SVM

Clases linealmente no Separables

Aprendizaje Automático

Dr. Luis Eduardo Falcón Morales
ITESM
Campus Guadalajara

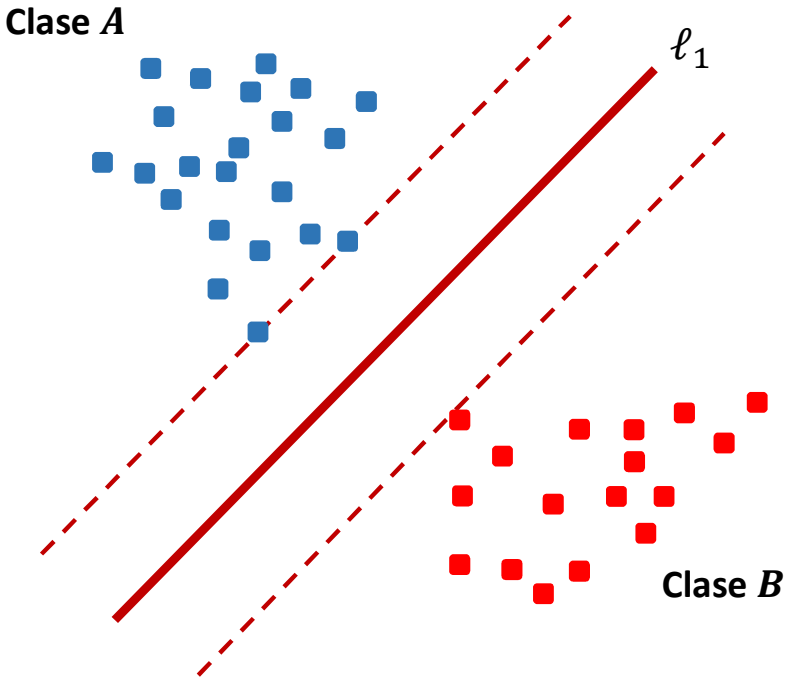
SVM para casos linealmente no separables:

En ocasiones los conjuntos no son linealmente separables, por lo que deberá ajustarse el método SVM para su aplicación en dichos casos.

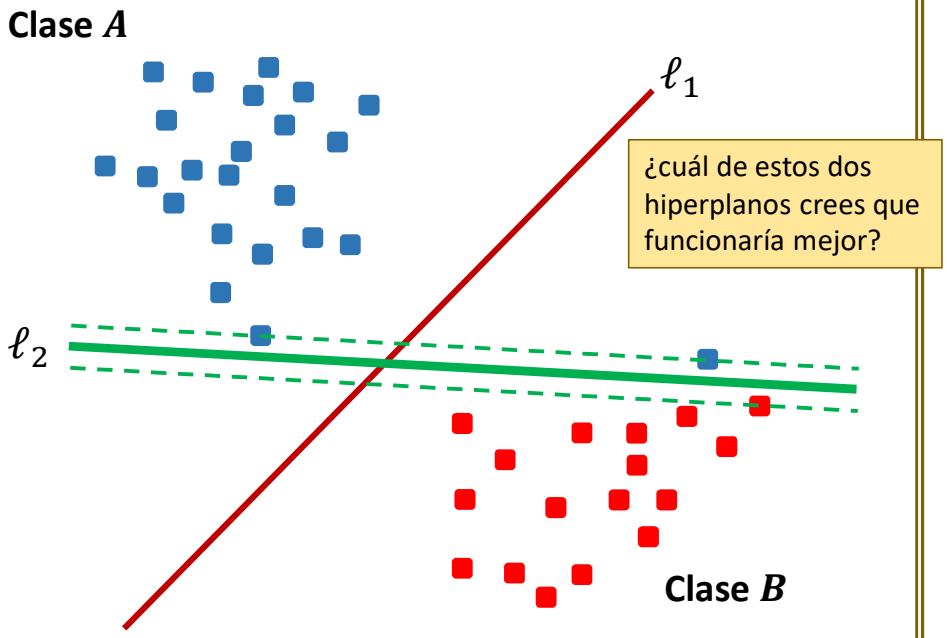
Inclusive aunque en ocasiones las clases sean linealmente separables, se hará también un ajuste al método para evitar que la solución sea muy sensible al ruido.

Aunque podamos encontrar un Hiperplano de Margen Máximo (HMM) de dos clases linealmente separables, queremos que su variabilidad o sensibilidad no sea muy grande con respecto a nuevas observaciones.

Además, mientras menor sea el margen del HMM, mayor probabilidad de tener sobre-entrenamiento.



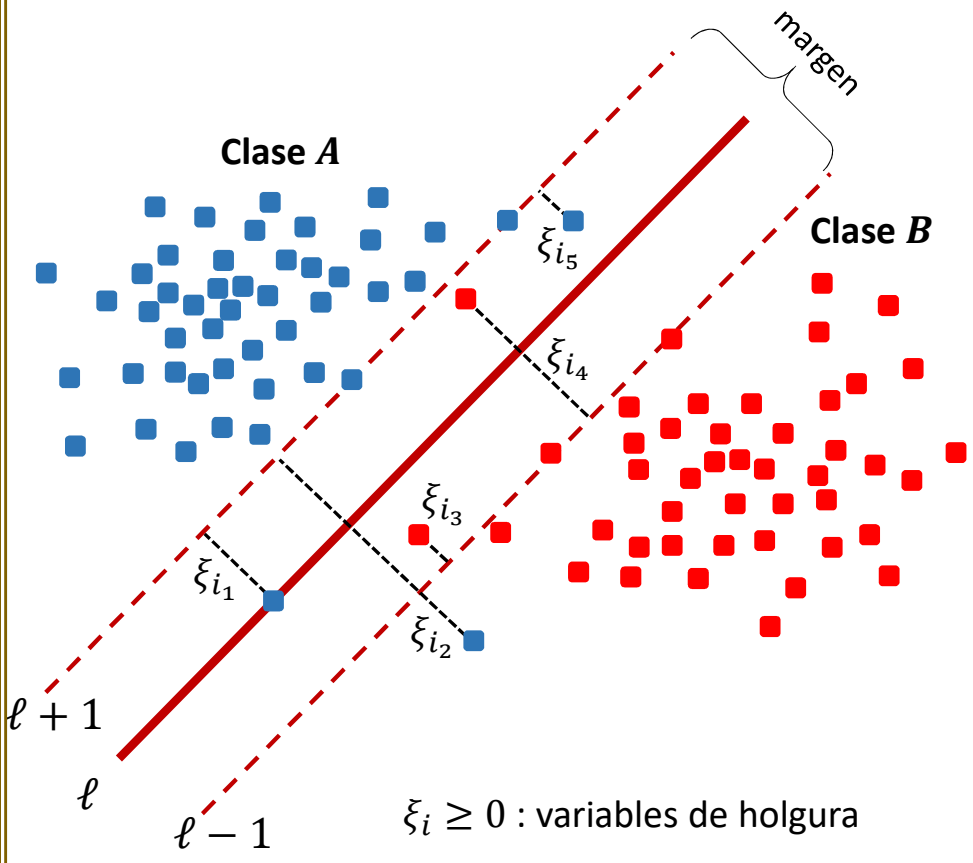
El hiperplano HMM fue encontrado.



Se agrega un solo dato atípico/outlier y el nuevo hiperplano (HMM) cambia drásticamente: ℓ_2. Estrictamente ℓ_1 ya no es un HMM, pero ¿convendría cambiarlo por ℓ_2?

Caso: Clases No Linealmente Separables

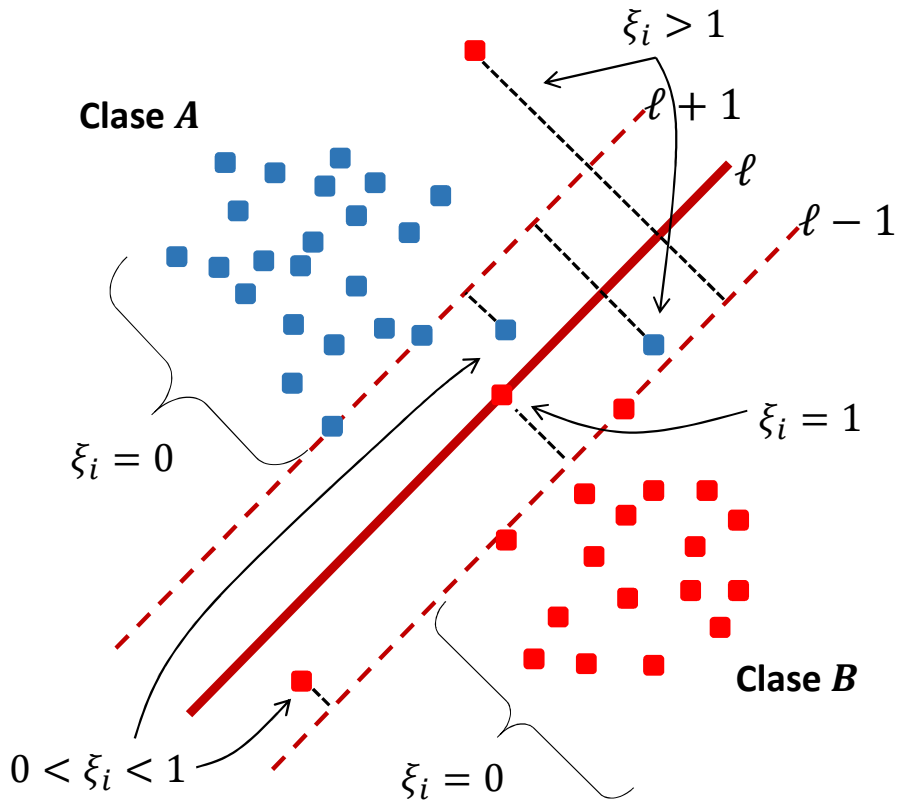
Variables de Holgura



$\xi_i \geq 0$: variables de holgura

Significado Geométrico de las Variables de Holgura

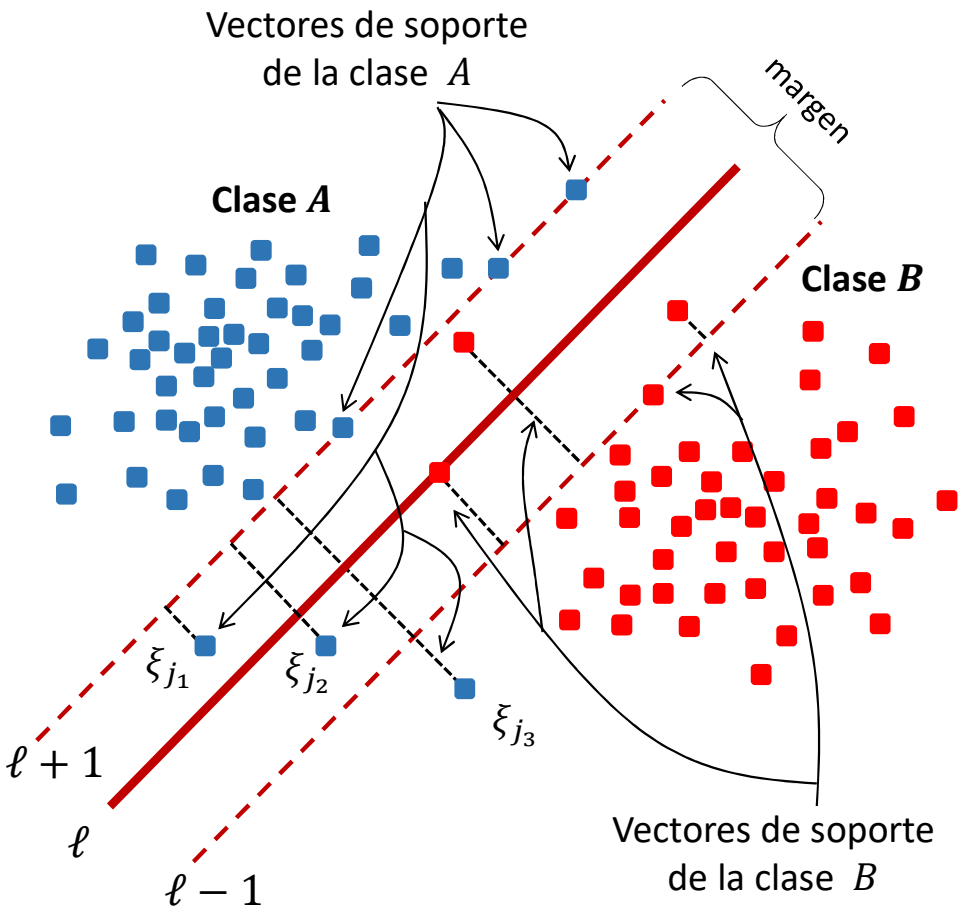
El objetivo ahora será **maximizar** el margen, a la vez que se **minimicen** los valores de las variables de holgura.



- Estrictamente la solución general del problema de encontrar un HMM con las restricciones de las variables de holgura es un problema NP-hard con respecto a la dimensión del espacio de vectores.

- Sin embargo, se incluirán varias restricciones para encontrar una solución que sea viable, aunque estrictamente no sea la solución general inicial.

Caso: Clases No Linealmente Separables



Si las clases no son linealmente separables se introducen las llamadas variables de holgura, ξ_k, (*slack variables*) las cuales nos llevan a crear un margen suave (*soft margin*) sobre el cual se les permite permanecer a las variables mal clasificadas.

El problema de optimización (minimización) queda ahora como sigue, donde nuevamente se normalizan los términos constantes:

$$\min_{\vec{\omega},b,\xi_j} \; J(\vec{\omega}) = \frac{1}{2}\|\vec{\omega}\|^2 + C\sum_{j=1}^{n}\xi_j$$

sujeto a las restricciones

$$y_j\left(\vec{\omega}\cdot\vec{x}_j + b\right) \geq 1 - \xi_j, \qquad \xi_j \geq 0$$

donde $j = 1,2,\ldots,n$, $C \geq 0$, además de estar acotados los valores de las observaciones mal clasificadas mediante la sumatoria: $\sum_{j=1}^{n}\xi_j$

Veamos algunas características de las variables de holgura y la constante de penalización $C \geq 0$:

$$\min J(\vec{\omega}) = \frac{1}{2}\|\vec{\omega}\|^2 + C\sum_j \xi_j$$

- El valor $C = 0$, se asignaría cuando tengamos $\xi_j = 0,\ \forall j$, y estamos en el caso de conjuntos linealmente separables.

- La constante $C \neq 0$, se aplica así a todos aquellos puntos $\vec{x}_j$, que quedan mal clasificados o dentro del margen del HMM.

- La constante C es una especie de regularización, como en el caso de regresión.

- La constante C es llamada generalmente el **parámetro de complejidad o de parametrización**, porque ayuda e encontrar un balance entre la maximización del margen y la minimización de la cantidad de variables de holgura.

- Cuando la constante C es muy grande, la cantidad de variables de holgura debe disminuir, así como la distancia de que se alejen del HHM los vectores mal clasificados.

- Cuando la constante C es muy pequeña, la cantidad de variables de holgura podrá aumentar, así como la distancia a la que pueden alejarse del HHM los vectores mal clasificados.

- El valor de C generalmente se determina mediante *cross-validation*.

- C ayuda a prevenir el sobre-entrenamiento.

De manera análoga al caso de clases linealmente separables, se llega al problema dual del problema de las máquinas de vectores de soporte para el caso de clases no linealmente separables, es decir:

La función de costo a optimizar mediante los multiplicadores de Lagrange:

$$G\left(\vec{\omega}, b, \xi_j, \lambda_j, \mu_j\right) = \frac{1}{2}\|\vec{\omega}\|^2 + C\sum_{k=1}^{n}\xi_k - \sum_{k=1}^{n}\lambda_k\{y_k(\vec{\omega}\cdot\vec{x}_k + b) - (1-\xi_k)\} - \sum_{k=1}^{n}\mu_k\xi_k$$

Agrupando algebraicamente los términos que contienen los ξ_k se puede expresar lo anterior en términos de la función objetivo del caso de clases linealmente separables $\mathcal{L}$ visto previamente, es decir:

$$G\left(\vec{\omega}, b, \xi_j, \lambda_j, \mu_j\right) = \mathcal{L}\left(\vec{\omega}, b, \lambda_j\right) + \sum_{k=1}^{n}(C - \lambda_k - \mu_k)\xi_k$$

$$G(\vec{\omega}, b, \xi_j, \lambda_j, \mu_j) = \frac{1}{2}\|\vec{\omega}\|^2 + C\sum_{k=1}^{n}\xi_k - \sum_{k=1}^{n}\lambda_k\{y_k(\vec{\omega}\cdot\vec{x}_k + b) - (1-\xi_k)\} - \sum_{k=1}^{n}\mu_k\xi_k$$

Y nuevamente, de las condiciones necesarias de Karush-Kuhn-Tucker (KKT) para un mínimo local, obtenemos:

De las derivadas parciales:

$$\frac{\partial G}{\partial \vec{\omega}} = 0 \quad \Rightarrow \quad \vec{\omega} = \sum_{k=1}^{n}\lambda_k y_k \vec{x}_k$$

$$\frac{\partial G}{\partial b} = 0 \quad \Rightarrow \quad \sum_{k=1}^{n}\lambda_k y_k = 0$$

$$\frac{\partial G}{\partial \xi_j} = 0 \Rightarrow \quad C - \lambda_j - \mu_j = 0$$

para $j = 1,2,\ldots,m$.

Y de los multiplicadores de Lagrange para $j = 1,2,\ldots,m$

$$\lambda_j\{y_j(\vec{\omega}\cdot\vec{x}_j + b) - 1 + \xi_j\} = 0,$$

$$\mu_j\xi_j = 0$$

$$\lambda_k \geq 0,\ \mu_k \geq 0,$$

Observa que como $C = \lambda_j + \mu_j$, además de que $\lambda_k \geq 0,\ \mu_k \geq 0,$ entonces en particular $0 \leq \lambda_k \leq C$, la cual es la única restricción que cambiará con respecto al problema dual del modelo de clases linealmente separables.

Problema Dual de Optimización: Clases Linealmente No Separables

Nuevamente , de manera análoga
se obtiene el problema dual:

$$\max_{\lambda_1,\ldots,\lambda_n} \sum_{k=1}^{n} \lambda_k - \frac{1}{2}\left\{\sum_{k=1}^{n}\sum_{j=1}^{n} \lambda_k \lambda_j y_k y_j \vec{x}_k \cdot \vec{x}_j\right\}$$

sujeto a las restricciones:

$$\begin{cases} 0 \le \lambda_k \le C, \text{ para } k = 1,2,\ldots,n \\ \sum_{k=1}^{n} \lambda_k y_k = 0 \end{cases}$$

a partir del conjunto de datos de entrada:

$$\{(x_{k1}, x_{k2}, \ldots, x_{km}, y_k)\}_{k=1}^{n}$$

donde $\vec{x}_k = (x_{k1}, x_{k2}, \ldots, x_{km})$

$$y_k = \begin{cases} +1 & si \quad \vec{x}_k \in A \\ -1 & si \quad \vec{x}_k \in B \end{cases}$$

Y una vez obtenidos los multiplicadores de Lagrange reconstruimos el hiperplano separador:

$$\vec{\omega} = \sum_{k=1}^{n} \lambda_k y_k \vec{x}_k \ , \qquad b = y_k - \vec{\omega} \cdot \vec{x}_k$$

Y entonces dado un nuevo dato de entrada

$$\vec{z} = (z_1, z_2, \ldots, z_m)$$

podremos determinar la clase a la cual pertenece.
Por ejemplo, suponiendo que tenemos un problema biclase linealmente no separable, entonces:

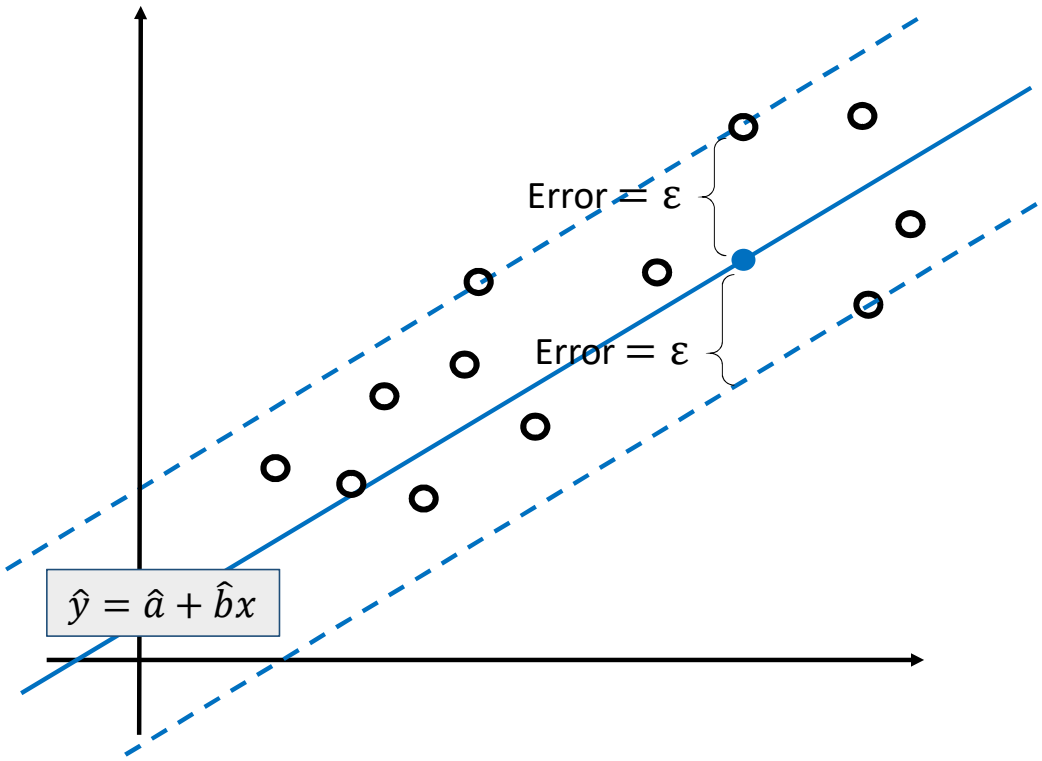
$$g(\vec{z}) = signo\{\vec{\omega} \cdot \vec{z} + b\} = \begin{cases} \ge 0 & entonces \quad \vec{z} \in A \\ < 0 & entonces \quad \vec{z} \in B \end{cases}$$

Complejidad de algoritmo SVM

- La complejidad durante el proceso de entrenamiento es $\sigma(n^2)$ donde n es el número de datos de entrenamiento.

- La complejidad durante el proceso de ejecución del modelo es $\sigma(kd)$ donde k es el número de vectores de soporte y d la dimensión de los datos de entrada.

SVM en problemas de Regresión:
Regresión de Vectores de Soporte
Support Vector Regression : SVR



$$\min_{\vec{\omega},b} \; J(\vec{\omega}) = \frac{1}{2}\|\vec{\omega}\|^2$$

sujeto a las restricciones

$$y_j - \vec{\omega} \cdot \vec{x}_j - b \leq \varepsilon$$

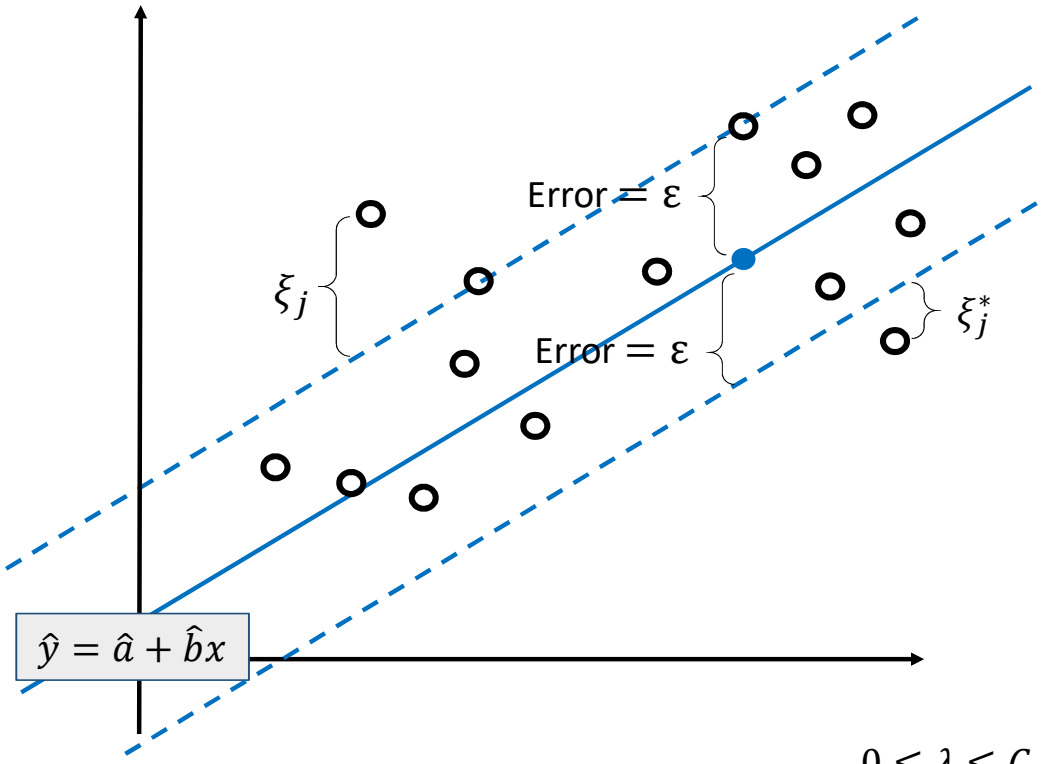
$$\vec{\omega} \cdot \vec{x}_j + b - y_j \leq \varepsilon$$

donde $\varepsilon \geq 0$

Sin considerar variables de holgura

$$\hat{y} = \sum_{j=1}^{n} \lambda_k \langle \vec{x}_k, \vec{x} \rangle + b$$

Support Vector Regression : SVR



$$\min_{\vec{\omega},b,\xi_j} \; J(\vec{\omega}) = \frac{1}{2}\|\vec{\omega}\|^2 + C\sum_{j=1}^{n}\left(\xi_j + \xi_j^*\right)$$

sujeto a las restricciones

$$y_j - \vec{\omega}\cdot\vec{x}_j - b \le \left(\varepsilon + \xi_j\right)$$

$$\vec{\omega}\cdot\vec{x}_j + b - y_j \le \left(\varepsilon + \xi_j^*\right)$$

donde $\varepsilon, \xi_j, \xi_j^* \ge 0$

Considerando variables de holgura

$$\hat{y} = \sum_{j=1}^{n}(\lambda_k - \lambda_k^*)\langle\vec{x}_k, \vec{x}\rangle + b$$

$0 \le \lambda \le C$
donde $\lambda = max\{\lambda_k, \lambda_k^*\}$