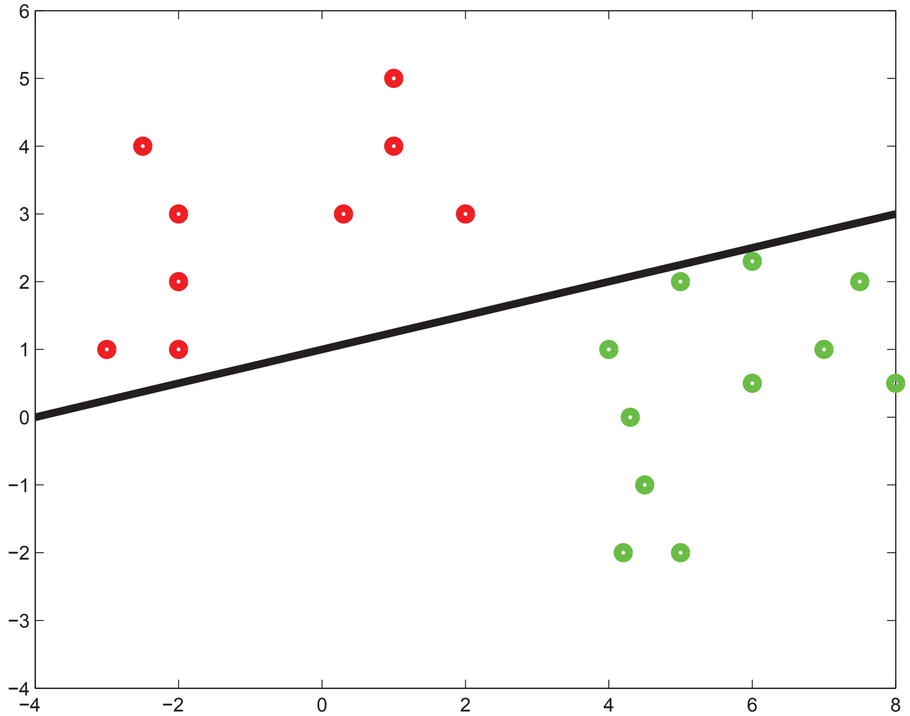
$k_1\mathbf{wx^T} + k_1 k\ c > 0$ A un lado de la recta frontera

$k_1\mathbf{wx^T} + k_1 k\ c < 0$ A otro lado de la recta frontera

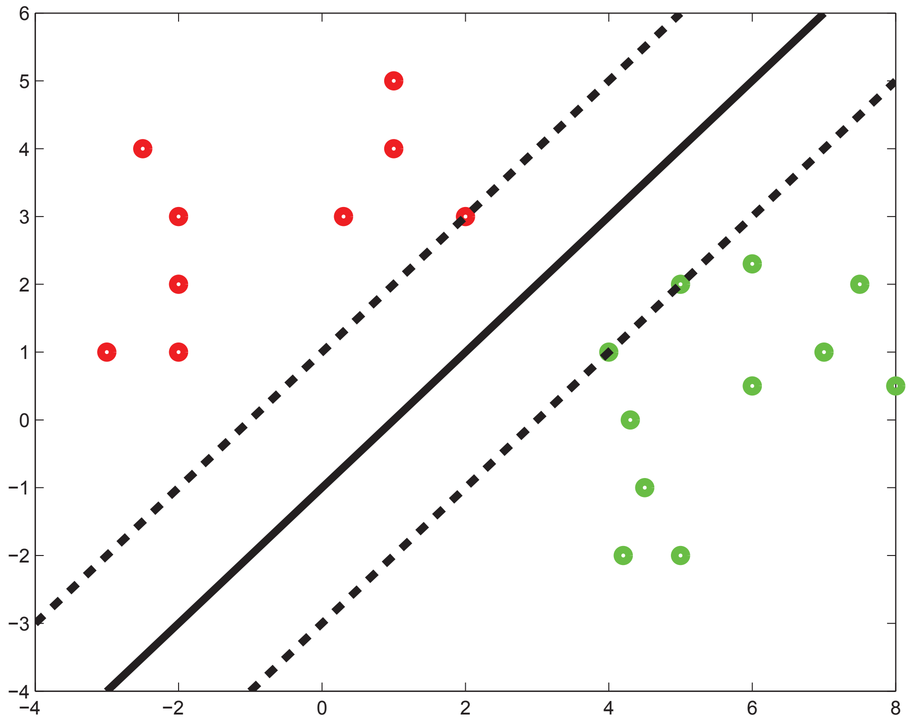
Ejemplo



Ejemplo

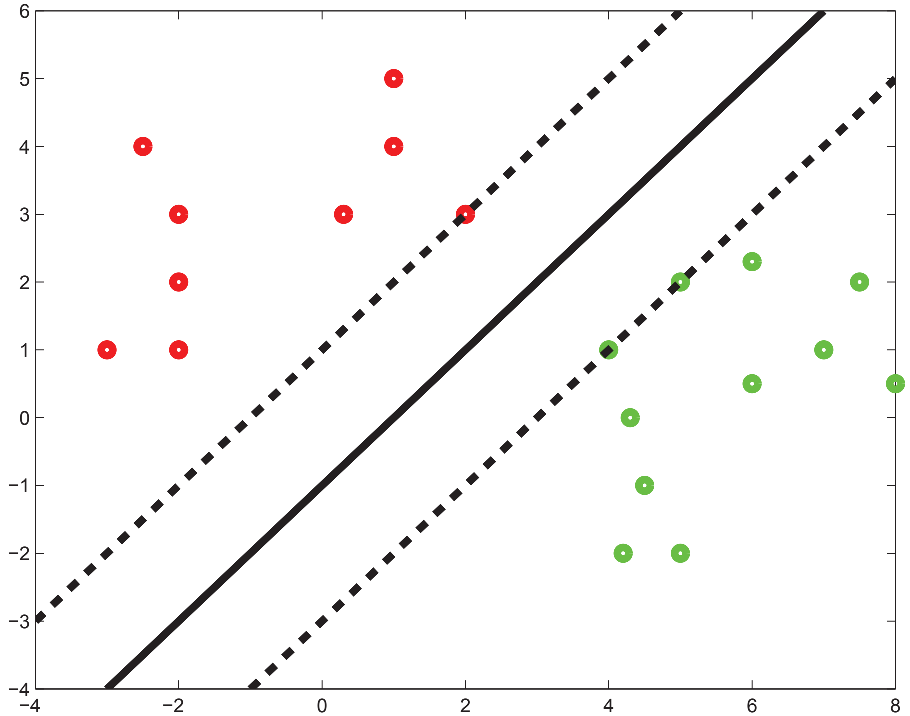


Ejemplo



Por convenio,
la ec. de la recta vale ± 1 en los vectores más cercanos

Conjuntos Linealmente separables



$$\mathbf{w}\mathbf{x}^{\mathbf{T}} + b = 0$$

$\mathbf{x_i}$: Vector de entrada de entrenamiento

$y_i \in \{-1, 1\}$: Etiqueta del vector $\mathbf{x_i}$

Support Vector Machines

$d'_+ = \frac{|1-b|}{||\mathbf{w}||}$ distancia de la recta - - - (+) hasta el origen

$d'_- = \frac{|-1-b|}{||\mathbf{w}||}$ distancia de la recta - - - (-) hasta el origen

$d_+ = \frac{1}{||\mathbf{w}||}$ distancia de la recta al vector más cercano

$d_- = \frac{1}{||\mathbf{w}||}$ distancia de la recta al vector más cercano

$$Margen = d_+ + d_- = \frac{2}{||\mathbf{w}||}$$

Objetivo: $\mathbf{w} = arg_{\mathbf{w}} \quad max(\frac{2}{||\mathbf{w}||})$

Optimización

$$\mathbf{w} = arg_{\mathbf{w}} \quad max(\frac{2}{||\mathbf{w}||}) = arg_{\mathbf{w}} min(\frac{1}{2}||\mathbf{w}||^2)$$

Sujeto a las siguientes restricciones (que el conjunto está bien separado):

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \geq 0 \qquad \forall i = 1, ..., l$$

Multiplicadores de Lagrange

Se introduce un multiplicador de Lagrange positivo

$$\alpha_i \quad \forall i = 1, ..., l$$

por cada condición, se resta de la función objetivo,
se suman los multiplicadores y se minimiza la función restante

$$L_p \equiv \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{l} \alpha_i y_i(\mathbf{w}^T \mathbf{x}_i + b) + \sum_{i=1}^{l} \alpha_i$$

$$\alpha_i \geq 0$$

Problema de programación cuadrática convexo

Derivando

$$\frac{\partial L_p}{\partial w_h} = w_h - \textstyle\sum_{i=1}^{l} \alpha_i y_i x_h = 0 \quad \forall h = 1, ..., N$$

De donde se obtiene:

$$\mathbf{w} = \textstyle\sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L_p}{\partial b} = \textstyle\sum_{i=1}^{l} \alpha_i y_i = 0$$

De donde se obtiene:

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

Simplificando

$$L_p \equiv \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{l} \alpha_i y_i (\mathbf{w}^T \mathbf{x}_i + b) + \sum_{i=1}^{l} \alpha_i$$

$$L_p \equiv \frac{1}{2}\mathbf{w}^T \mathbf{w} - \sum_{i=1}^{l} (\mathbf{w}^T \alpha_i y_i \mathbf{x}_i + \alpha_i y_i b) + \sum_{i=1}^{l} \alpha_i$$

$$L_p \equiv \frac{1}{2}\sum_{i=1}^{l} \alpha_i y_i \mathbf{w}^T \mathbf{x}_i - \sum_{i=1}^{l} \alpha_i y_i \mathbf{w}^T \mathbf{x}_i - \sum_{i=1}^{l} \alpha_i y_i b + \sum_{i=1}^{l} \alpha_i$$

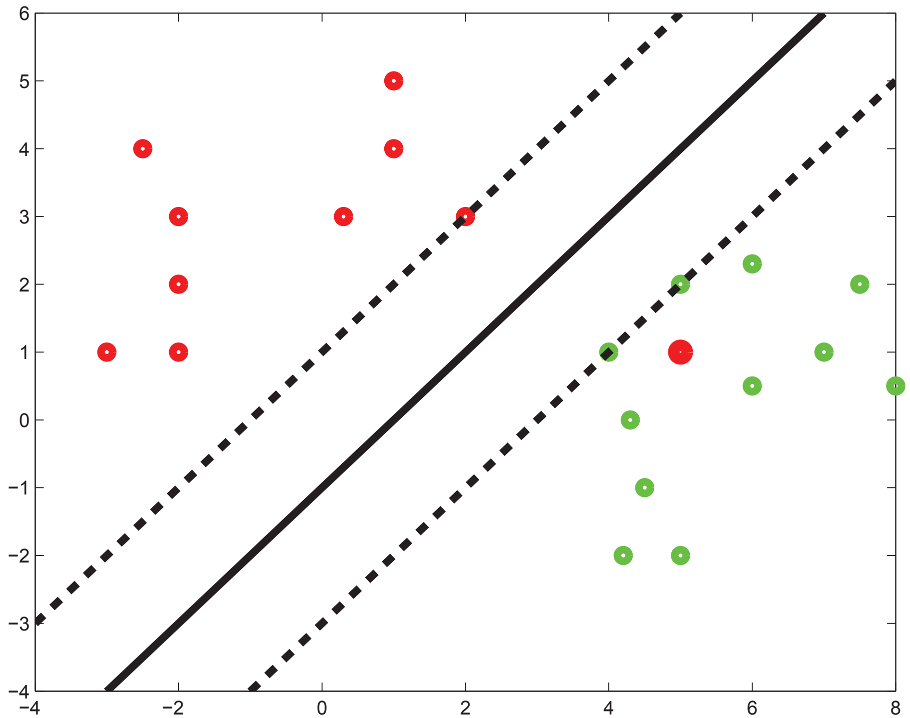
$$L_p \equiv \frac{1}{2}\sum_{i=1}^{l} \alpha_i y_i \mathbf{w}^T \mathbf{x}_i - \sum_{i=1}^{l} \alpha_i y_i \mathbf{w}^T \mathbf{x}_i + \sum_{i=1}^{l} \alpha_i$$

$$L_p \equiv -\frac{1}{2}\sum_{i=1}^{l} \alpha_i y_i \mathbf{w}^T \mathbf{x}_i + \sum_{i=1}^{l} \alpha_i$$

$$L_p \equiv \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l} \alpha_i y_i \mathbf{w}^T \mathbf{x}_i$$

$$L_p \equiv \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j$$

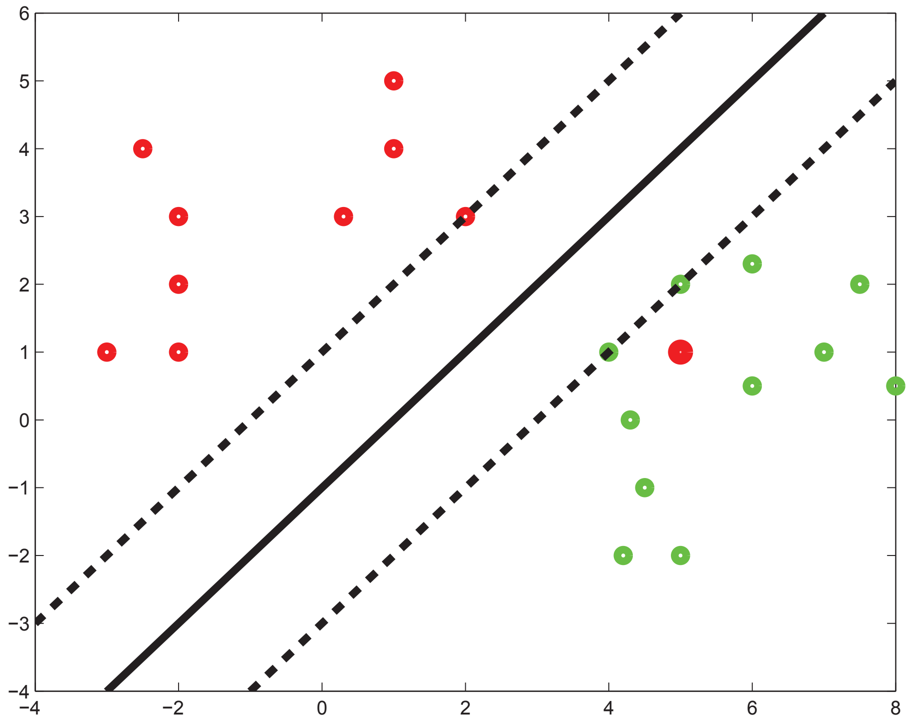
No Separable



$$\mathbf{w} = arg_{\mathbf{w}} \quad max\left(\frac{2}{||\mathbf{w}||}\right) = arg_{\mathbf{w}} min\left(\frac{1}{2}||\mathbf{w}||^2\right)$$

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 \geq 0 \qquad \forall i = 1, ..., l$$

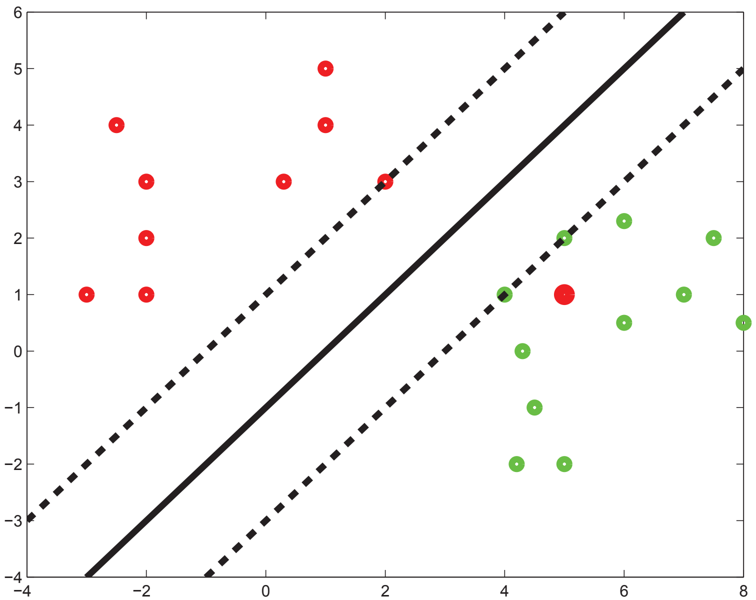
No Separable



$$\mathbf{w} = arg_{\mathbf{w}} \quad max(\frac{2}{||\mathbf{w}||}) = arg_{\mathbf{w}} min(\frac{1}{2}||\mathbf{w}||^2)$$

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \qquad \forall i = 1, ..., l$$

No Separable



$$\mathbf{w} = arg_{\mathbf{w}} min(\frac{1}{2}||\mathbf{w}||^2 + C(\sum_{i=1}^{l} \xi_i)^k)$$

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 + \xi_i \qquad \forall i = 1, ..., l$$

$$\xi_i \geq 0$$

Separable: Función de Optimizacion

$$L_p \equiv \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{l} \alpha_i y_i(\mathbf{w}^T\mathbf{x}_i + b) + \sum_{i=1}^{l} \alpha_i$$

No Separable: Función de Optimizacion

$$L_p \equiv \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{l} \xi_i - \sum_{i=1}^{l} \alpha_i y_i(\mathbf{w}^T\mathbf{x}_i + b) + \sum_{i=1}^{l} \alpha_i - \sum_{i=1}^{l} \alpha_i\xi_i - \sum_{i=1}^{l} \mu_i x_i$$

La solución en ambos casos es la misma: $\mathbf{w} = \sum_{i=1}^{N_s} \alpha_i y_i x_i$

Función de decisión: $sgn(\mathbf{w}\mathbf{x}^{\mathbf{T}} + b)$

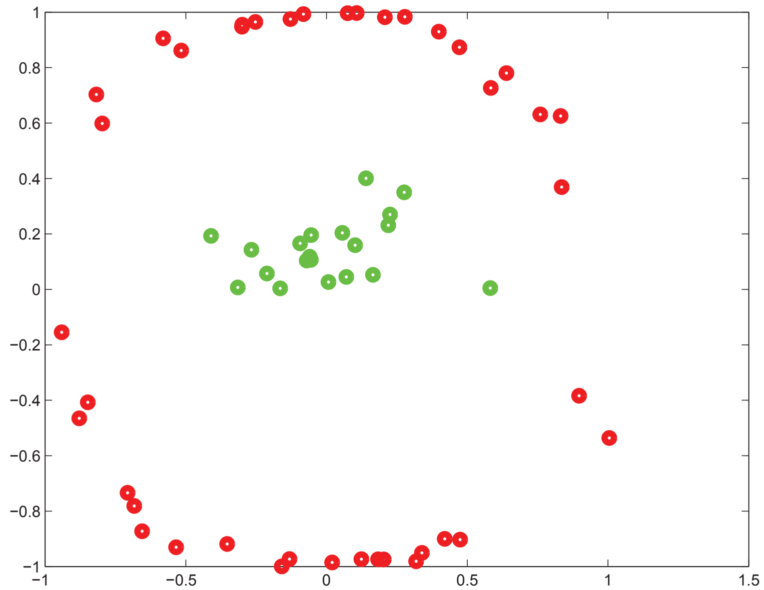
$\alpha_i \neq 0$: Support Vectors

Separable: $\alpha_i \geq 0$

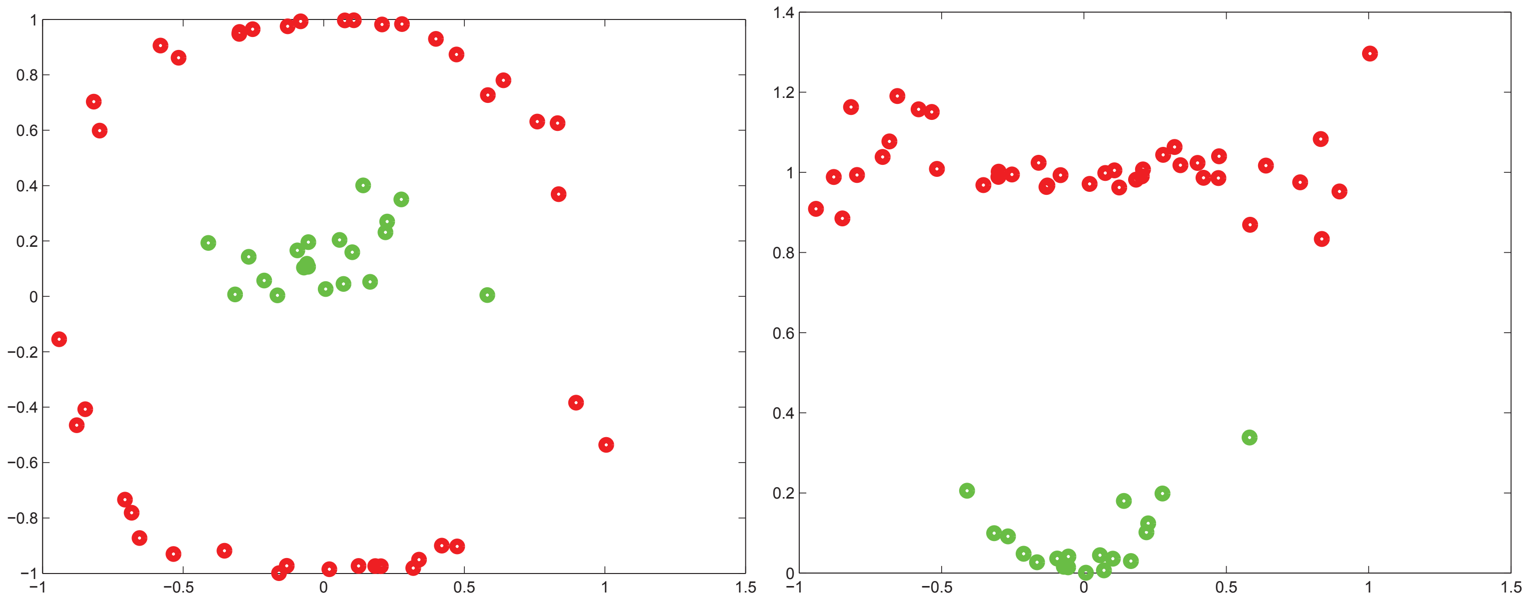
No separable: $0 \leq \alpha_i \leq C$ $\quad \alpha_i = C, \quad \xi_i \neq 0:$ Muestra no separable

$0 < \alpha_i < C, \quad \xi_i = 0:$ Muestra separable

Kernel No lineal



Kernel No lineal



$$x_1(1) = x(1)$$

$$x_1(2) = x^2(1) + x^2(2)$$

$$x_1 = \Phi(x)$$

Kernel no lineal

En ambos casos, lineal separable y no separable,

la función de optimización es la misma con diferentes restricciones

$$L_p \equiv \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \mathbf{x^T}_j$$

La solución en ambos casos es la misma: $\mathbf{w} = \sum_{i=1}^{N_s} \alpha_i y_i x_i$

Función de decisión: $sgn(\mathbf{w}\mathbf{x}^{\mathbf{T}} + b)$

Función de decisión: $sgn(\sum_{\mathbf{i=1}}^{\mathbf{N_s}} \alpha_{\mathbf{i}} \mathbf{y_i} \mathbf{x_i} \mathbf{x}^{\mathbf{T}} + b)$

Kernel no lineal

En ambos casos, lineal separable y no separable,

la función de optimización es la misma con diferentes restricciones

$$L_p \equiv \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \mathbf{x^T}_j$$

La solución en ambos casos es la misma: $\mathbf{w} = \sum_{i=1}^{N_s} \alpha_i y_i x_i$

Función de decisión: $sgn(\mathbf{w}\mathbf{x}^\mathbf{T} + b)$

Función de decisión: $sgn(\sum_{\mathbf{i=1}}^{\mathbf{N_s}} \alpha_\mathbf{i} \mathbf{y_i} \mathbf{x_i} \mathbf{x}^\mathbf{T} + b)$

Kernel no lineal

En ambos casos, lineal separable y no separable,

la función de optimización es la misma con diferentes restricciones

$$L_p \equiv \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

La solución en ambos casos es la misma: $\mathbf{w} = \sum_{i=1}^{N_s} \alpha_i y_i x_i$

Función de decisión: $sgn(\mathbf{w}\mathbf{x}^{\mathbf{T}} + b)$

Función de decisión: $sgn(\sum_{\mathbf{i=1}}^{\mathbf{N_s}} \alpha_{\mathbf{i}} \mathbf{y_i} \mathbf{K}(\mathbf{x_i}, \mathbf{x}) + b)$

Now if there were a "kernel function" K such that $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$,
we would only need to use K in the training algorithm,
and would never need to explicitly even know what Φ is.

Ref.: A Tutorial on Support Vector Machines for Pattern Recognition
CHRISTOPHER J.C. BURGES