APRENDIZAJE DE MÁQUINA

MODELOS NO LINEALES I

AGENDA

01 **Máquina de vectores de soporte**

FALTA

02 **K Vecinos**

FALTA

03 **Árboles de decisión**

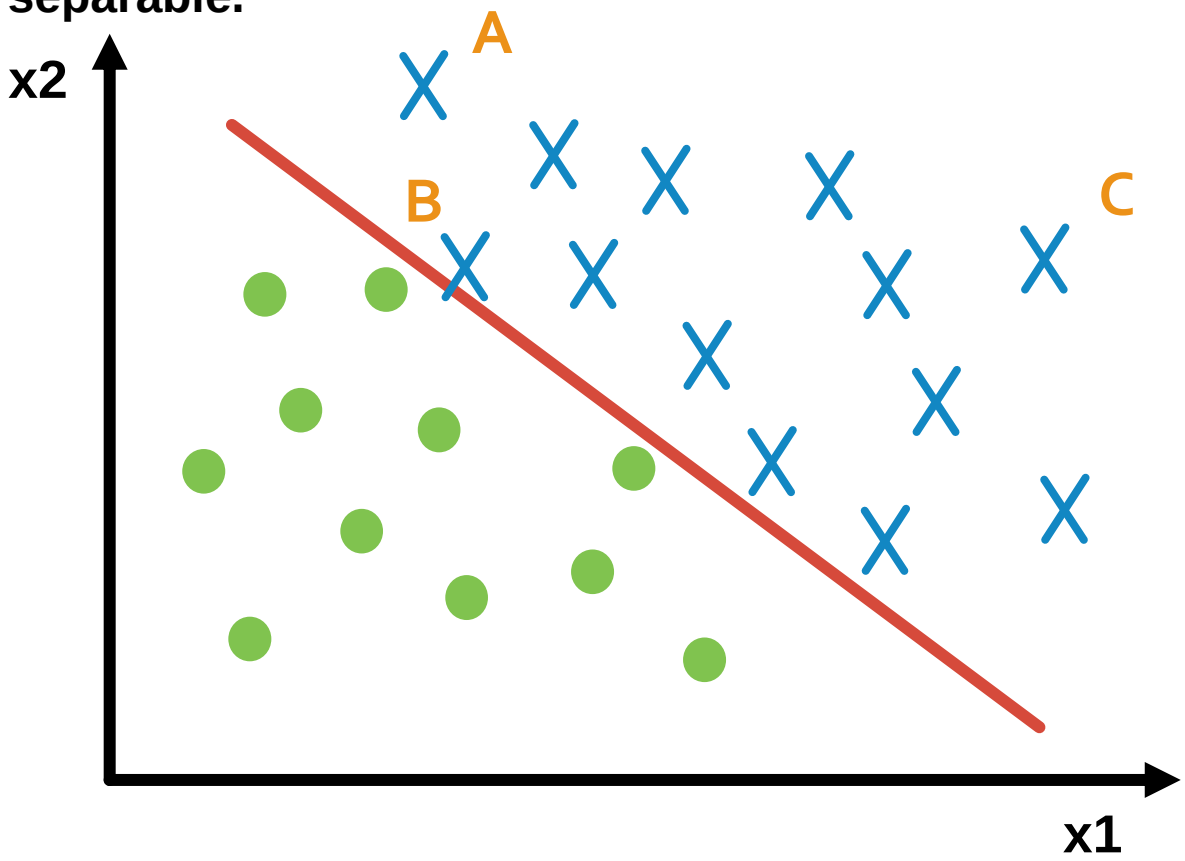
FALTA

MÁQUINA DE VECTORES DE SOPORTE

MÁQUINA DE SV

INTRODUCCIÓN

Se **asume** que **tenemos** un **conjunto de datos de entrenamiento** que es **linealmente separable.**



MÁQUINA DE SV

CAMBIO DE NOTACIÓN

Para **proceder**, se va a **realizar** un **cambio** de **notación**:

$$\boldsymbol{y} \in \{\boldsymbol{-1}, \boldsymbol{1}\}$$

$$\boldsymbol{h_{w,b}} \in \{\boldsymbol{-1}, \boldsymbol{1}\}$$

Por lo **tanto** se tiene que:

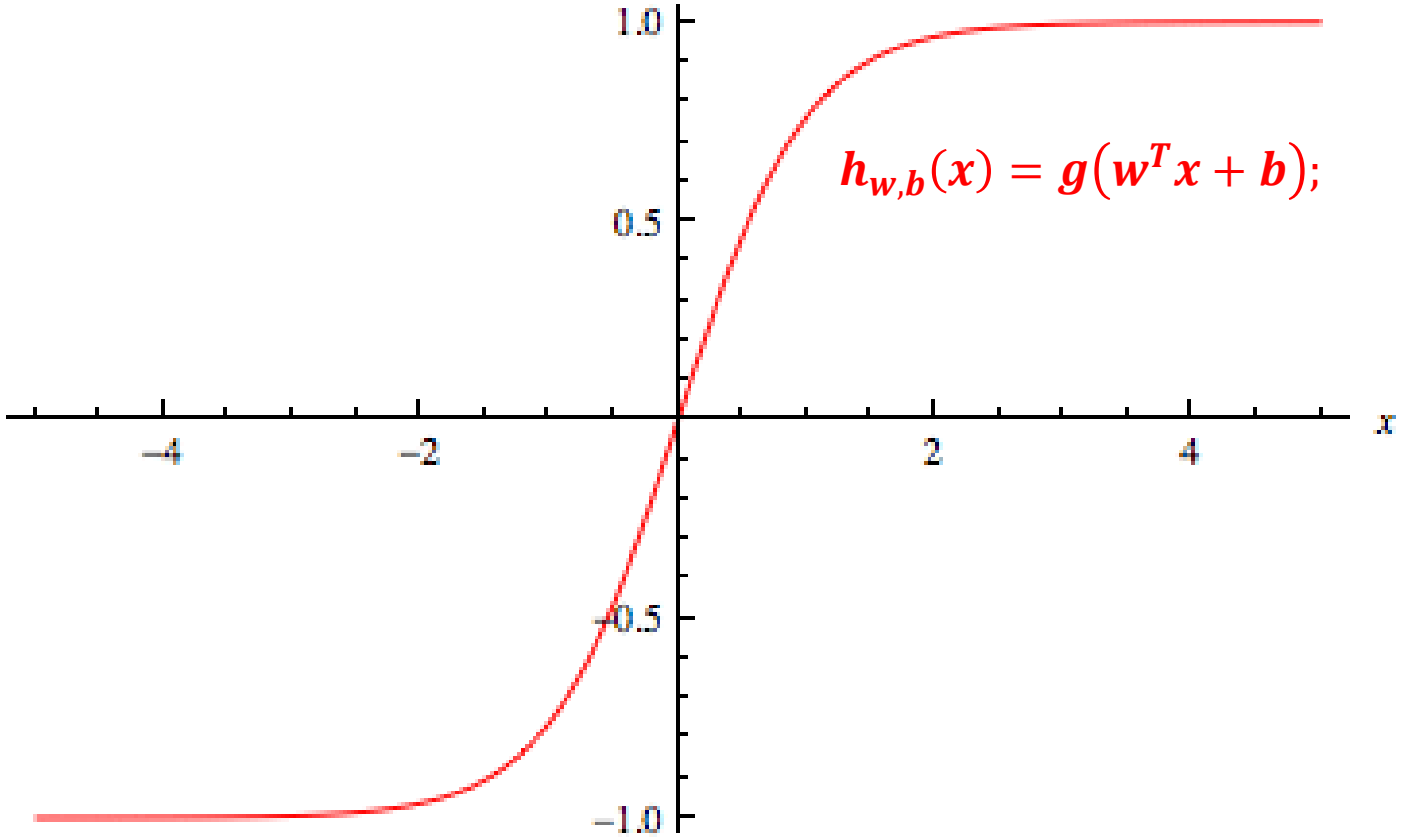
$$\boldsymbol{g(z)} \begin{cases} \boldsymbol{1\ sí\ z \geq 0} \\ \boldsymbol{-1\ de\ otra\ forma} \end{cases}$$

$$\boldsymbol{h_{w,b}(x) = g\left(w^T x + b\right); w \in \mathbb{R}^n, x \in \mathbb{R}^n}$$

(Se manipula el sesgo de manera independiente)

MÁQUINA DE SV

INTUICIÓN DE MARGEN



$$h_{w,b}(x) = g(w^T x + b);$$

Se tienen **dos** cuestiones:

$$w^T x \to \infty$$
$$g(z) \to 1$$

$$w^T x \to -\infty$$
$$g(z) \to -1$$

MÁQUINA DE SV

MARGEN FUNCIONAL

Se define el **margen funcional respecto** a un **dato** de **entrenamiento** $\left(x^{(i)}, y^{(i)}\right)$ **como:**

$$\widehat{\gamma^{(i)}} = y^{(i)}\left(w^T x^{(i)} + b\right)$$

Un **margen funcional grande** representa una **predicción correcta** y **certera**.

Variable de respuesta $y^{(i)}$	Margen funcional $\widehat{\gamma^{(i)}}$
1	$w^T x^{(i)} \ll 0$ $\widehat{\gamma^{(i)}} \ll 0$
	$w^T x^{(i)} \gg 0$ $\widehat{\gamma^{(i)}} \gg 0$
-1	$w^T x^{(i)} \ll 0$ $\widehat{\gamma^{(i)}} \gg 0$
	$w^T \gg 0$ $\widehat{\gamma^{(i)}} \ll 0$

MÁQUINA DE SV

MARGEN FUNCIONAL

El problema del **margen funcional**, es que **solo depende** del **signo** de $y^{(i)}$ **y no** de la **magnitud** de $w^T x^{(i)} + b$

Es decir, puedo **"forzar arbitrariamente"** que la predicción sea la correcta tan solo escalando w**.**

$$\gamma^{(i)} = y^{(i)}\left(w^T x^{(i)} + b\right)$$

$\gamma^{(i)} \gg 0$ **sí y solo sí:**

$$w = kw$$
$$b = kb$$
$$k \gg 0$$

Una **solución** es **normalizar** el **vector** w **(se verá más adelante):**

$$\frac{w}{\|w\|_2}$$

MÁQUINA DE SV

MARGEN FUNCIONAL

En este caso, nos **centraremos** en el **dato** de **entrenamiento,** que **proporciona** el **peor margen funcional** de **todos, es decir**:

$$\hat{\gamma} = min_{i=1,2,\ldots,m}\ \widehat{\gamma^{(i)}}$$

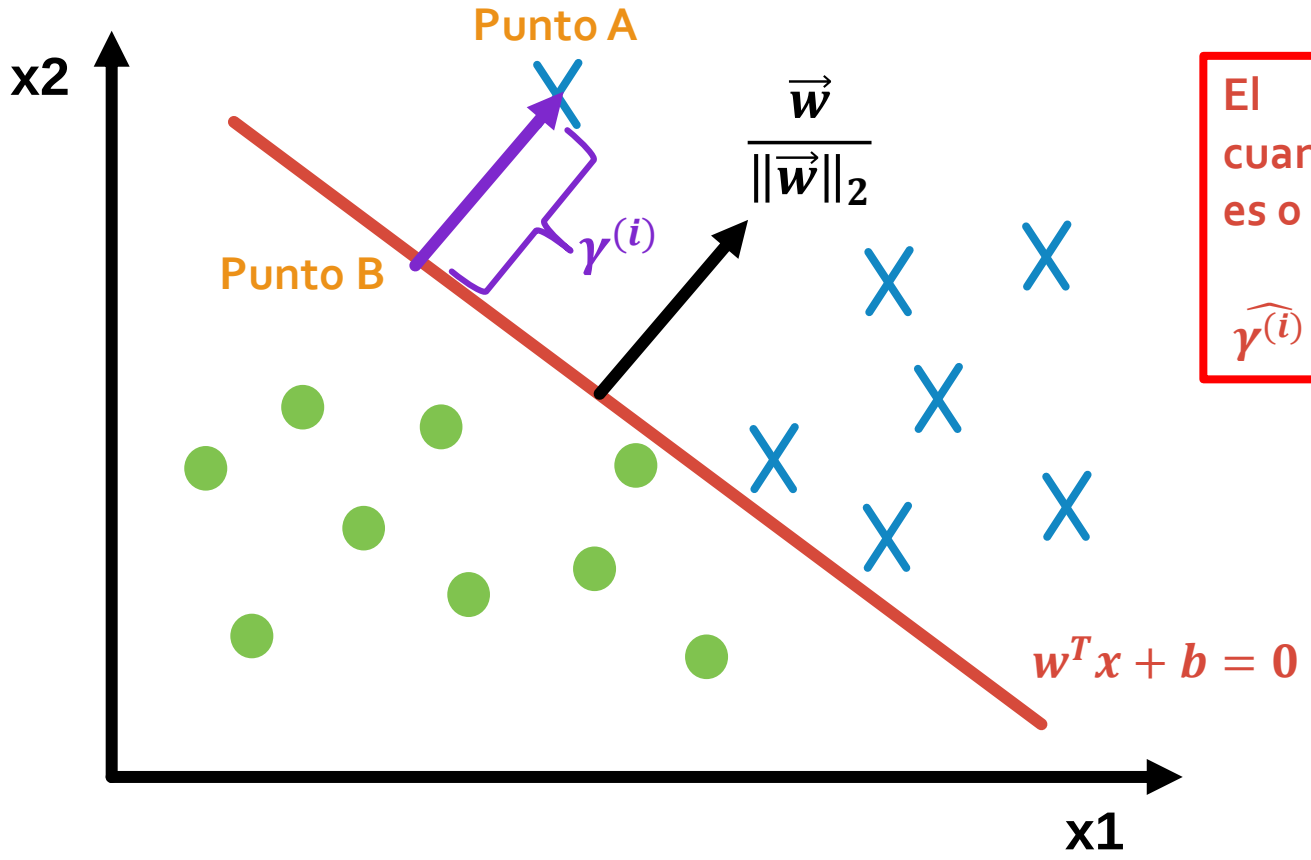
En consecuencia el **margen funcional** se **define** como:

El mínimo de todos los márgenes funcionales contemplando todos los datos de entrenamiento $\left(x^{(i)}, y^{(i)}\right)$

MÁQUINA DE SV

MARGEN GEOMÉTRICO

Para definir el **margen geométrico** $\gamma^{(i)}$ se **visualiza** lo **siguiente:**



El hiperplano se forma cuando el margen funcional es o

$$\widehat{\gamma^{(i)}} = y^{(i)}\left(w^T x^{(i)} + b\right) = 0$$

MÁQUINA DE SV

MARGEN GEOMÉTRICO

El vector $\boldsymbol{w}$ es **perpendicular** al **hiperplano** formado por $\boldsymbol{w^T x + b = 0}$ por la **siguiente razón:**

Sí $\boldsymbol{b = 0}$ entonces **se tiene que:**

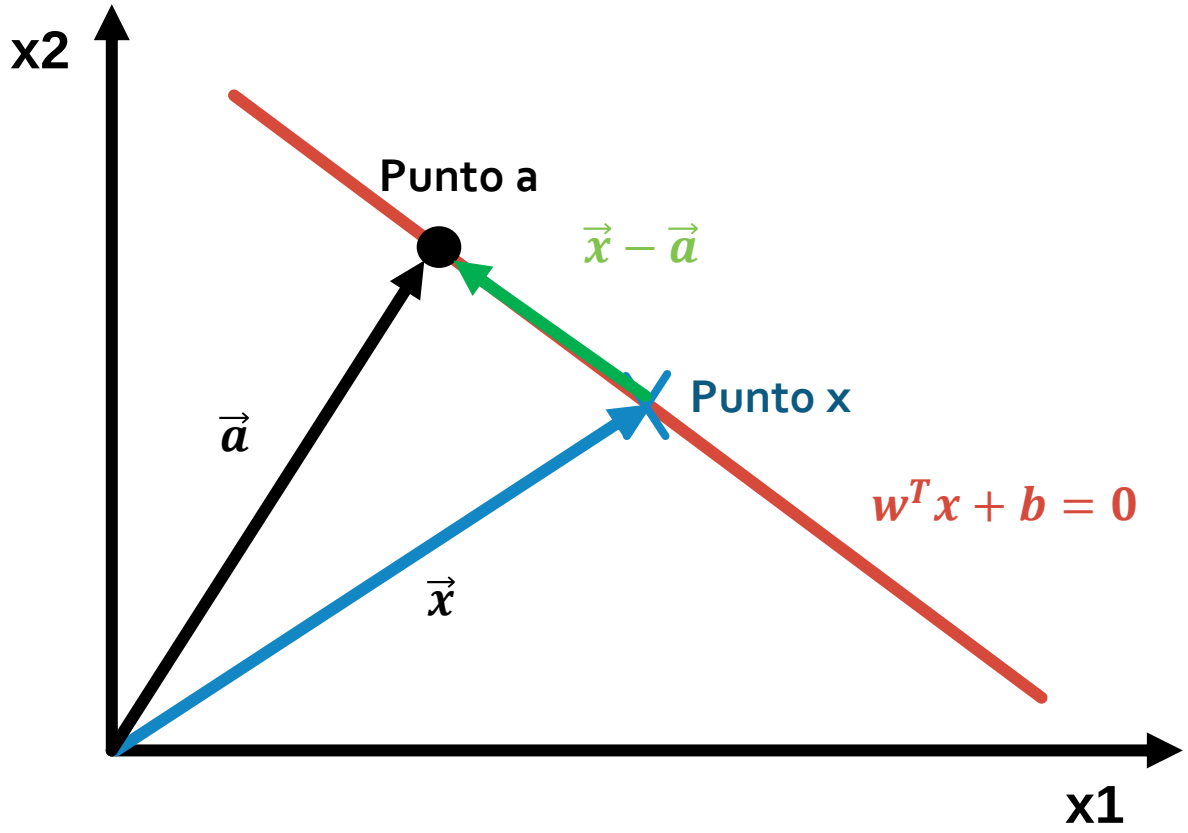
$$\boldsymbol{w^T x + b = 0}$$
$$\boldsymbol{w^T x = 0}$$

Por lo tanto para que $\boldsymbol{w^T x = 0}$ **sea posible, ambos vectores** deben ser **perpendiculares**

MÁQUINA DE SV

MARGEN GEOMÉTRICO

Para ver el caso de $b \neq 0$, se **toma** un **punto** a en **el plano, definido por el vector** $\vec{a}$:



MÁQUINA DE SV

MARGEN GEOMÉTRICO

Por lo tanto, la **ecuación** del **plano también** se puede **expresar** como:

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

$$\boldsymbol{w^T(\vec{x} - \vec{a}) = 0}$$

Desarrollando se tiene**:**

$$\boldsymbol{w^T\vec{x} - w^T\vec{a} = 0}$$

La **única manera** de que **esto** se **cumpla** es:

$$\boldsymbol{w, \vec{x}\ \text{sean}\ \perp}$$
$$\boldsymbol{w, \vec{a}\ \text{sean}\ \perp}$$

MÁQUINA DE SV

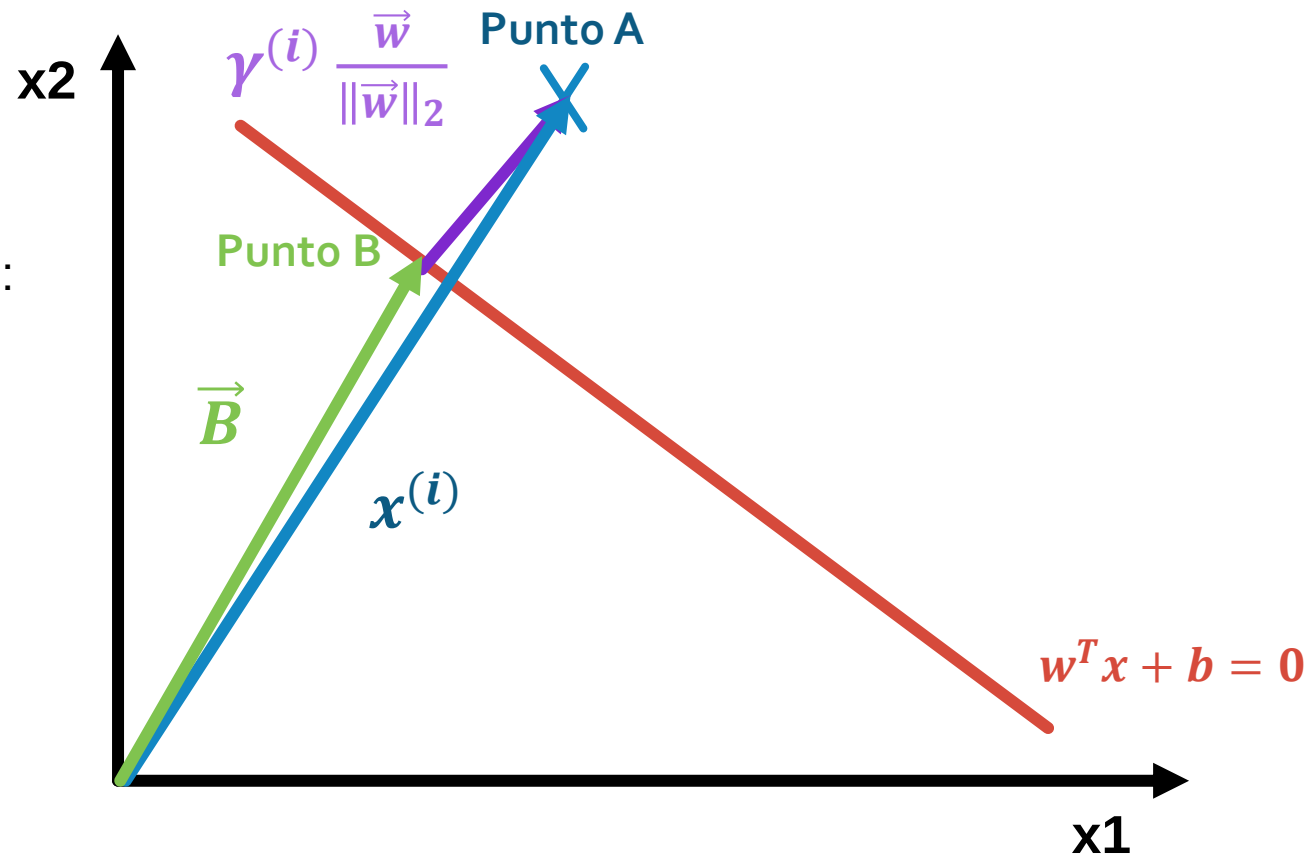
MARGEN GEOMÉTRICO

Una vez demostrado que $\boldsymbol{w}, \vec{x}$ **son** $\perp$**,** se **define** lo siguiente para **obtener el margen geométrico:**

$$\overrightarrow{\boldsymbol{B}} + \boldsymbol{\gamma}^{(i)} \frac{\vec{w}}{\|\vec{w}\|_2} = \boldsymbol{x}^{(i)}$$

Se obtiene el **punto B** en el **hiperplano**:

$$\overrightarrow{\boldsymbol{B}} = \boldsymbol{x}^{(i)} - \boldsymbol{\gamma}^{(i)} \frac{\vec{w}}{\|\vec{w}\|_2}$$



MÁQUINA DE SV

MARGEN GEOMÉTRICO

Como el punto $\vec{B}$ **se encuentra en el hiperplano, se debe satisfacer la ecuación del hiperplano:**

$$w^T \vec{B} + b = 0$$

$$w^T \left(x^{(i)} - \gamma^{(i)} \frac{\vec{w}}{\|\vec{w}\|_2} \right) + b = 0$$

Se **calcula** el **margen geométrico**:

$$\gamma^{(i)} = \frac{w^T x^{(i)} + b}{\|\vec{w}\|_2}$$

MÁQUINA DE SV

MARGEN GEOMÉTRICO

Para tener en **cuenta** el **signo**, de manera general, el **margen geométrico** estaría **dado por**:

$$\gamma^{(i)} = y^{(i)} \left[\frac{w^T x^{(i)} + b}{\|\vec{w}\|_2} \right]$$

Nos damos cuenta que el **margen geométrico** es **invariante** a la **escala** de w por la **normalización**.

De igual forma**, contemplando todos** los **datos** de **entrenamiento,** se define el **peor escenario:**

$$\gamma = min_{i=1,2,\ldots,m} \gamma^{(i)}$$

MÁQUINA DE SV

MARGEN GEOMÉTRICO Y FUNCIONAL

Relacionando **ambos márgenes**, **geométrico** y **funcional**, se tiene que:

$$\gamma^{(i)} = \frac{\widehat{\gamma^{(i)}}}{\|\vec{w}\|_2}$$

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Una vez **definidos ambos márgenes**, se puede expresar el **objetivo** de **clasificación** como un **problema** de **maximización** del **margen geométrico**.

$$\boldsymbol{max_{\gamma,w,b}\gamma}$$

Sujeto a:

$$\boldsymbol{y^{(i)}\left(w^T x^{(i)} + b\right) \geq \gamma; i = 1, \dots, m}$$

$$\boldsymbol{\|w\| = 1}$$

Con $\|\boldsymbol{w}\| = \boldsymbol{1}$ garantizamos que el **margen geométrico** es **igual** al **margen funcional**.

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Debido a que la **restricción** $\|\boldsymbol{w}\| = \mathbf{1}$ define un **problema** de **optimización NO convexo** (**encontrar pesos** $\boldsymbol{w}$ en una **esfera**) se transforma el problema para **eliminar** la **restricción**:

$$\boldsymbol{max}_{\boldsymbol{\gamma},\boldsymbol{w},\boldsymbol{b}} \frac{\widehat{\boldsymbol{\gamma}}}{\|\boldsymbol{w}\|}$$

Sujeto a:

$$\boldsymbol{y}^{(\boldsymbol{i})}\left(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}^{(\boldsymbol{i})} + \boldsymbol{b}\right) \geq \widehat{\boldsymbol{\gamma}}; \boldsymbol{i} = \mathbf{1}, \ldots, \boldsymbol{m}$$

De todas formas se **sigue** teniendo un **problema NO convexo,** pero ahora éste se **encuentra** en la **función objetivo.**

$$\frac{\widehat{\boldsymbol{\gamma}}}{\|\boldsymbol{w}\|}$$

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Si ahora **imponemos** la **restricción de escalamiento**, que determina que el **margen funcional** debe ser $\hat{\gamma} = 1$.

$$max_{\gamma,w,b} \frac{1}{\|w\|}$$

Sujeto a:

$$y^{(i)}\left(w^T x^{(i)} + b\right) \geq \hat{1}; i = 1, \ldots, m$$

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Nos damos cuenta que la **maximización** es lo **mismo** que un **problema** de **minimización**:

$$\boldsymbol{max_{\gamma,w,b}} \frac{\mathbf{1}}{\|\boldsymbol{w}\|} = \boldsymbol{min_{\gamma,w,b}} \frac{\mathbf{1}}{\mathbf{2}} \|\boldsymbol{w}\|^{\mathbf{2}}$$

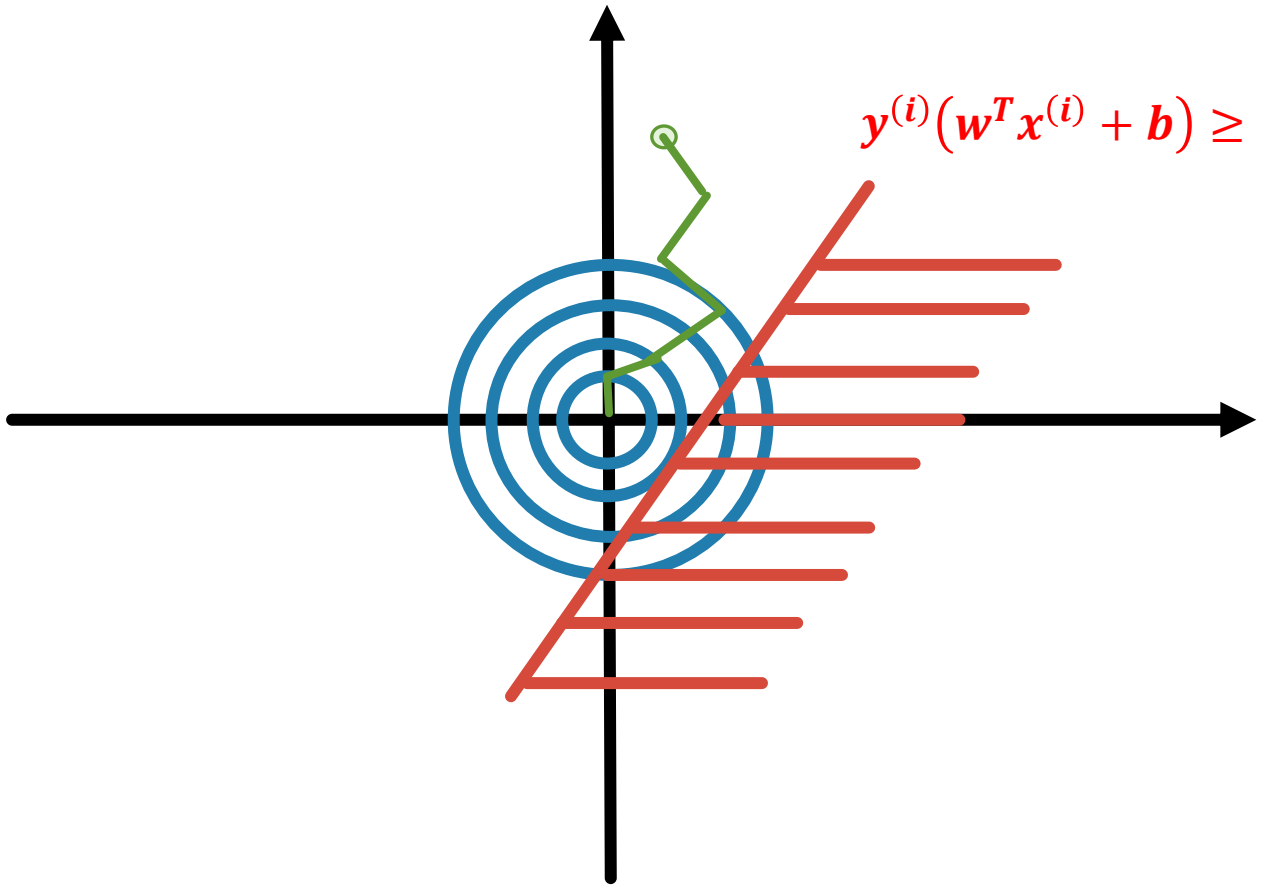
Sujeto a:

$$\boldsymbol{y^{(i)}}\left(\boldsymbol{w^T x^{(i)}} + \boldsymbol{b}\right) \geq \mathbf{1}; \boldsymbol{i} = \mathbf{1}, \ldots, \boldsymbol{m}$$

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Ejemplo gráfico:



MÁQUINA DE SV

MULTIPLICADORES DE LAGRANGE

Ahora veremos como **solucionar problemas** de **optimización sujetos** a **restricciones**. Consideremos el **siguiente problema** de **optimización**:

$$\boldsymbol{min_w\, f(w)}$$

Sujeto a:

$$\boldsymbol{h_i(w) = 0; i = 1, \dots, l}$$

Este **problema** se **puede solucionar** con **multiplicadores** de **Lagrange.** El **lagrangiano** estaría dado por:

$$\boldsymbol{L(w, \beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w)}$$

donde $\boldsymbol{\beta_i}$ son los **multiplicadores** de **Lagrange**.

MÁQUINA DE SV

MULTIPLICADORES DE LAGRANGE

Para **encontrar** $\boldsymbol{w}$ y $\boldsymbol{\beta}$:

$$\frac{\partial L}{\partial w_i} = 0$$

$$\frac{\partial L}{\partial \beta_i} = 0$$

MÁQUINA DE SV

PROBLEMA PRIMAL

Se define un **problema** de **optimización** más **general**, que considere **tanto igualdades** como **desigualdades**: el **Problema Primal**

$$\boldsymbol{min_w\, f(w)}$$

Sujeto a:

$$\boldsymbol{h_i(w) = 0; i = 1, \dots, l}$$

$$\boldsymbol{g_i(w) \leq 0; i = 1, \dots, k}$$

El **lagrangiano** estaría dado por (donde $\boldsymbol{\alpha_i}$ y $\boldsymbol{\beta_i}$ son los **multiplicadores** de **Lagrange**):

$$\boldsymbol{L(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)}$$

MÁQUINA DE SV

PROBLEMA PRIMAL

Se va a **definir** el **problema primal** (donde **P** se refiere a **Primal**)

$$\boldsymbol{\theta_P(w) = max_{\alpha,\beta:\alpha_i \geq 0} L(w,\alpha,\beta)}$$

$$\boldsymbol{\theta_P(w) = max_{\alpha,\beta:\alpha_i \geq 0}\, f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)}$$

Sí $\boldsymbol{w}$ llega a **violar cualquiera** de las **dos restricciones** $\boldsymbol{h_i(w) \neq 0}$ y $\boldsymbol{g_i(w) > 0}$:

$$\boldsymbol{\theta_P(w) = \infty}$$

MÁQUINA DE SV

PROBLEMA PRIMAL

Por el **contrario**, si $\boldsymbol{w}$ satisface **ambas restricciones** $\boldsymbol{h_i(w) = 0}$ y $\boldsymbol{g_i(w) \leq 0}$:

$$\boldsymbol{\theta_P(w) = f(w)}$$

En resumen:

$$\boldsymbol{\theta_P(w) = \begin{cases} f(w) \; si \; se \; satisfacen \\ \infty \; no \; se \; satisfacen \end{cases}}$$

MÁQUINA DE SV

PROBLEMA PRIMAL

Ahora se define la **minimización** de $\boldsymbol{\theta_P(w)}$:

$$\boldsymbol{min_w \theta_P(w) = min_w max_{\alpha,\beta:\alpha_i \geq 0} L(w,\alpha,\beta)}$$

Nos damos cuenta que **volvemos** al **problema inicial** de **minimizar** respecto a $\boldsymbol{w}$.

De manera específica, el valor primal $\boldsymbol{p*}$ es:

$$\boldsymbol{p* = min_w \theta_P(w)}$$

MÁQUINA DE SV

PROBLEMA DUAL

Vamos a ver un **problema semejante**, pero ahora **minimizamos primero respecto a w**:

$$\theta_D(\alpha, \beta) = min_w L(w, \alpha, \beta)$$

El subíndice D indica que estamos definiendo el **problema dual** de esta manera:

$$max_{\alpha,\beta:\alpha_i \geq 0} \theta_D(\alpha, \beta) = max_{\alpha,\beta:\alpha_i \geq 0} min_w L(w, \alpha, \beta)$$

Vemos que **solo** se **cambia** el **orden** de **minimización** y **maximización**. El **valor dual** está **definido** como:

$$d* = max_{\alpha,\beta:\alpha_i \geq 0} \theta_D(\alpha, \beta)$$

MÁQUINA DE SV

PROBLEMA DUAL y PRIMAL

A partir de **pruebas matemáticas** se llega a la **conclusión** de que:

$$\boldsymbol{d*} \leq \boldsymbol{p*}$$

$$\boldsymbol{max\ min} \leq \boldsymbol{min\ max}$$

Para **ciertas condiciones** se cumple **a veces que**:

$$\boldsymbol{d*} = \boldsymbol{p*}$$

MÁQUINA DE SV

CONDICIONES KKT

Para que se cumpla $\boldsymbol{d*} = \boldsymbol{p*}$ se deben cumplir ciertas condiciones. **Antes** de **definirlas**, vamos a **realizar** las **siguientes suposiciones**:

1. Se define que las funciones $\boldsymbol{f}$ y $\boldsymbol{g_i}$ son **convexas** (sí y solo sí la **Hessiana** de la **función** es **positiva semi-definida**).

2. Las funciones $\boldsymbol{h_i}$ son **afines (lineales).**

3. Las funciones $\boldsymbol{g_i}$ son **factibles,** es decir $\exists\, \boldsymbol{w}$: $\boldsymbol{g_i(w) < 0}, \forall\, \boldsymbol{i}$.

Dadas las suposiciones deben existir $\boldsymbol{w*}, \boldsymbol{\alpha*}, \boldsymbol{\beta*}$**,** tal que $\boldsymbol{w*}$ **satisfaga** el **problema primal,** así como $\boldsymbol{\alpha*}, \boldsymbol{\beta*}$ **satisfagan** el **problema dual**.

Es decir **que** se **cumpla**:

$$\boldsymbol{d* = p*}$$

MÁQUINA DE SV

CONDICIONES KKT

Para esto, $\boldsymbol{w}*, \boldsymbol{\alpha}*, \boldsymbol{\beta}*$ deben **satisfacer** las **condiciones Karush-Kuhn-Tucker (KKT):**

$$
\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, n \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \ldots, k \\
g_i(w^*) &\leq 0, \quad i = 1, \ldots, k \\
\alpha^* &\geq 0, \quad i = 1, \ldots, k
\end{aligned}
$$

MÁQUINA DE SV

CONDICIONES KKT

Sí existen parámetros $\boldsymbol{w*, \alpha*, \beta*}$ que **satisfagan** las **condiciones Karush-Kuhn-Tucker (KKT), se garantiza** la **solución** para **ambos** problemas**,** el **primal** y el **dual.**

$$\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \ldots, n$$

$$\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \ldots, l$$

$$\alpha_i^* g_i(w^*) = 0, \quad i = 1, \ldots, k$$

$$g_i(w^*) \leq 0, \quad i = 1, \ldots, k$$

$$\alpha^* \geq 0, \quad i = 1, \ldots, k$$

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Regresando al **problema** del **clasificador marginal**, se sabe que **teníamos** el **siguiente problema** de **optimización** (**primal**):

$$\boldsymbol{max_{w,b}} \frac{\mathbf{1}}{\|\boldsymbol{w}\|} = \boldsymbol{min_{w,b}} \frac{\mathbf{1}}{\mathbf{2}} \|\boldsymbol{w}\|^{\mathbf{2}}$$

Sujeto a:

$$\boldsymbol{y^{(i)}}\left(\boldsymbol{w^T x^{(i)}} + \boldsymbol{b}\right) \geq \mathbf{1}; \boldsymbol{i} = \mathbf{1}, \ldots, \boldsymbol{m}$$

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Transformando la **restricción** en **base** a la **nomenclatura** usada en la **dualidad de Lagrange:**

$$\boldsymbol{max_{w,b}\frac{1}{\|w\|} = min_{w,b}\frac{1}{2}\|w\|^2}$$

Sujeto a:

$$\boldsymbol{g_i(w,b) = -y^{(i)}\left(w^T x^{(i)} + b\right) + 1 \leq 0; i = 1, \ldots, m}$$

Observamos**,** que **solo** existe **una desigualdad,** por lo que **solo** tendremos los **multiplicadores** de **Lagrange** $\boldsymbol{\alpha_i}$.

De igual forma, **no solo** tenemos el **parámetro** $\boldsymbol{w}$**,** pero **también** el **parámetro** $\boldsymbol{b}$.

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Construimos el **Lagrangiano** como**:**

$$L(w, b, \alpha) = f(w) + \sum_{i=1}^{m} \alpha_i g_i(w)$$

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i \left[-y^{(i)}\left(w^T x^{(i)} + b\right) + 1\right]$$

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i \left[-\hat{y}^{(i)} + 1\right]$$

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Debido a la **condición** de **complementariedad (restricción activa)** en **KKT:**

$$\alpha_i^* g_i(w^*) = 0, \quad i = 1, \ldots, k$$

Para los **multiplicadores** de **Lagrange** $\boldsymbol{\alpha_i > 0}$ se tiene que:

$$\boldsymbol{g_i(w) = \left[-\hat{\gamma}^{(i)} + 1\right] = 0}$$

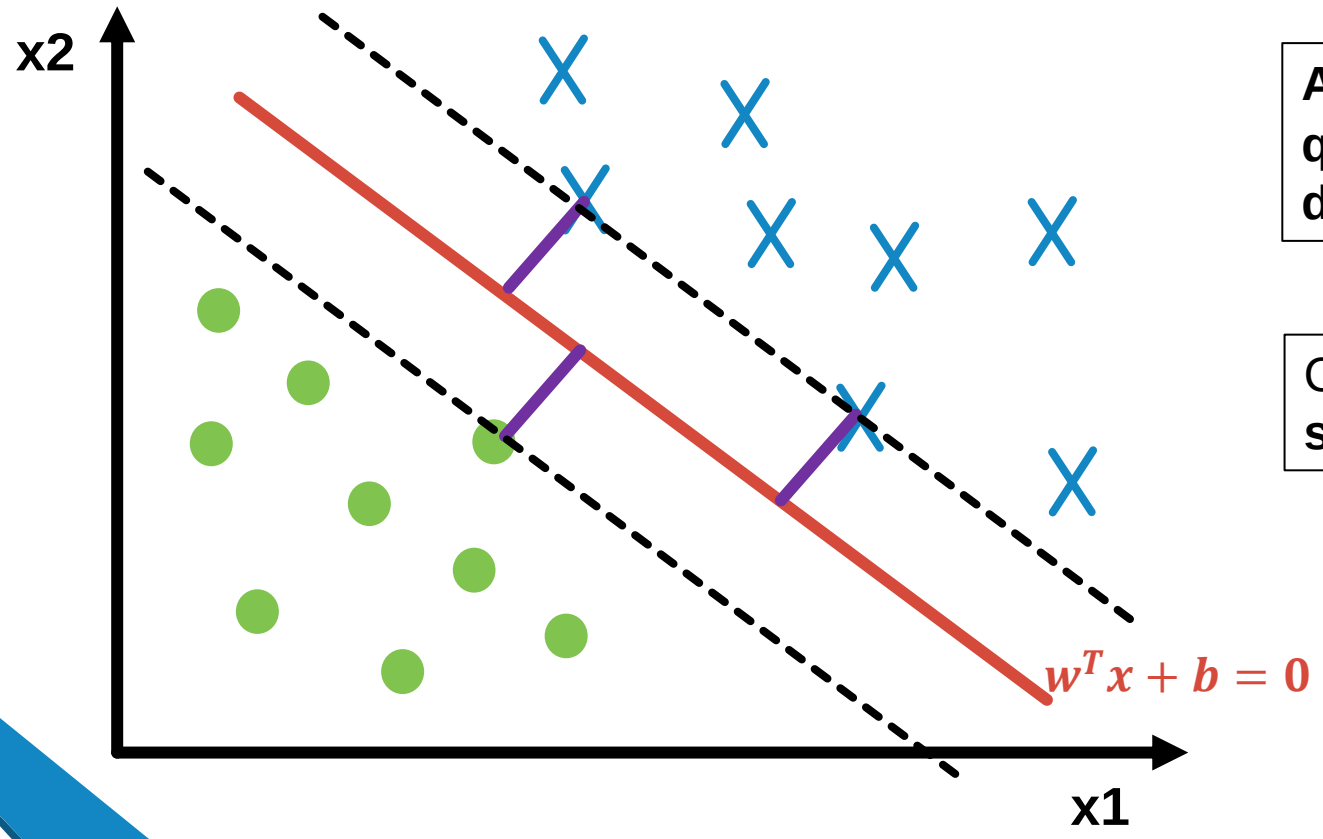
Por lo que el **margen funcional** $\boldsymbol{\gamma^{(i)}}$ tendrá que ser **igual** a **1** para el dato de entrenamiento $\boldsymbol{i}$**:**

$$\boldsymbol{\hat{\gamma}^{(i)} = 1}$$

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Lo que esto significa es que los **datos** de **entrenamiento** con margen $\gamma^{(i)} = 1$ serán los **más cercanos** al **hiperplano** y por lo general sus $\alpha_i \neq 0$ **porque** $g_i = 0$.



A estos datos de entrenamiento, que cumplen con $\gamma^{(i)} = 1$, se les denominarán vectores de soporte.

Como hay **pocos vectores** de **soporte**, la **mayoría** de $\alpha_i = 0$.

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Se **desarrollará** ahora la **versión dual** del **problema**:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)} \left(w^T x^{(i)} + b \right) - 1 \right]$$

Se **define** $\theta_D(\alpha)$**:**

$$\theta_D(\alpha) = min_{w,b} L(w, b, \alpha)$$

Se **minimiza** respecto **a** w**:**

$$\nabla_{w,} L(w, b, \alpha) = w - \sum_{i=1}^{m} \alpha_i \left[y^{(i)} \left(x^{(i)} \right) \right] = 0$$

$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Se **minimiza respecto** a $\boldsymbol{b}$:

$$\frac{\partial L(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\alpha})}{\partial \boldsymbol{b}} = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

Se tienen **ambas restricciones**:

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

$$\frac{\partial L(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\alpha})}{\partial \boldsymbol{b}} = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Se **sustituye** el **mínimo** $\boldsymbol{w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}}$ en el **Lagrangiano:**

$$\boldsymbol{L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)} \left(w^T x^{(i)} + b \right) - 1 \right]}$$

$$\boldsymbol{L(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{m} \alpha_i \left[y^{(i)} \left(w^T x^{(i)} + b \right) - 1 \right]}$$

$$\boldsymbol{L(w, b, \alpha) = \frac{1}{2} \left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \right)^T \sum_{j=1}^{m} \alpha_j y^{(j)} x^{(j)} - \sum_{i=1}^{m} \alpha_i \left[y^{(i)} \left(w^T x^{(i)} + b \right) - 1 \right]}$$

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

$$L(w,b,\alpha) = \frac{1}{2}\left(\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}\right)^T \sum_{j=1}^{m}\alpha_j y^{(j)}x^{(j)} - \sum_{i=1}^{m}\alpha_i\left[y^{(i)}\left(\left(\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}\right)^T x^{(i)} + b\right) - 1\right]$$

$$L(w,b,\alpha) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i y^{(i)}\alpha_j y^{(j)}x^{(i)^T}x^{(j)} - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i y^{(i)}\alpha_j y^{(j)}x^{(i)^T}x^{(j)} - \sum_{i=1}^{m}\alpha_i y^{(i)}b + \sum_{i=1}^{m}\alpha_i$$

$$L(w,b,\alpha) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i y^{(i)}\alpha_j y^{(j)}x^{(i)^T}x^{(j)} - b\sum_{i=1}^{m}\alpha_i y^{(i)} + \sum_{i=1}^{m}\alpha_i$$

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Se **sustituye** ahora la **restricción** de $\boldsymbol{b}$ de la **última ecuación** $\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$:

$$L(w, b, \alpha) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i y^{(i)} \alpha_j y^{(j)} x^{(i)^T} x^{(j)} - b\sum_{i=1}^{m} \alpha_i y^{(i)} + \sum_{i=1}^{m} \alpha_i$$

$$L(w, b, \alpha) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i y^{(i)} \alpha_j y^{(j)} x^{(i)^T} x^{(j)} - 0 + \sum_{i=1}^{m} \alpha_i$$

$$min_{w,b} L(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i y^{(i)} \alpha_j y^{(j)} x^{(i)^T} x^{(j)}$$

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Recordamos la **forma** del **problema dual,** para el **caso** del **clasificador marginal:**

$$max_{\alpha} W(\alpha) = max_{\alpha} min_{w,b} L(w, b, \alpha)$$

$$max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j x^{(i)^T} x^{(j)}$$

Sujeto a:

$$\alpha_i \geq 0, i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i \, y^{(i)} = 0$$

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Podemos **observar** que el **problema** dual **satisface** las **condiciones KKT**.

$$\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \ldots, n$$

$$\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \ldots, l$$

$$\alpha_i^* g_i(w^*) = 0, \quad i = 1, \ldots, k$$

$$g_i(w^*) \leq 0, \quad i = 1, \ldots, k$$

$$\alpha^* \geq 0, \quad i = 1, \ldots, k$$

SE SATISFACE $d* = p*$

$max\ min = min\ max$

PODEMOS RESOLVER EL PROBLEMA PRIMAL RESOLVIENDO EL DUAL

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Recordemos que **ya** tenemos la **expresión** de $\boldsymbol{w}$**,** pero está en **función** de $\boldsymbol{\alpha_i}$:

$$\boldsymbol{w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}}$$

Cómo se **satisface** $\boldsymbol{d* = p*}$ podemos **resolver** el **problema dual** para **obtener** $\boldsymbol{\alpha_i}$:

$$\boldsymbol{max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j {x^{(i)}}^T x^{(j)}}$$

Finalmente, con las $\boldsymbol{\alpha_i}$ podemos obtener $\boldsymbol{b}$ **resolviendo** el **problema primal** con la $\boldsymbol{w}$ obtenida.

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Para obtener $\boldsymbol{b}$ sabemos que la **ecuación** del **plano** es $\boldsymbol{w^T x^{(i)} + b = 0}$, donde:

$$\boldsymbol{w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}}$$

Despejamos $\boldsymbol{b}$:

$$\boldsymbol{b = -w^T x^{(i)}}$$

Como $\boldsymbol{\alpha_i = 0}$ para **vectores** que **no son** de **soporte,** esto **solo aplica** para **vectores** de **soporte** (en este caso, **vectores** que están **más cerca** del **hiperplano**):

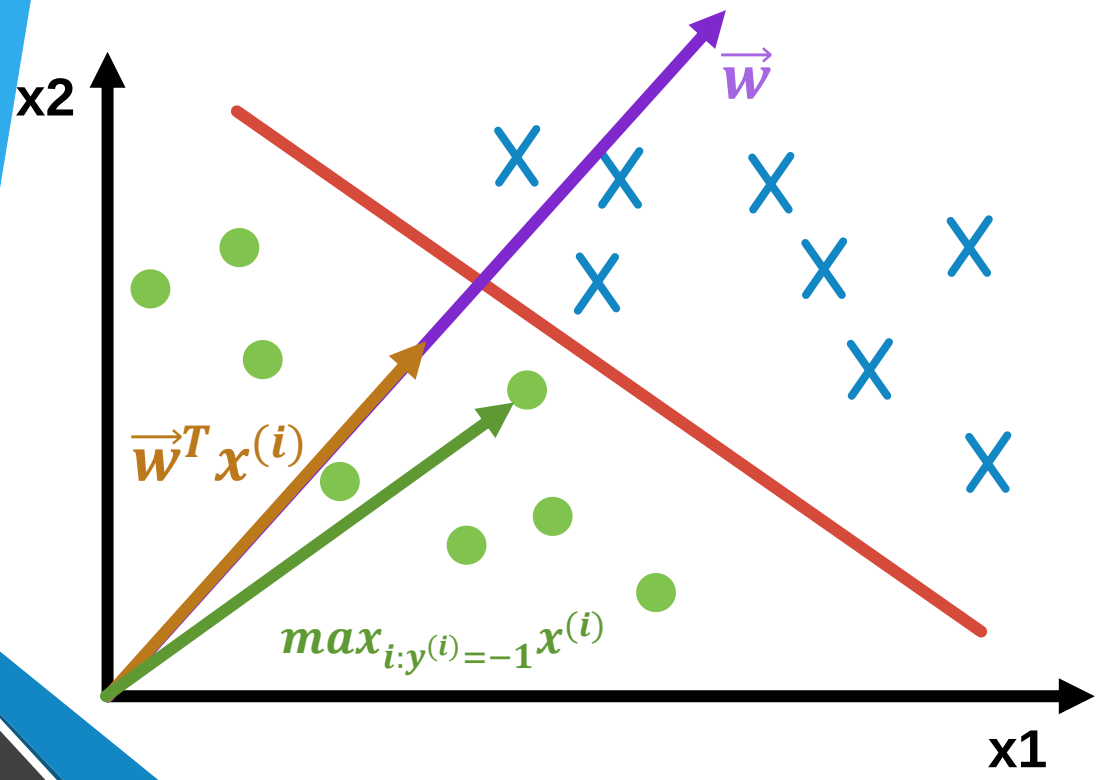
$$\boldsymbol{max_{i:y^{(i)}=-1} - w^T x^{(i)}}$$

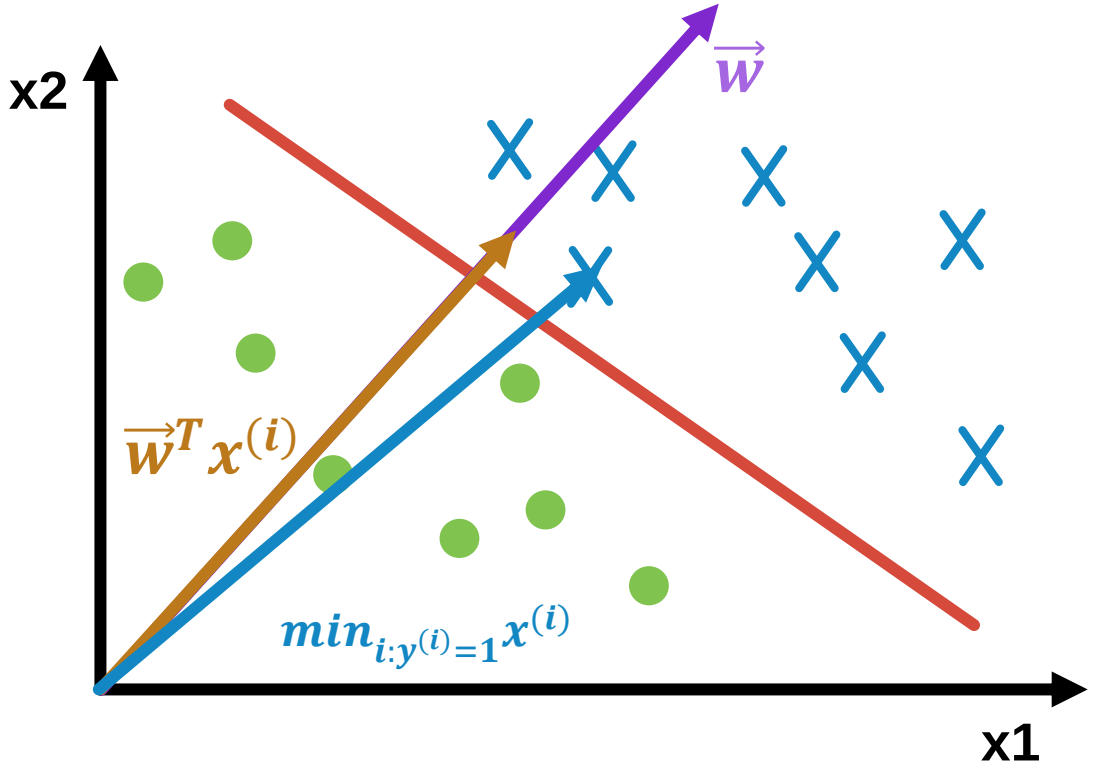
$$\boldsymbol{min_{i:y^{(i)}=1} w^T x^{(i)}}$$

MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Se representan ambos vectores $max_{i:y^{(i)}=-1} - w^T x^{(i)}$ y $min_{i:y^{(i)}=1} w^T x^{(i)}$:



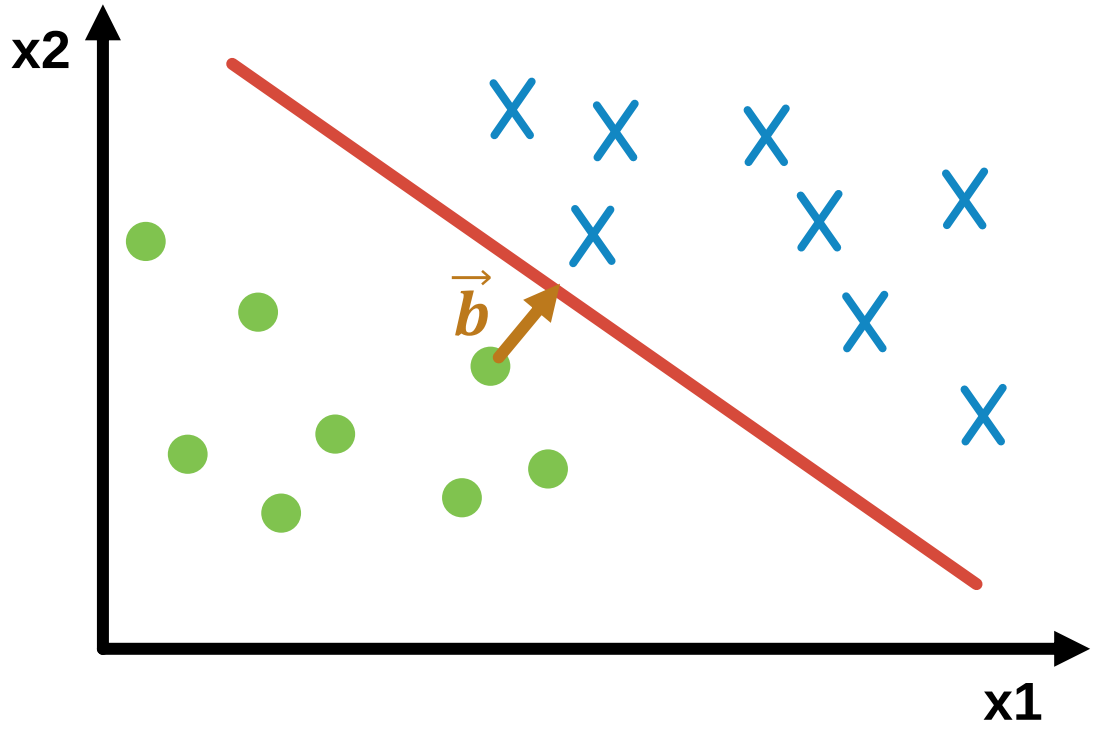


MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Por lo tanto ***b*** **sería**:

$$b^* = -\frac{\max_{i:y^{(i)}=-1} w^{*T} x^{(i)} + \min_{i:y^{(i)}=1} w^{*T} x^{(i)}}{2}.$$



MÁQUINA DE SV

CLASIFICADOR MARGINAL ÓPTIMO

Cómo ya **calculamos** todos los **parámetros**, vamos a **expresar** las **predicciones** en función de **productos puntos**:

$$h_{w,b}(x) = g\left(w^T x + b\right) = g\left(\left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}\right)^T x + b\right)$$

$$h_{w,b}(x) = g\left(\sum_{i=1}^{m} \alpha_i y^{(i)} \left\langle x^{(i)}, x \right\rangle + b\right)$$

Como $\alpha_i = 0$ para **vectores** que **no son** de **soporte.** Es decir, para **hacer** una **nueva predicción** se **calcularían** las **proyecciones solamente** con los **vectores de soporte.**

MÁQUINA DE SV

KERNELS

Para **comenzar** vamos a **realizar** un **mapeado** de las **características** $x^{(i)}$ a un **espacio dimensional más alto**. Vamos a ver el **caso** con **una** sola **característica** $x^{(i)} \in \mathbb{R}$:

$$\boldsymbol{\phi}\left(x^{(i)}\right) = \begin{bmatrix} x \\ x^2 \\ x^3 \end{bmatrix}$$

De esta manera, podemos **cambiar todo** lo que hemos hecho en el **algoritmo** de la siguiente **forma:**

$$\left\langle x^{(i)}, x^{(j)} \right\rangle \rightarrow \left\langle \boldsymbol{\phi}\left(x^{(i)}\right), \boldsymbol{\phi}\left(x^{(j)}\right) \right\rangle = \boldsymbol{\phi}\left(x^{(i)}\right)^T \boldsymbol{\phi}\left(x^{(j)}\right)$$

MÁQUINA DE SV

KERNELS

Al **producto punto** de las **características** en el **espacio altamente dimensional** se le llama **Kernel**:

$$K(x, z) = \phi(x)^T \phi(z)$$

Por lo tanto**,** el **algoritmo** va a **calcular** las **características** en **este espacio** de **mayores dimensiones.**

En principio, si ϕ **realiza** un mapeo a un **espacio altamente dimensional** pareciera que **calcular** $K(x, z)$ sería **muy costoso** computacionalmente. De hecho, sería **imposible** si se **llegara** a realizar un **mapeo** en **infinitas dimensiones**.

MÁQUINA DE SV

KERNELS

Aún así, existen **casos especiales** donde el **costo computacional no** se **dispara**. Son estos **casos** los que nos interesan para **poder incluirlos** en la **máquina** de **vectores** de **soporte**.

EJEMPLO:

Se tienen **dos vectores** $\boldsymbol{x, z} \in \mathbb{R}^{\boldsymbol{n}}$ por lo tanto, el **kernel** estaría dado por:

$$\boldsymbol{K(x,z) = \left(x^T z\right)^2}$$

$$\boldsymbol{K(x,z) = \left(\sum_{i=1}^{n} x_i z_i\right)\left(\sum_{j=1}^{n} x_j z_j\right)}$$

MÁQUINA DE SV

KERNELS

EJEMPLO:

$$K(x,z) = \left(\sum_{i=1}^{n} x_i z_i\right)\left(\sum_{j=1}^{n} x_j z_j\right)$$

$$K(x,z) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j z_i z_j$$

$$K(x,z) = \sum_{i,j=1}^{n} x_i x_j z_i z_j$$

MÁQUINA DE SV

KERNELS

EJEMPLO:

$$K(x, z) = \sum_{i,j=1}^{n} x_i x_j z_i z_j$$

Se **observa** que se este **kernel corresponde** a la siguiente **transformación ϕ:**

$$\phi = \begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ \vdots \\ x_1 x_n \\ \vdots \\ x_n x_n \end{bmatrix} \in \mathbb{R}^{n^2}$$

MÁQUINA DE SV

KERNELS

Es importante ver que, aunque calcular $\boldsymbol{\phi}$ necesita $\boldsymbol{n^2}$ **operaciones,** para el caso del **kernel solo** se necesitan $\boldsymbol{2n+1}$ **operaciones**

$$\boldsymbol{\phi} = \begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ \vdots \\ x_1 x_n \\ \vdots \\ x_n x_n \end{bmatrix} \in \mathbb{R}^{n^2}$$

$$K(x, z) = \sum_{i,j=1}^{n} x_i x_j z_i z_j = \left(x^T z\right)^2$$

MÁQUINA DE SV
KERNELS

EJEMPLO 2:

De manera general, se puede obtener un **kernel** para poder **mapear polinomios** de **cualquier orden** $\boldsymbol{d}$:

$$\boldsymbol{K(x,z) = \left(x^T z + c\right)^d}$$

Veamos el **ejemplo** con $\boldsymbol{d = 2}$ y $\boldsymbol{n = 3}$:

$$\boldsymbol{K(x,z) = \left(x^T z + c\right)^2}$$

$$\boldsymbol{K(x,z) = \left(x^T z\right)^2 + 2cx^T z + c^2}$$

MÁQUINA DE SV

KERNELS

EJEMPLO 2:

$$K(x,z) = \left(x^T z\right)^2 + 2cx^T z + c^2$$

$$K(x,z) = \sum_{i,j=1}^{n} x_i x_j z_i z_j + \sum_{i=1}^{n} 2cx_i z_i + c^2$$

$$K(x,z) = \sum_{i,j=1}^{n} x_i x_j z_i z_j + \sum_{i=1}^{n} 2cx_i z_i + c^2$$

$$K(x,z) = \sum_{i,j=1}^{n} x_i x_j z_i z_j + \sum_{i=1}^{n} \left(\sqrt{2c}x_i\right)\left(\sqrt{2c}z_i\right) + c^2$$

MÁQUINA DE SV

KERNELS

EJEMPLO 2:

$$\boldsymbol{K(x,z) = \sum_{i,j=1}^{n} x_i x_j z_i z_j + \sum_{i=1}^{n} (\sqrt{2c}x_i)(\sqrt{2c}z_i) + c^2}$$

Observamos la **transformación $\boldsymbol{\phi}$.**

$$\phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \\ \sqrt{2c}x_1 \\ \sqrt{2c}x_2 \\ \sqrt{2c}x_3 \\ c \end{bmatrix}$$

MÁQUINA DE SV
KERNELS

Recordando que para el **caso general**:

$$\boldsymbol{K(x,z) = \left(x^T z + c\right)^d}$$

La **matriz** $\boldsymbol{\phi}$ requeriría de $\boldsymbol{O\left(n^d\right)}$ **cálculos** para computarse**, pero el kernel solo necesita** $\boldsymbol{O(n)}$ **cálculos.**

Por lo tanto**,** se ha **encontrado** una **forma** de trabajar con **vectores** en un **espacio** de **infinitas dimensiones sin** tener que **representarlos explícitamente.**

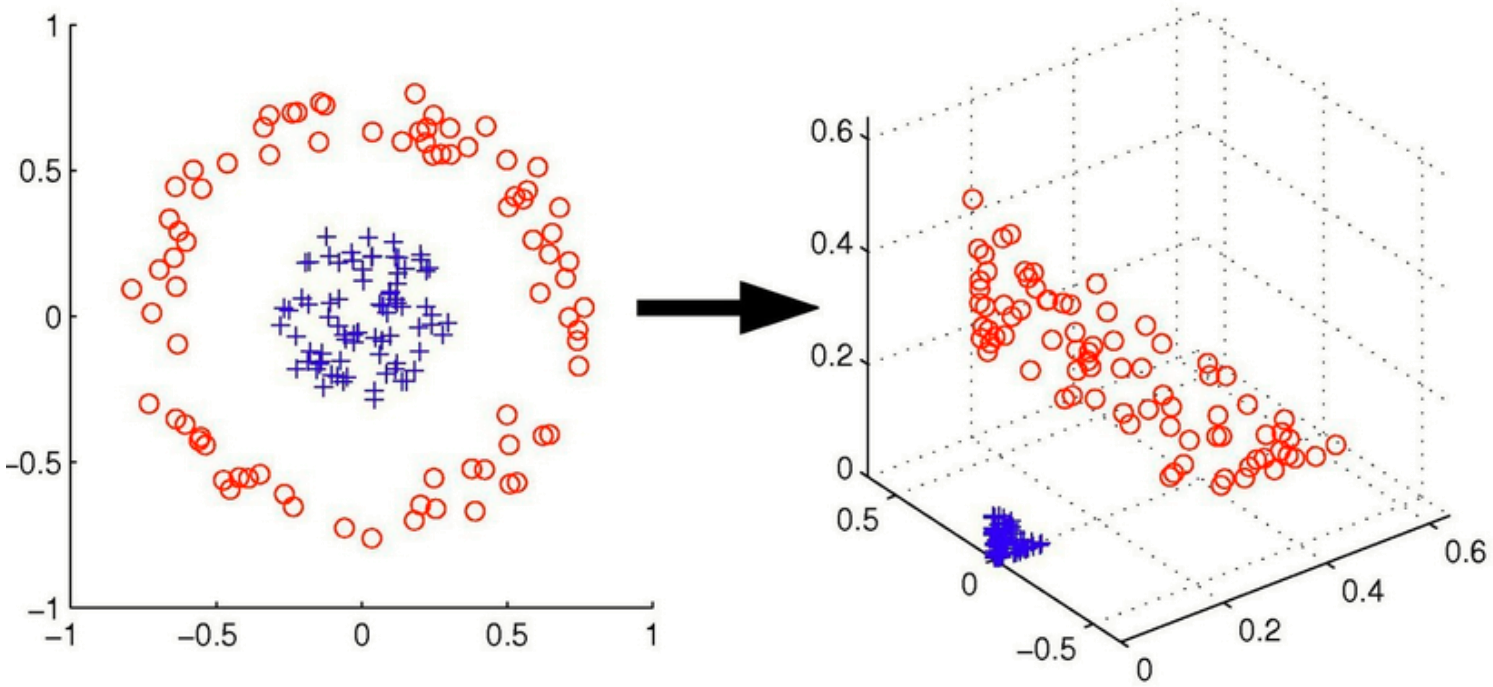
MÁQUINA DE SV
KERNELS

¿PORQUÉ NECESITAMOS EL KERNEL?

¿PORQUÉ TRANSFORMAR LOS VECTORES?

MÁQUINA DE SV

KERNELS



$$\boldsymbol{\phi} = \begin{bmatrix} \boldsymbol{x_1} \\ \boldsymbol{x_2} \end{bmatrix} \in \mathbb{R}^2$$

$$\boldsymbol{\phi} = \begin{bmatrix} \boldsymbol{x_1 x_1} \\ \boldsymbol{\sqrt{2} x_1 x_2} \\ \boldsymbol{x_2 x_2} \end{bmatrix} \in \mathbb{R}^3$$

MÁQUINA DE SV

KERNELS

Ahora veamos un **kernel muy importante:** el **kernel Gaussiano** o **kernel** de **función de base radial:**

$$K(x,z) = \phi(x)^T \phi(z) = e^{-\left(\frac{\|x-z\|^2}{2\sigma^2}\right)}$$

Se puede **entender**, de **manera intuitiva**, que:

- Sí x y z se encuentran **muy cercanos**, el **kernel** $K(x,z)$ será **1**.
- Sí x y z se encuentran **muy lejanos**, el **kernel** $K(x,z)$ será **0**.

Este **kernel corresponde** a un **mapeado** $\phi(x) \in \mathbb{R}^{\infty}$

MÁQUINA DE SV

KERNELS

Vamos a **demostrar** que $\boldsymbol{K(x,z) = \phi(x)^T \phi(z) = e^{-\left(\frac{\|x-z\|^2}{2\sigma^2}\right)}}$ nos da un **mapeado** $\boldsymbol{\phi(x) \in \mathbb{R}^\infty}$.

Veremos el **ejemplo** más sencillo $\boldsymbol{x, z \in \mathbb{R}}$ y $\boldsymbol{\sigma^2 = \frac{1}{2}}$.

$$\boldsymbol{K(x,z) = \phi(x)^T \phi(z) = e^{-\left(\frac{(x-z)^2}{2\sigma^2}\right)} = e^{-(x-z)^2}}$$

$$\boldsymbol{K(x,z) = e^{-x^2+2xz-z^2} = e^{-x^2}\, e^{-z^2}\, e^{2xz}}$$

Recordando que $\boldsymbol{e^k = \sum_{n=0}^{\infty} \frac{k^n}{n!}}$:

$$\boldsymbol{K(x,z) = e^{-x^2+2xz-z^2} = e^{-x^2} e^{-z^2} \sum_{n=0}^{\infty} \frac{(2xz)^n}{n!}}$$

MÁQUINA DE SV

KERNELS

$$K(x,z) = e^{-x^2} e^{-z^2} \sum_{n=0}^{\infty} \frac{2^n x^n z^n}{n!}$$

$$K(x,z) = \sum_{n=0}^{\infty} \frac{2^n \left[e^{-x^2} x^n\right]\left[e^{-z^2} z^n\right]}{n!} = \phi(x)^T \phi(z)$$

Por lo tanto:

$$\phi(x) = \begin{bmatrix} e^{-x^2} \\ \frac{\sqrt{2^1}}{1!} e^{-x^2} x \\ \frac{\sqrt{2^2}}{2!} e^{-x^2} x^2 \\ \frac{\sqrt{2^3}}{3!} e^{-x^2} x^3 \\ \vdots \end{bmatrix}$$

MÁQUINA DE SV

KERNELS

Hasta el momento se han **planteado kernels**, pero **no** se ha **establecido** cuando un **kernel es válido** o **no**. Es decir, que $\forall \boldsymbol{x}, \boldsymbol{z} \; \exists \; \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{\phi}(\boldsymbol{x})^{\boldsymbol{T}} \boldsymbol{\phi}(\boldsymbol{z})$.

Para esto, se **deben cumplir** las **dos condiciones** de **Mercer**:

1. **Simetría del kernel:**

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{K}(\boldsymbol{z}, \boldsymbol{x})$$

2. La **matriz K** es **positiva semi definida**:

$$\boldsymbol{K} = \begin{bmatrix} \boldsymbol{K}\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(1)}\right) & \boldsymbol{K}\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\right) & \dots & \boldsymbol{K}\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(m)}\right) \\ \boldsymbol{K}\left(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)}\right) & \boldsymbol{K}\left(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(2)}\right) & \dots & \boldsymbol{K}\left(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(m)}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{K}\left(\boldsymbol{x}^{(m)}, \boldsymbol{x}^{(1)}\right) & \boldsymbol{K}\left(\boldsymbol{x}^{(m)}, \boldsymbol{x}^{(2)}\right) & \dots & \boldsymbol{K}\left(\boldsymbol{x}^{(m)}, \boldsymbol{x}^{(m)}\right) \end{bmatrix}$$

MÁQUINA DE SV

KERNELS

Teorema de Mercer:

Establezcamos que $\boldsymbol{K}: \mathbb{R}^{\boldsymbol{n}} \mathbf{x} \mathbb{R}^{\boldsymbol{n}} \rightarrow \mathbb{R}$**.** Entonces, para que $\boldsymbol{K}$ sea un **kernel válido** de Mecer es **necesario y suficiente** que **para cualquier** $\left\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)}\right\}, (\boldsymbol{m} < \infty)$, la correspondiente **matriz** de **kernel** sea simétrica positiva semi definida.

Aplicaciones

Realmente, los **kernels no solo** pueden aplicarse **para SVM**, pero para **cualquier algoritmo** que pueda ser **representado** en **términos** de **productos punto**.

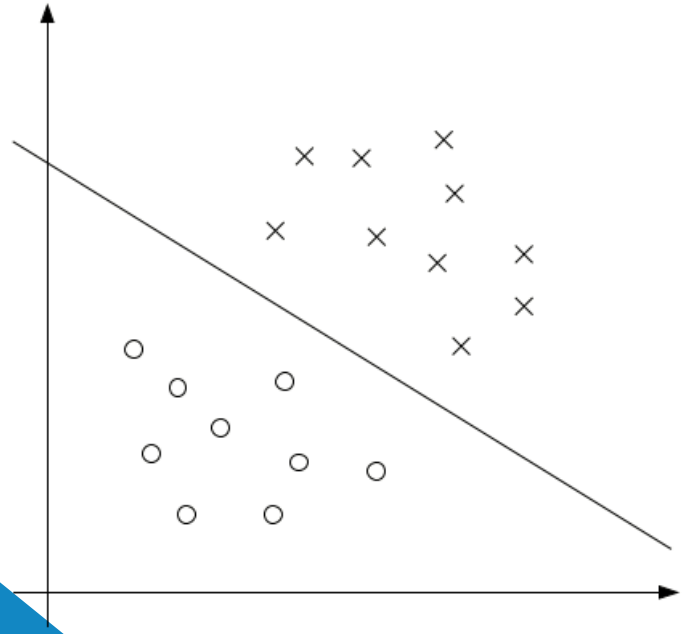
- **Regresión lineal.**
- **Regresión logística.**
- **Perceptrón.**
- **Procesos gaussianos.**
- **PCA.**

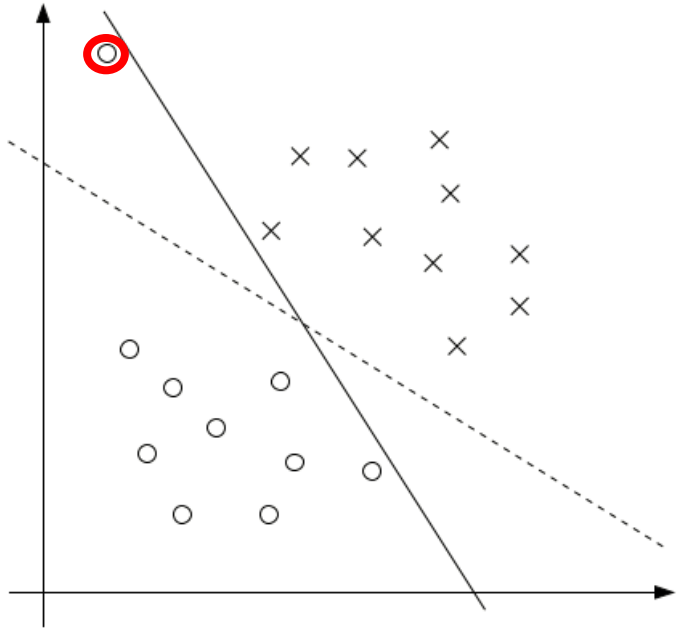
MÁQUINA DE SV

SVM CON REGULARIZACIÓN L1

Aún cuando los **kernels** pueden **trasladar** el **problema** a un **espacio** donde los datos sean **separables linealmente**, **no existe** una **garantía** para que esto **siempre ocurra**.

Además, existe el **problema** de los **datos atípicos** que pueden llegar a **cambiar** el **hiperplano drásticamente** dejando **poco margen**.



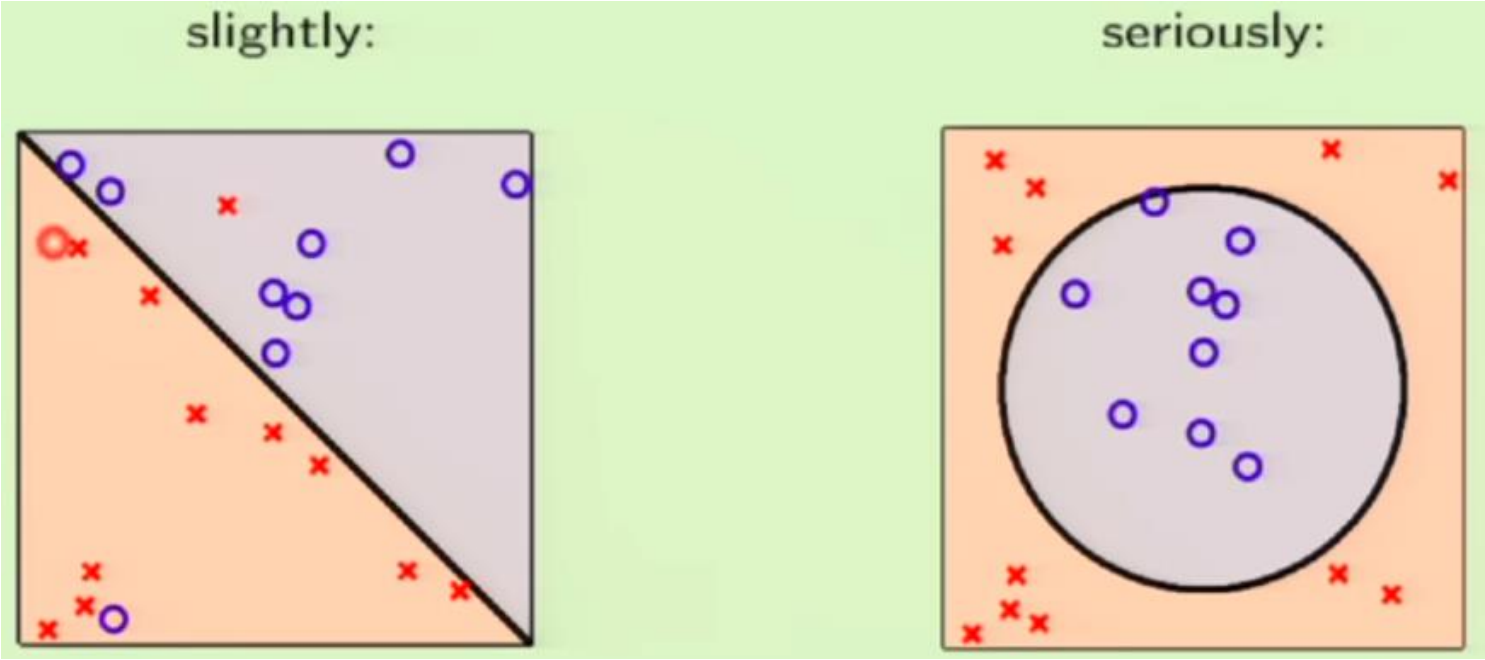


MÁQUINA DE SV

SVM CON REGULARIZACIÓN L1

De **manera general**, puede haber **dos tipos** de **no linealidad**:

- **Poca no linealidad** por datos atípicos (**solución**: **regularización**).
- **No linealidad inherente** a los datos (**solución**: **kernels**).



MÁQUINA DE SV

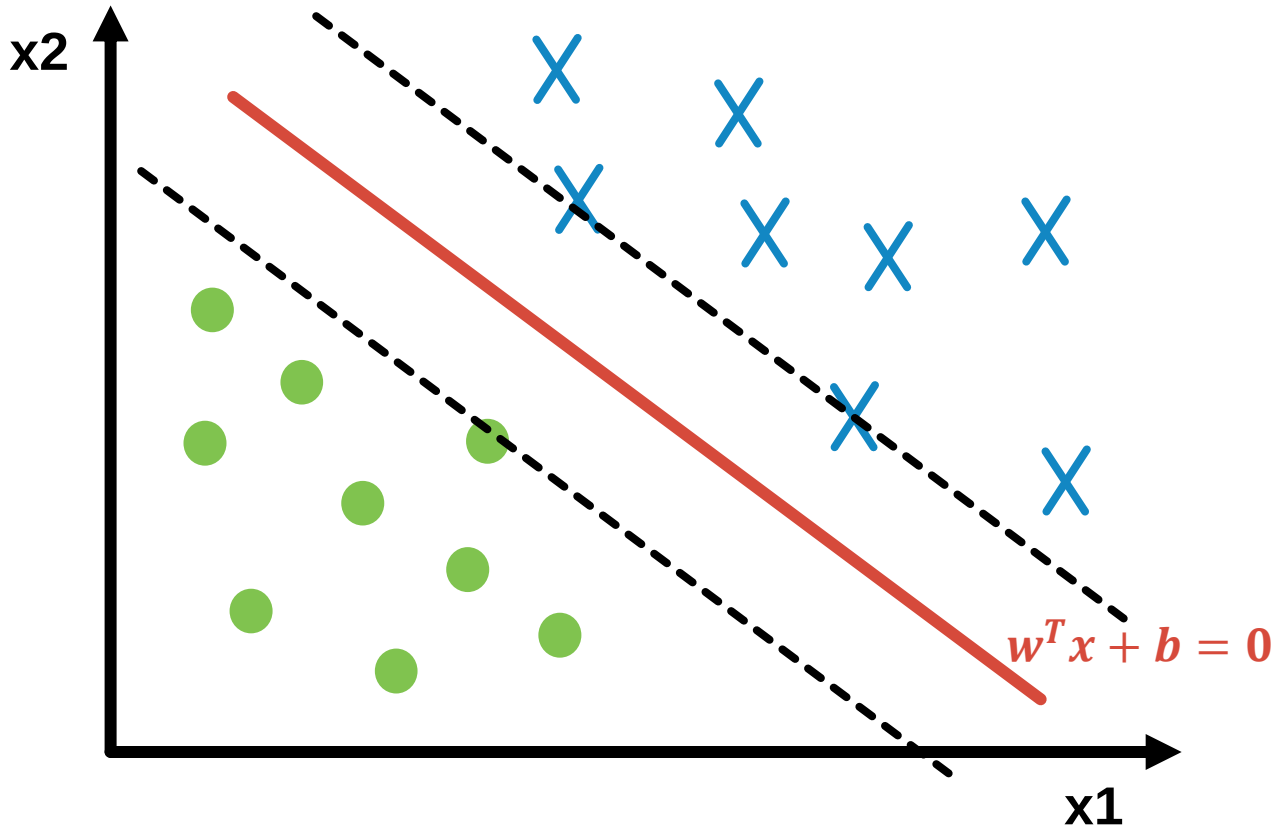
SVM CON REGULARIZACIÓN L1

Vamos a **resolver** el **problema** de los **datos atípicos.** Para esto**, recordemos** cual era el **problema original**:

$$min_{w,b}\frac{1}{2}\|w\|^2$$

Sujeto a:

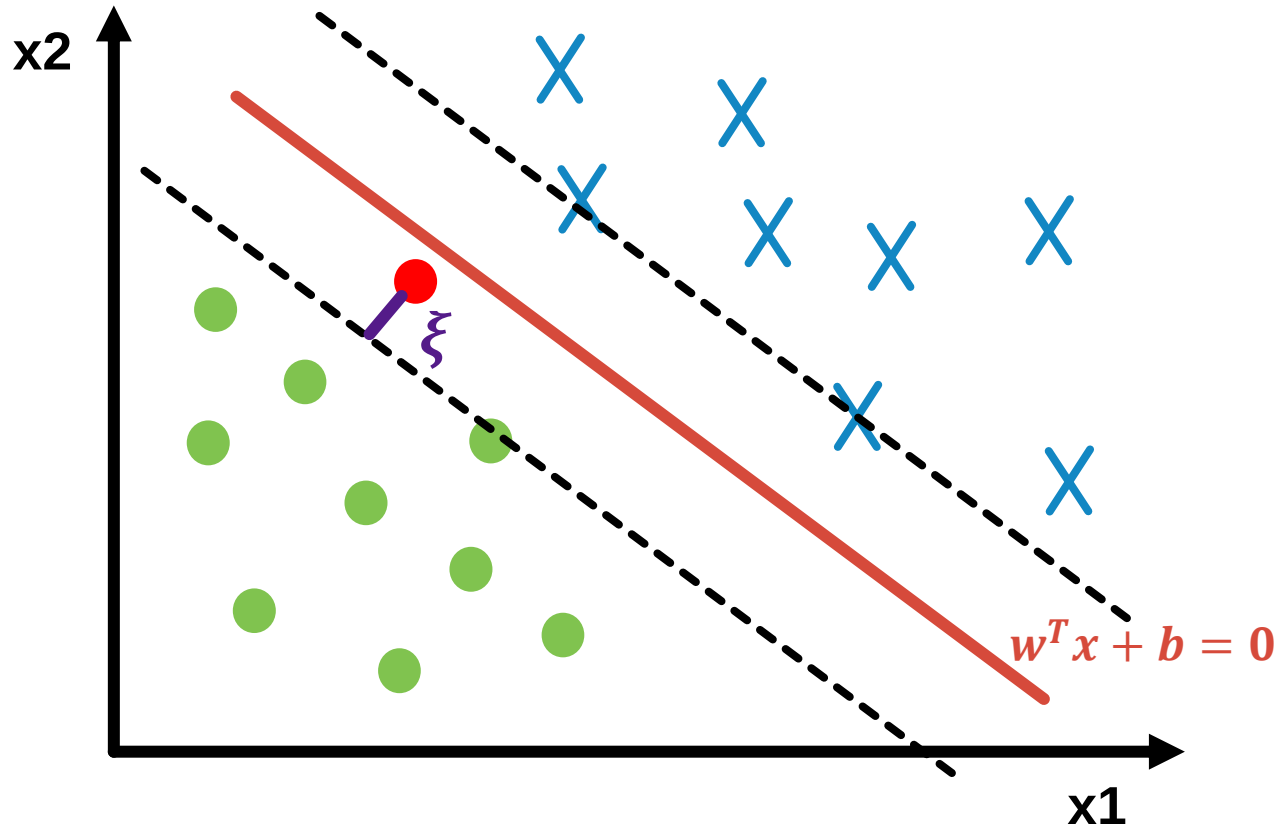
$$y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1; i = 1, \dots, m$$



MÁQUINA DE SV

SVM CON REGULARIZACIÓN L1

Ahora consideraremos **un error** como una **violación del margen** por una **cantidad $\xi \geq 0$:**



MÁQUINA DE SV

SVM CON REGULARIZACIÓN L1

Consideramos el **total** de **violación generada** por **todos** los **datos** como **(regularización L1):**

$$\boldsymbol{error\ total} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{\xi_i}$$

El **problema** se **convierte** en:

$$\boldsymbol{min_{w,b}} \frac{\mathbf{1}}{\mathbf{2}} \|\boldsymbol{w}\|^{\mathbf{2}} + \boldsymbol{C} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{\xi_i}$$

Sujeto a:

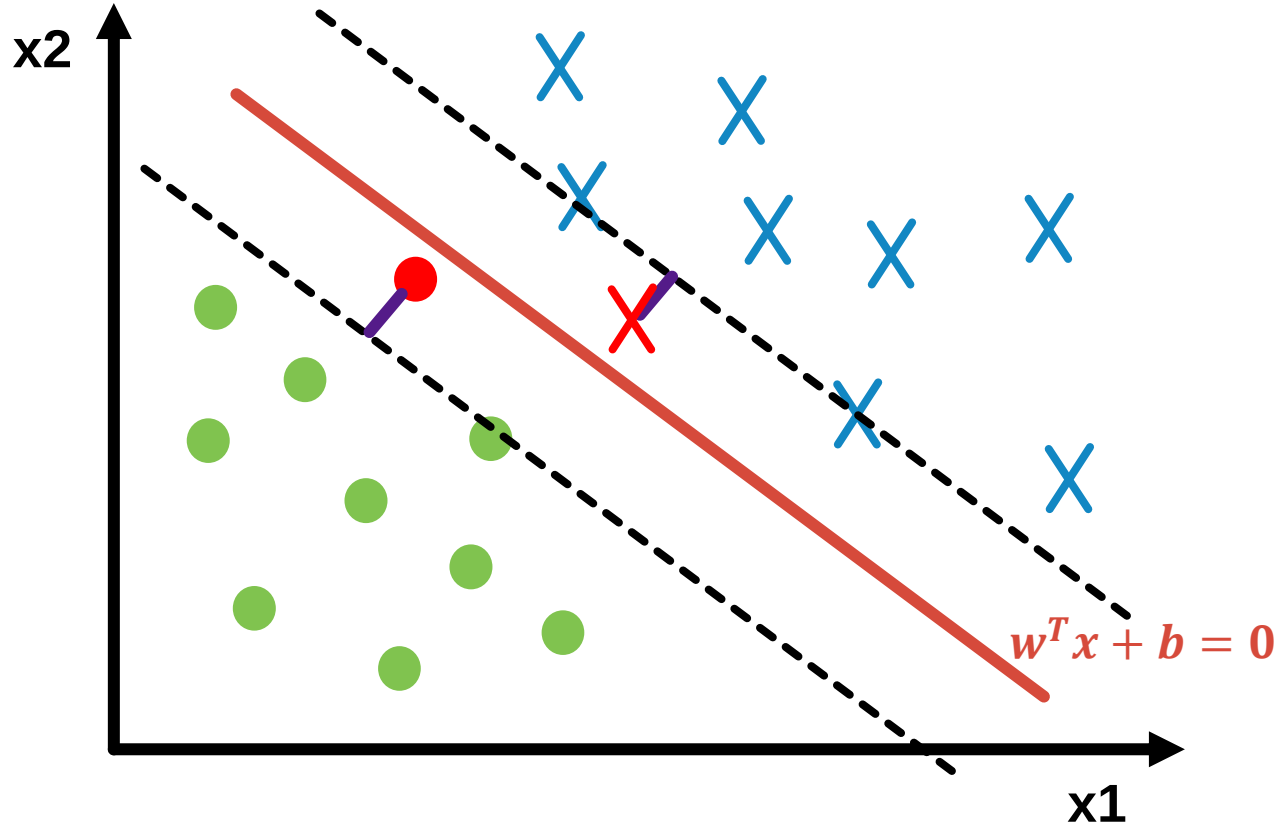
$$\boldsymbol{y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i; i = 1, \ldots, m}$$

$$\boldsymbol{\xi_i > 0}$$

MÁQUINA DE SV

SVM CON REGULARIZACIÓN L1

Es decir, ahora sí **se permiten datos** con **margen funcional < 1** que **violen** el **hiperplano**, solo que **cada vez** que **exista** una **violación** por un dato, **se penalizara** la **función objetivo** por una **cantidad** $C\xi_i$.



MÁQUINA DE SV

SVM CON REGULARIZACIÓN L1

Con esta **nueva función objetivo**, se construye el **Lagrangiano**:

$$L(w, b, \xi, \alpha, r) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i\left[y^{(i)}\left(w^T x^{(i)} + b\right) - 1 + \xi_i\right] - \sum_{i=1}^{m} r_i\xi_i$$

donde los **multiplicadores** de **Lagrange** son α_i y r_i.

MÁQUINA DE SV

SVM CON REGULARIZACIÓN L1

Si se **resuelve** de nuevo el **problema dual, derivando** respecto a $\boldsymbol{w}$, $\boldsymbol{b}$ y $\boldsymbol{\xi_i}$ se obtiene lo siguiente:

$$\frac{\partial L(w, b, \xi, \alpha, r)}{\partial b} = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

$$\nabla_w L(w, b, \xi, \alpha, r) = w - \sum_{i=1}^{m} \alpha_i \left[y^{(i)} \left(x^{(i)} \right) \right] = 0$$

$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

$$\frac{\partial L(w, b, \xi, \alpha, r)}{\partial \xi_i} = C - \alpha_i - r_i = 0$$

MÁQUINA DE SV

SVM CON REGULARIZACIÓN L1

Sustituyendo $\boldsymbol{w}$:

$$\boldsymbol{min\ L(w,b,\xi,\alpha,r)} = \frac{\boldsymbol{1}}{\boldsymbol{2}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{m}}\boldsymbol{\alpha_i y^{(i)}\alpha_j y^{(j)} x^{(i)^T} x^{(j)}} + \boldsymbol{C}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\boldsymbol{\xi_i} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\boldsymbol{r_i\xi_i} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\boldsymbol{\alpha_i\xi_i} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\boldsymbol{\alpha_i}\left[\boldsymbol{y^{(i)}}\left(\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\boldsymbol{\alpha_i y^{(i)} x^{(i)}}\right)^{\boldsymbol{T}}\boldsymbol{x^{(i)}} + \boldsymbol{b}\right) - \boldsymbol{1}\right]$$

$$\boldsymbol{min\ L(w,b,\xi,\alpha,r)} = \frac{\boldsymbol{1}}{\boldsymbol{2}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{m}}\boldsymbol{\alpha_i y^{(i)}\alpha_j y^{(j)} x^{(i)^T} x^{(j)}} + \boldsymbol{(C - r_i - \alpha_i)}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\boldsymbol{\xi_i} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\boldsymbol{\alpha_i}\left[\boldsymbol{y^{(i)}}\left(\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\boldsymbol{\alpha_i y^{(i)} x^{(i)}}\right)^{\boldsymbol{T}}\boldsymbol{x^{(i)}} + \boldsymbol{b}\right) - \boldsymbol{1}\right]$$

Sustituyendo $\boldsymbol{C - \alpha_i - r_i = 0}$:

$$\boldsymbol{min\ L(w,b,\xi,\alpha,r)} = \frac{\boldsymbol{1}}{\boldsymbol{2}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{m}}\boldsymbol{\alpha_i y^{(i)}\alpha_j y^{(j)} x^{(i)^T} x^{(j)}} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\boldsymbol{\alpha_i}\left[\boldsymbol{y^{(i)}}\left(\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\boldsymbol{\alpha_i y^{(i)} x^{(i)}}\right)^{\boldsymbol{T}}\boldsymbol{x^{(i)}} + \boldsymbol{b}\right) - \boldsymbol{1}\right]$$

MÁQUINA DE SV

SVM CON REGULARIZACIÓN L1

$$\min L(w, b, \xi, \alpha, r) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i y^{(i)} \alpha_j y^{(j)} x^{(i)^T} x^{(j)} - \sum_{i=1}^{m} \alpha_i \left[y^{(i)} \left(\left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \right)^T x^{(i)} + b \right) - 1 \right]$$

$$\min L(w, b, \xi, \alpha, r) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i y^{(i)} \alpha_j y^{(j)} x^{(i)^T} x^{(j)} - \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i y^{(i)} \alpha_j y^{(j)} x^{(i)^T} x^{(j)} - b\sum_{i=1}^{m} \alpha_i y^{(i)} + \sum_{i=1}^{m} \alpha_i$$

$$L(w, b, \alpha) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i y^{(i)} \alpha_j y^{(j)} x^{(i)^T} x^{(j)} - b\sum_{i=1}^{m} \alpha_i y^{(i)} + \sum_{i=1}^{m} \alpha_i$$

Sustituyendo $\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$:

$$\min L(w, b, \xi, \alpha, r) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i y^{(i)} \alpha_j y^{(j)} x^{(i)^T} x^{(j)} + \sum_{i=1}^{m} \alpha_i$$

MÁQUINA DE SV

SVM CON REGULARIZACIÓN L1

Se tiene el **mismo problema dual** pero **con regularización**:

$$\begin{aligned} \max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m \\ & \sum_{i=1}^{m} \alpha_i y^{(i)} = 0, \end{aligned}$$

Lo **único** que **cambia** es la **restricción** para $\boldsymbol{\alpha_i}$**.**

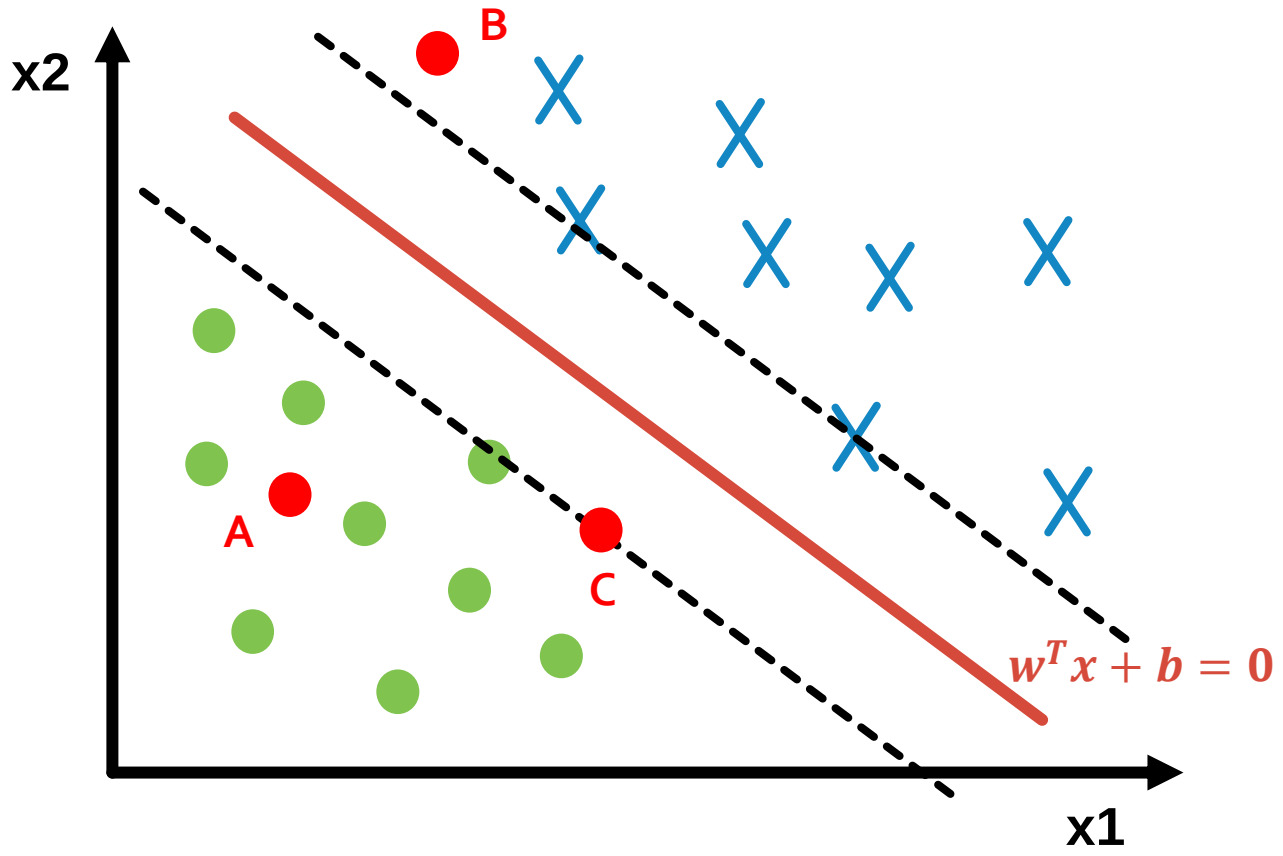
MÁQUINA DE SV

SVM CON REGULARIZACIÓN L1

Las **nuevas condiciones KKT complementarias** serían:

$$\alpha_i = 0 \Rightarrow y^{(i)}(w^T x^{(i)} + b) \geq 1 \text{ A}$$
$$\alpha_i = C \Rightarrow y^{(i)}(w^T x^{(i)} + b) \leq 1 \text{ B}$$
$$0 < \alpha_i < C \Rightarrow y^{(i)}(w^T x^{(i)} + b) = 1. \text{ C}$$

A = vectores **NO** de **soporte**
B = vectores de **soporte NO marginales**
C = vectores de **soporte marginales**):



MÁQUINA DE SV

ALGORITMO SMO

Hasta el momento **no hemos** discutido la **forma** en **como solucionar** el **problema dual**.

$$\max_\alpha \quad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0,$$

Para esto, se verá antes **otro algoritmo** denominado **ascenso** de **coordenadas** que permita **hacer eficiente** la **optimización**.

Esta es **otra** de las **razones** por las que se **desarrolló** la **forma dual** del **problema** de **optimización**.

MÁQUINA DE SV

ALGORITMO SMO

ASCENSO DE COORDENADAS

Supongamos que se **tiene** el siguiente **problema** de **optimización sin restricciones**:

$$\max_{\alpha} W(\alpha_1, \alpha_2, \ldots, \alpha_m).$$

Se va a **maximizar** la **función** $\boldsymbol{W(\alpha)}$ respecto a **una sola** $\boldsymbol{\alpha_i}$ mientras que **los demás parámetros** se van a **dejar fijos**.

Se **repite** el **proceso** con **las demás** $\boldsymbol{\alpha}$ hasta que **converja**.

MÁQUINA DE SV

ALGORITMO SMO

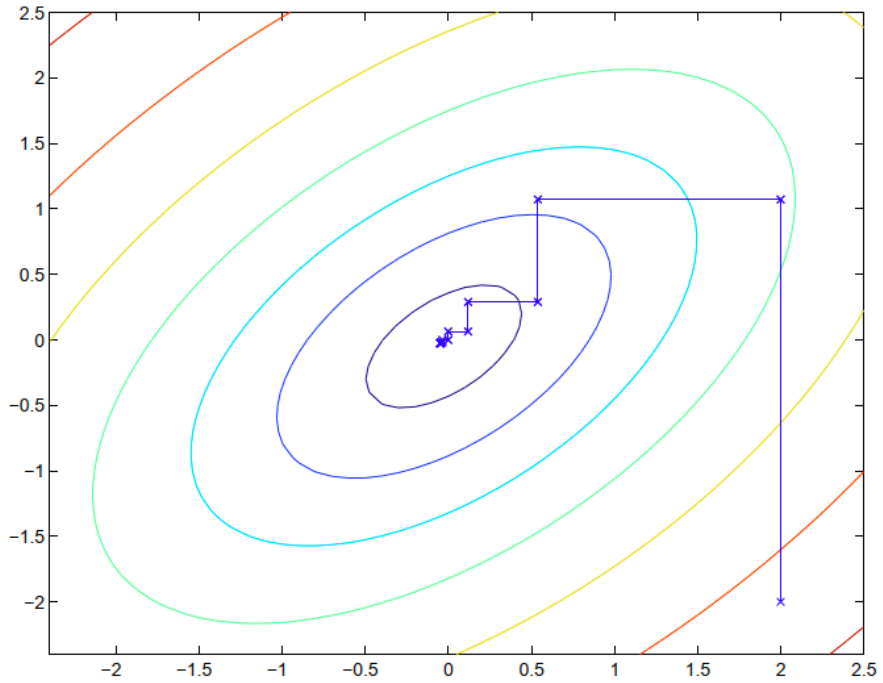
Loop until convergence: {

For $i = 1, \ldots, m$, {

$\alpha_i := \arg\max_{\hat{\alpha}_i} W(\alpha_1, \ldots, \alpha_{i-1}, \hat{\alpha}_i, \alpha_{i+1}, \ldots, \alpha_m)$.

}

}



MÁQUINA DE SV

ALGORITMO SMO

El **problema** del **algoritmo** de **ascenso** de **coordenadas** es que **no** toma en **consideración restricciones**. En el **problema dual** se tenía:

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0,$$

La idea del **algoritmo SMO (Sequential Minimal Optimization)** es que se **cambiarán dos** $\boldsymbol{\alpha_i}$ y $\boldsymbol{\alpha_j}$ **simultáneamente**.

MÁQUINA DE SV

ALGORITMO SMO

Repeat till convergence {

1. Select some pair α_i and α_j to update next (using a heuristic that tries to pick the two that will allow us to make the biggest progress towards the global maximum).
2. Reoptimize $W(\alpha)$ with respect to α_i and α_j, while holding all the other α_k's $(k \neq i, j)$ fixed.

}

MÁQUINA DE SV

ALGORITMO SMO

Se **seleccionan dos** α_1 y α_2 que **serán variables** mientras que **todas las demás** α serán tratadas como **constantes**. Sí esto es cierto entonces **se tiene** para la siguiente **restricción** lo siguiente:

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0,$$

$$\alpha_1 y^{(1)} + \alpha_2 y^{(2)} = -\sum_{i=3}^{m} \alpha_i y^{(i)}.$$

$$\alpha_1 y^{(1)} + \alpha_2 y^{(2)} = \zeta.$$

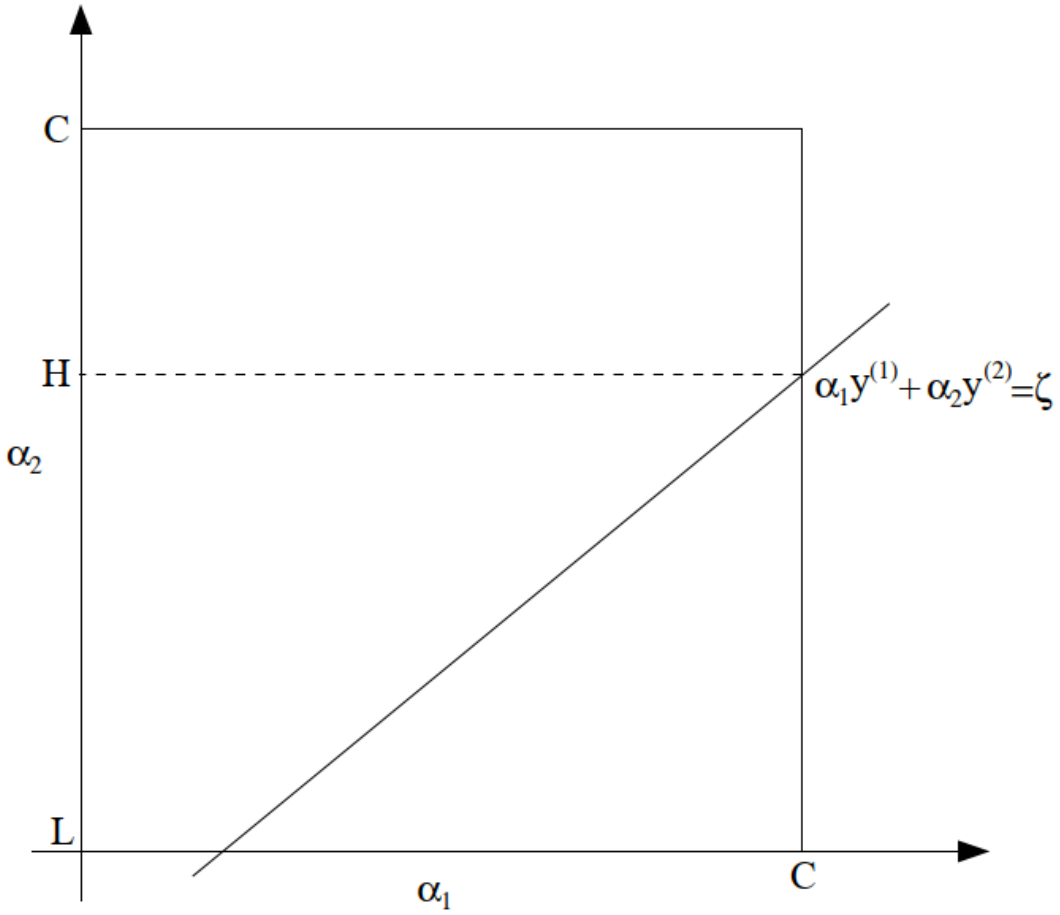
MÁQUINA DE SV

ALGORITMO SMO

Gráficamente las **dos restricciones** se podría **representar** de la **siguiente forma**:

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0,$$

$$0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m$$



MÁQUINA DE SV

ALGORITMO SMO

Podemos **representar** α_1 en **función** de α_2, por lo que:

$$\alpha_1 = (\zeta - \alpha_2 y^{(2)}) y^{(1)}.$$

Lo que implica que **nuestro objetivo** se puede **representar como**:

$$W(\alpha_1, \alpha_2, \ldots, \alpha_m) = W((\zeta - \alpha_2 y^{(2)}) y^{(1)}, \alpha_2, \ldots, \alpha_m).$$

$$\boldsymbol{max_\alpha W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j {x^{(i)}}^T x^{(j)}}$$

$$\boldsymbol{max_\alpha W(\alpha) = \left(\xi - \alpha_2 y^{(2)}\right) y^{(1)} + \alpha_2 - \frac{1}{2} \sum_{i=1}^{2} \sum_{j=1}^{2} y^{(i)} y^{(j)} \alpha_i \alpha_j {x^{(i)}}^T x^{(j)}}$$

MÁQUINA DE SV

ALGORITMO SMO

Podemos **representar** $\boldsymbol{\alpha_1}$ en **función** de $\boldsymbol{\alpha_2}$, por lo que:

$$\alpha_1 = (\zeta - \alpha_2 y^{(2)}) y^{(1)}.$$

Lo que implica que **nuestro objetivo** se puede **representar como**:

$$W(\alpha_1, \alpha_2, \ldots, \alpha_m) = W((\zeta - \alpha_2 y^{(2)}) y^{(1)}, \alpha_2, \ldots, \alpha_m).$$

$$\boldsymbol{W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j x^{(i)^T} x^{(j)}}$$

$$\boldsymbol{W(\alpha) = \left(\xi - \alpha_2 y^{(2)}\right) y^{(1)} + \alpha_2 - \frac{1}{2} \sum_{i=1}^{2} \sum_{j=1}^{2} y^{(i)} y^{(j)} \alpha_i \alpha_j x^{(i)^T} x^{(j)}}$$

MÁQUINA DE SV

ALGORITMO SMO

Vemos que $\boldsymbol{W(\alpha)}$ es una **función cuadrática** en $\boldsymbol{\alpha_2}$

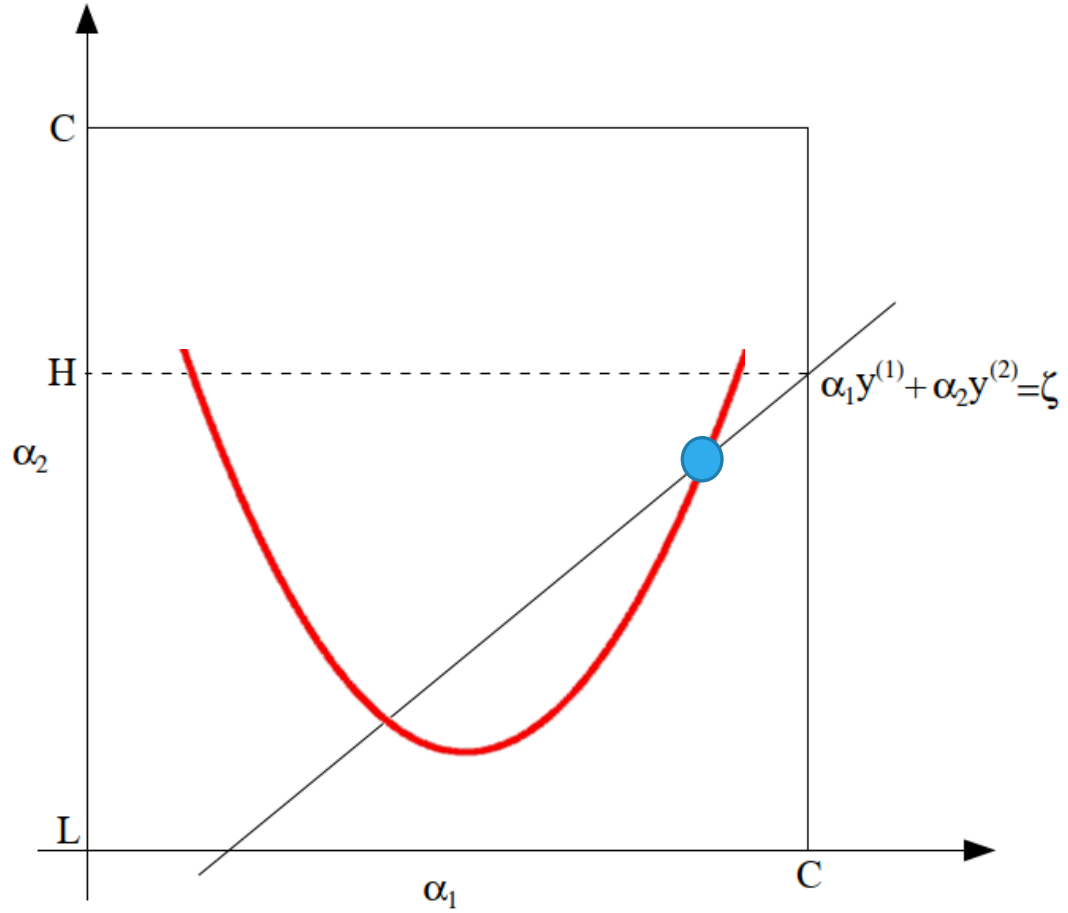
$$\boldsymbol{W(\alpha) = \left(\xi - \alpha_2 y^{(2)}\right) y^{(1)} + \alpha_2 - \frac{1}{2} \sum_{i=1}^{2} \sum_{j=1}^{2} y^{(i)} y^{(j)} \alpha_i \alpha_j x^{(i)^T} x^{(j)}}$$

$$\boldsymbol{W(\alpha) = a{\alpha_2}^2 + b\alpha_2 + c}$$

Es una **función fácil** de **optimizar**.

MÁQUINA DE SV

ALGORITMO SMO



Se **sustituye** el **valor óptimo** de $\boldsymbol{\alpha_2}$ en la **recta** para obtener $\boldsymbol{\alpha_1}$

MÁQUINA DE SV

ALGORITMO SMO

Vemos que $\boldsymbol{W(\alpha)}$ es una **función cuadrática** en $\boldsymbol{\alpha_2}$

$$\boldsymbol{W(\alpha) = \left(\xi - \alpha_2 y^{(2)}\right) y^{(1)} + \alpha_2 - \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2} y^{(i)} y^{(j)} \alpha_i \alpha_j x^{(i)^T} x^{(j)}}$$

$$\boldsymbol{W(\alpha) = a{\alpha_2}^2 + b\alpha_2 + c}$$

Es una **función fácil** de **optimizar**. Una vez optimizada, el **último paso** sería:

K VECINOS

K V E C I N O S

F A L T A

A

ÁRBOLES DE DECISIÓN

ÁRBOLES DE DECISIÓN

FALTA

A