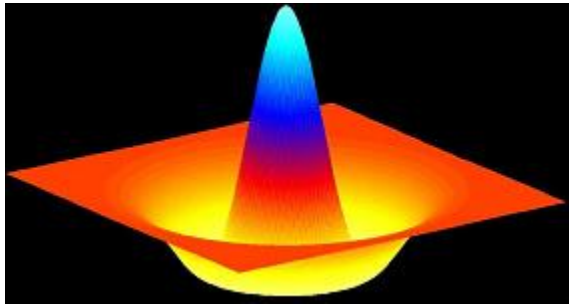


Procesamiento Digital de Imágenes

Clasificación usando Máquinas de Vectores Soporte (SVM)

Clasificación usando Vectores Soporte

Sin pérdida de generalidad, el problema de clasificación puede restringirse a un problema con dos clases.

El objetivo es separar las dos clases con una función que pueda inducirse a partir de los ejemplos disponibles de cada clase.

Es deseable que el clasificador generalice bien, es decir funcione bien con ejemplos no usados para su entrenamiento.

Consideremos el ejemplo de Fig. 1, donde puede verse que existen numerosos clasificadores lineales (hiperplanos) que separan los datos, pero donde hay un único hiperplano que maximiza la distancia entre él mismo y el punto más cercano de cada clase. Dicho hiperplano de denomina **hiperplano de separación óptimo**.

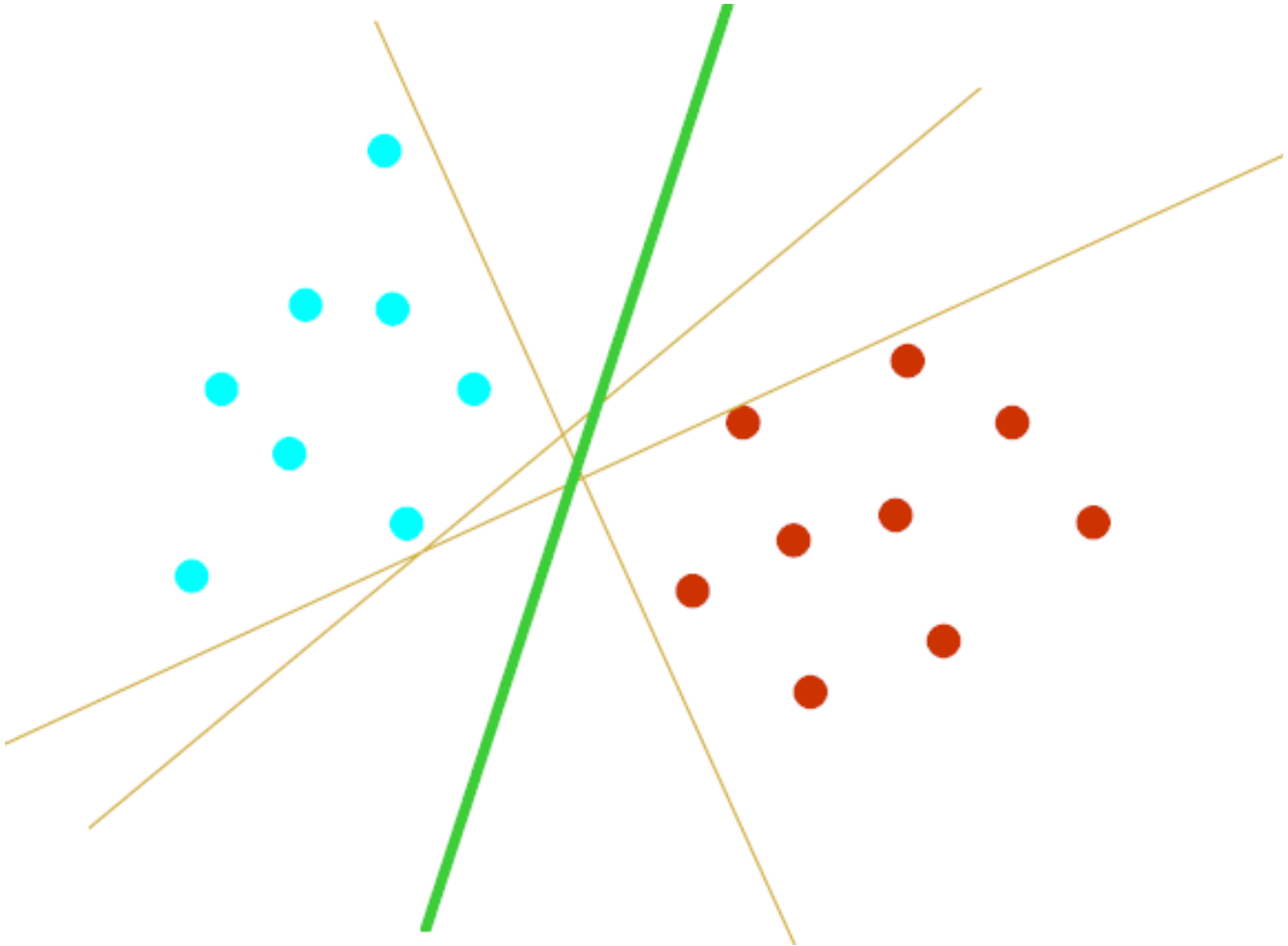


Figura 1: Hiperplano óptimo de separación (en verde).

Consideremos un conjunto de datos de entrenamiento pertenecientes a dos clases diferentes

$$(y_1, x_1), \cdots, (y_\ell, x_\ell), \quad x_j \in \mathbb{R}^n, y_j \in \{-1, +1\} \qquad (1)$$

La ecuación de un hiperplano en el espacio $\mathbb{R}^n$ puede escribirse como:

$$w^T x + b = 0 \qquad (2)$$

Un conjunto de datos se dice que están óptimamente separados por un hiperplano, si están separados sin error y la distancia entre el vector más cercano al hiperplano y el hiperplano es máxima.

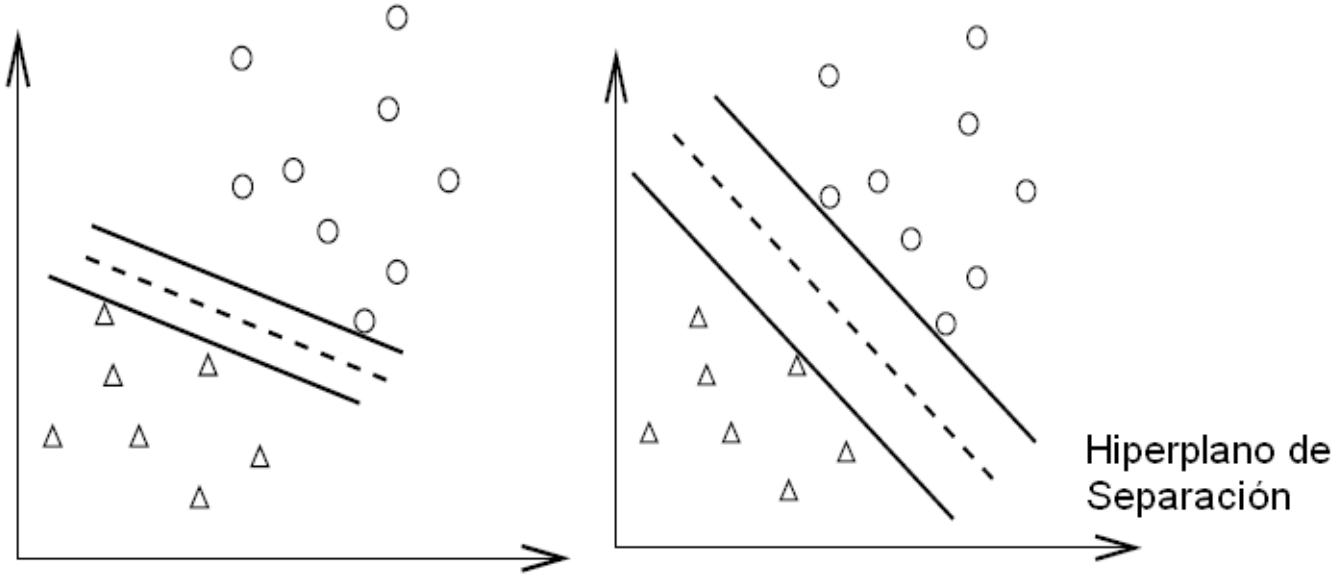


Fig. 2: Clases linealmente separables

Se tiene entonces que:

$$\begin{cases} w^T x_i + b > 0 \quad \text{si} \quad y_i = 1 \\ w^T x_i + b < 0 \quad \text{si} \quad y_i = -1 \end{cases} \tag{3}$$

La **función de decisión** resulta:

$$f(x) = sign\left(w^T x + b\right) \tag{4}$$

Existe cierta redundancia en los parámetros *w* y *b* en (2).

La ecuación (2) se suele normalizar imponiendo la condición

$$\min_{x_j} \left| w^T x + b \right| = 1 \qquad (5)$$

que indica que la norma del vector de peso *w* debe ser igual a la inversa de la distancia entre el punto más cercano al hiperplano y el hiperplano. Con esta normalización resulta:

$$\begin{cases} w^T x_i + b \geq 1 & \text{si} \quad y_i = 1 \\ w^T x_i + b \leq -1 & \text{si} \quad y_i = -1 \end{cases} \qquad (6)$$

O equivalentemente

$$y_i \left[w^T x_i + b \right] \geq 1 \qquad (7)$$

La distancia entre un punto *x* y el hiperplano (*w, b*) resulta entonces

$$d\left(w,b;x\right)=\frac{\left|w^{T}x+b\right|}{\left\|w\right\|} \tag{8}$$

El hiperplano de separación **óptimo** es el que **maximiza** el margen $\rho(w,b)$ sujeto a la restricción (7), definido como:

$$\rho(w,b)=\min_{\{x_i:y_i=1\}} d(w,b;x_i)+\min_{\{x_j:y_j=-1\}} d(w,b;x_j)$$

$$=\min_{\{x_i:y_i=1\}} \frac{\left|w^{T}x_i+b\right|}{\left\|w\right\|}+\min_{\{x_j:y_j=-1\}} \frac{\left|w^{T}x_j+b\right|}{\left\|w\right\|}$$

$$=\frac{1}{\left\|w\right\|}\left(\min_{\{x_i:y_i=1\}}\left|w^{T}x_i+b\right|+\min_{\{x_j:y_j=-1\}}\left|w^{T}x_j+b\right|\right)$$

$$=\frac{2}{\left\|w\right\|} \tag{9}$$

Es decir, el hiperplano que separa en forma óptima los datos es el que **minimiza** el funcional

$$\Phi(w) = \frac{1}{2}\|w\|^2 \qquad (10)$$

sujeto a las restricciones (7).

Hasta aquí se asumió que los datos de entrenamiento eran linealmente separables. Sin embargo, este no es el caso en general. Para contemplar posibles errores de clasificación, y que el problema tenga solución, se introduce un término adicional en la función de costo (10).

La Fig. 3 representa el caso general donde hay errores de clasificación (las clases no son linealmente separables por un hiperplano).

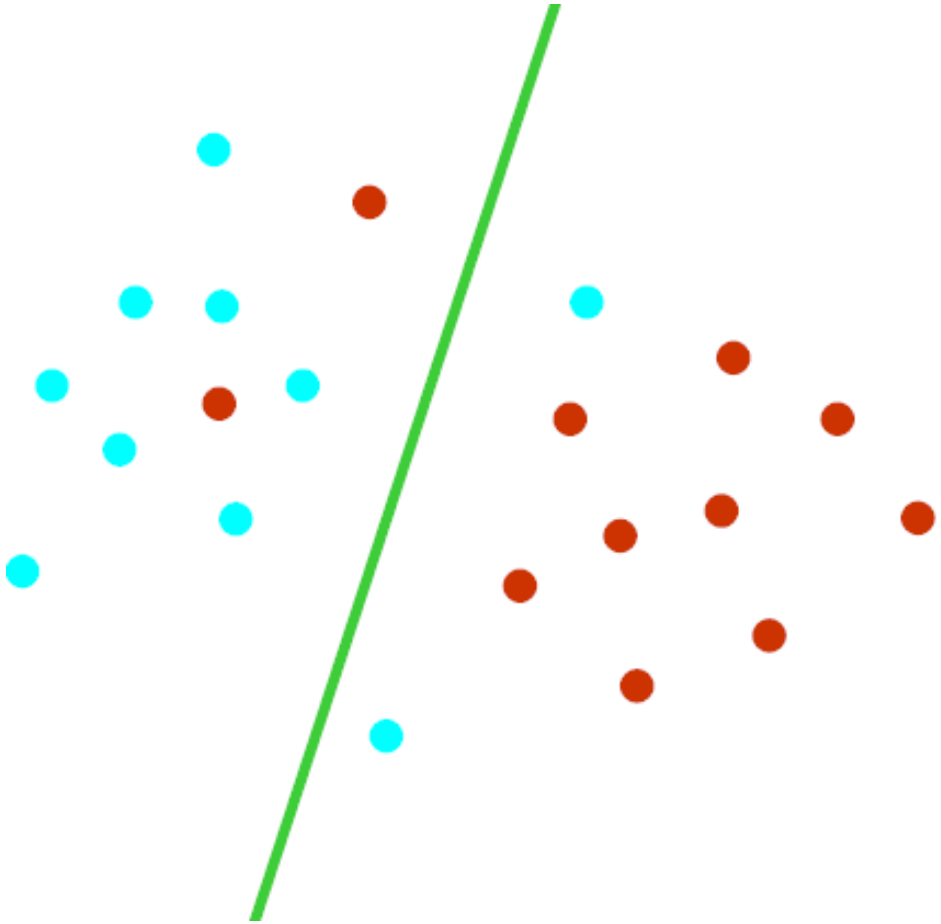


Figura 3: Hiperplano de separación óptimo generalizado (se permiten errores de clasificación).

La restricción (7) se escribe ahora como

$$y_i \left[w^T x_i + b \right] \geq 1 - \xi_i \quad , \quad i = 1, \cdots, \ell \tag{11}$$

donde $\xi_i \geq 0$ es una medida del error de clasificación (**slack variable**).

El hiperplano de separación óptimo generalizado está ahora determinado por el vector w que minimiza la función de costo

$$\Phi(w, \xi) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} \xi_i \tag{12}$$

sujeto a la restricción (11), donde $C > 0$ es una constante que pondera el compromiso entre la complejidad del clasificador y el error de clasificación.

El problema de clasificación consiste entonces en resolver el siguiente problema de optimización con restricciones

$$\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}\xi_i$$

$$\text{subject to} \quad y_i\left(w^T x_i + b\right) \geq 1 - \xi_i,$$

$$\xi_i \geq 0, \quad i = 1, \cdots, \ell \qquad (13)$$

- Si $\xi_i > 1 \Rightarrow x_i$ **no está en el lado correcto** del hiperplano de separación.
- Si C es grande la mayoría de los ξ_i son nulos

En muchos casos las clases no son separables por un hiper-plano, sino por una hipersuperficie (no lineal). Por ejemplo:

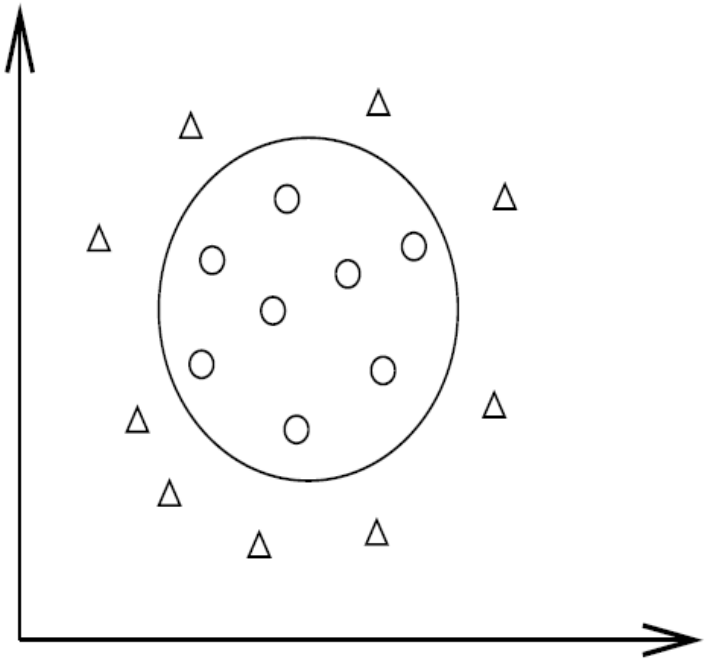


Fig. 4: Clases no separables linealmente (por un hiperplano).

Se realiza una transformación (no lineal) $\varphi(x)$ desde el espacio original de entrada (**input or *primal* space**), donde las clases no son separables linealmente, a un espacio de mayor dimensión, que podría ser infinito dimensional, (**feature or *dual* space**), donde si lo son.

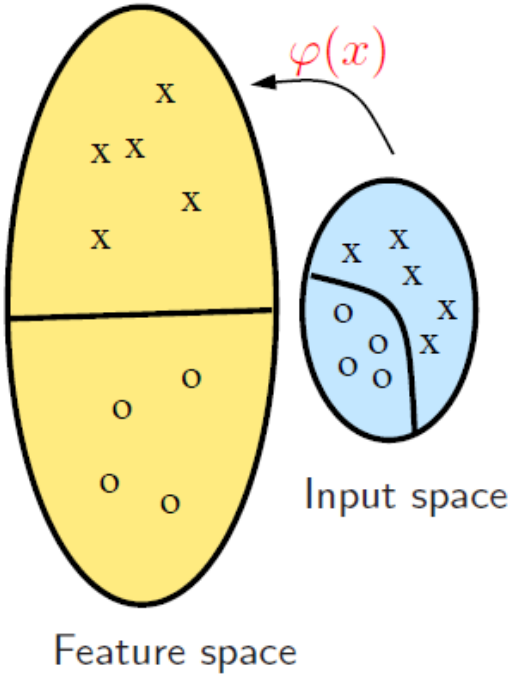


Fig. 5a: Transformación no lineal a un espacio de mayor dimensión.

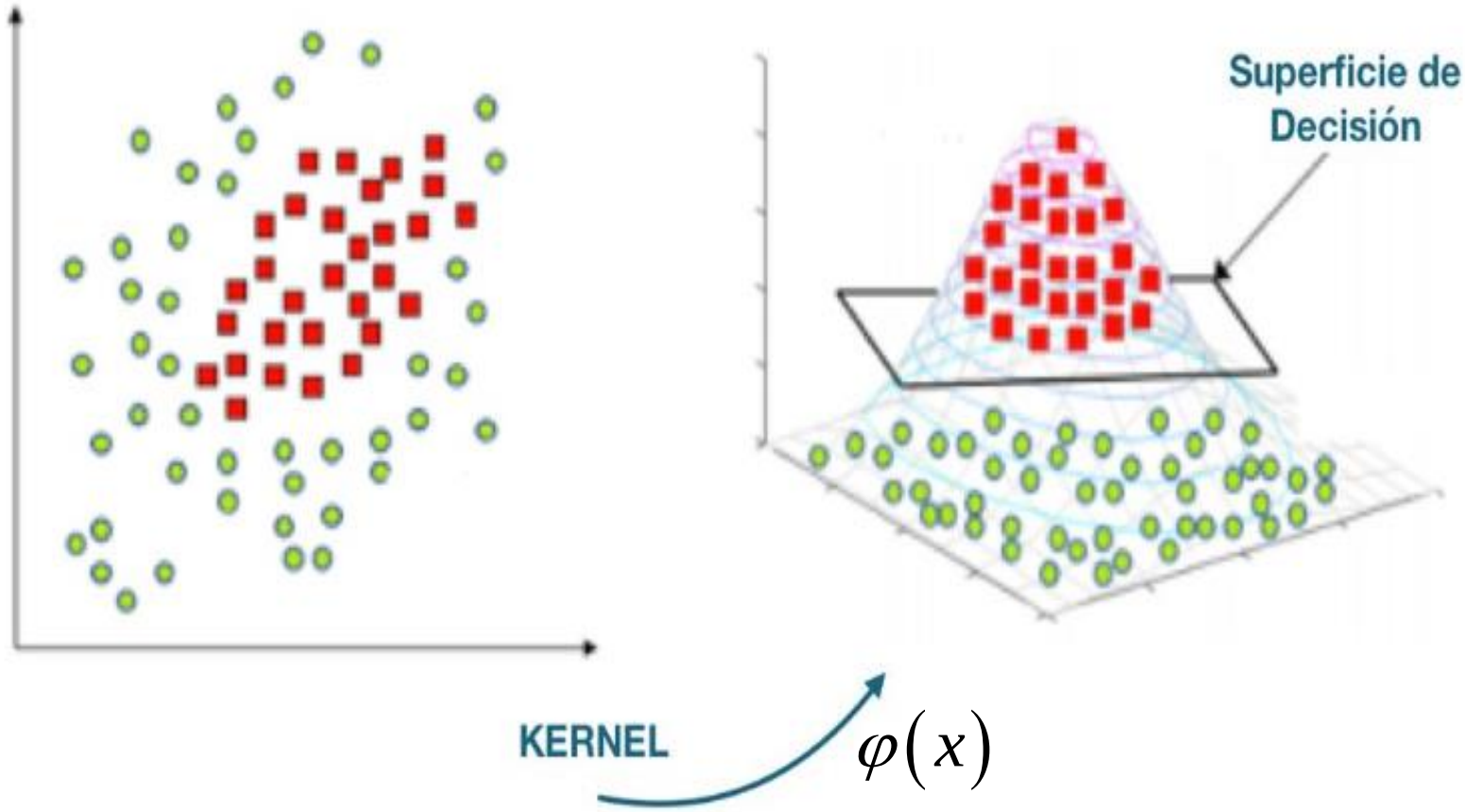


Fig. 5b: Ejemplo deTransformación no lineal a un espacio de mayor dimensión.

El problema en el espacio **primal** resulta ahora

$$\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}\xi_i$$

$$\text{subject to} \ \ y_i\left(w^T\varphi(x_i)+b\right) \geq 1-\xi_i,$$

$$\xi_i \geq 0, \quad i=1,\cdots,\ell \tag{14}$$

[Cortes and Vapnik, 1995]

C: compromiso entre “error de entrenamiento” y “generalización”.

La transformación $\varphi(x)$ no es necesario definirla explícitamente si se recurre al denominado **truco del kernel (kernel trick)**, definiendo una función de kernel

$$K(x_i, x_j) = \varphi(x_i)^T \varphi\left(x_j\right)$$

Kernels más usados:

$$K(x, x_i) = x_i^T x \qquad \text{(lineal)}$$

$$K(x, x_i) = \left(x_i^T x + \tau\right)^d \quad \text{con} \quad \tau \geq 0 \qquad \text{(polinomial)}$$

$$K(x, x_i) = \exp\left(-\left\|x - x_i\right\|_2^2 / \sigma^2\right) \qquad \text{(RBF Gaussiana)}$$

$$K(x, x_i) = \tanh\left(k x_i^T x + \theta\right) \qquad \text{(MLP)}$$

Normalmente el problema de optimización en el espacio primal (14), se puede resolver más fácilmente en el espacio dual, recurriendo a los **multiplicadores de Lagrange**. En el espacio dual, el problema es equivalente a

$$\max_{\alpha}\left[-\frac{1}{2}\sum_{i,j=1}^{\ell} y_i y_j K(x_i,x_j)\alpha_i\alpha_j + \sum_{j=1}^{\ell}\alpha_j\right]$$

$$\text{subject to}\begin{cases}\displaystyle\sum_{i=1}^{\ell}\alpha_i y_i \\ 0 \le \alpha_i \le c, \forall i\end{cases} \qquad (15)$$

donde α_i son los multiplicadores de Lagrange. El problema es un problema **QP (Quadratic Programming)** convexo.

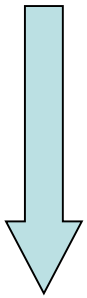
Representación del Clasificador

Espacio Primal: Representación paramétrica

$$y_i = sign(w^T \varphi(x_i) + b)$$

clasificador

$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$$



Kernel Trick

Espacio Dual: Representación no paramétrica

$$y = sign(\sum_{i=1}^{\#SV} \alpha_i y_i K(x, x_i) + b)$$

clasificador

SV: número de **vectores soporte.** Son una porción limitada de los datos de entrenamiento que se utilizan para definir la frontera de separación de las clases.

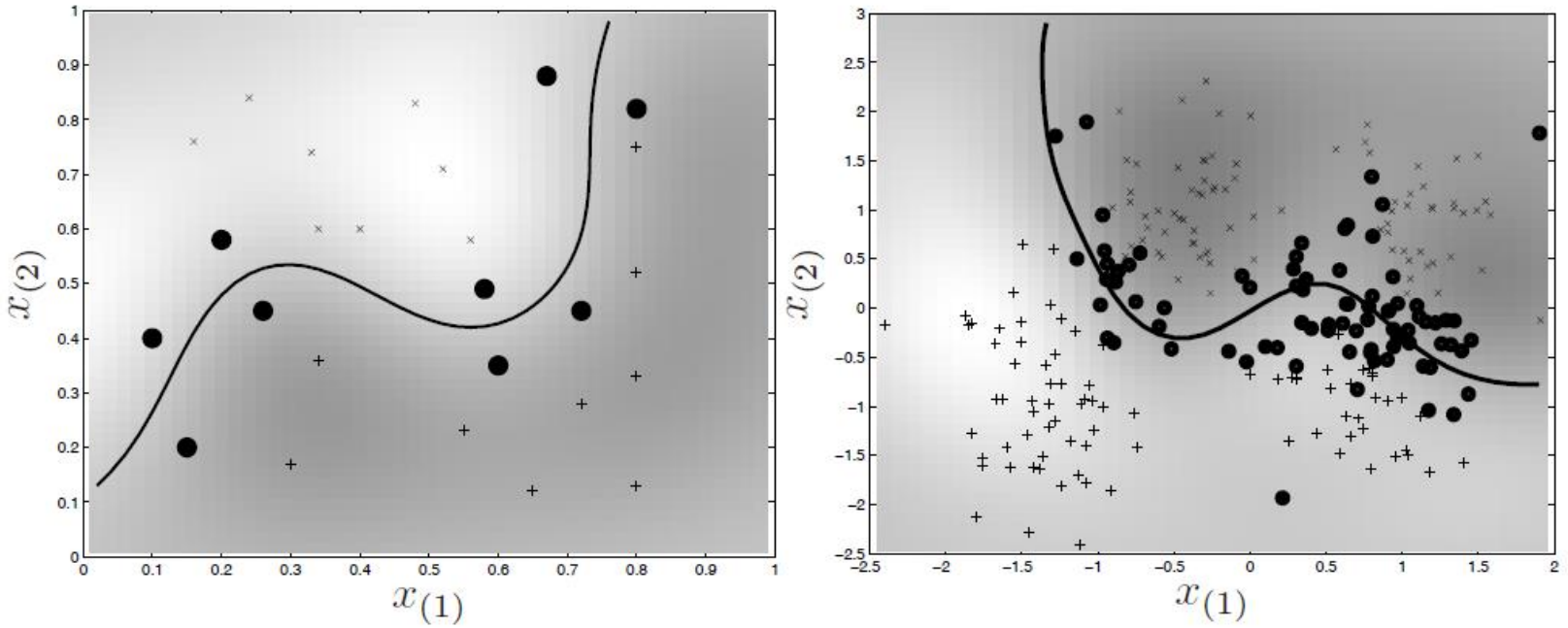


Fig. 6: Vectores soporte