MAQUINAS DE VECTORES DE SOPORTE

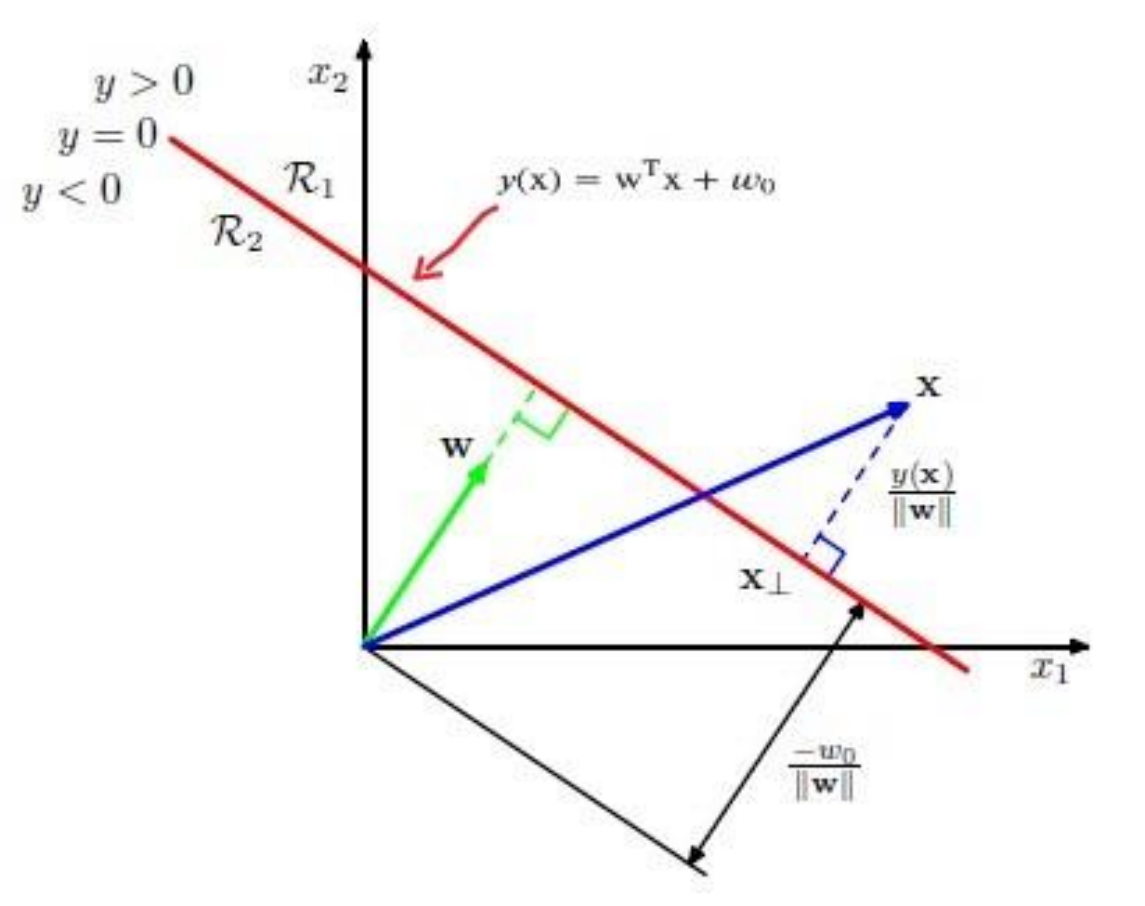
Dr. Jorge Hermosillo
Laboratorio de Semántica Computacional

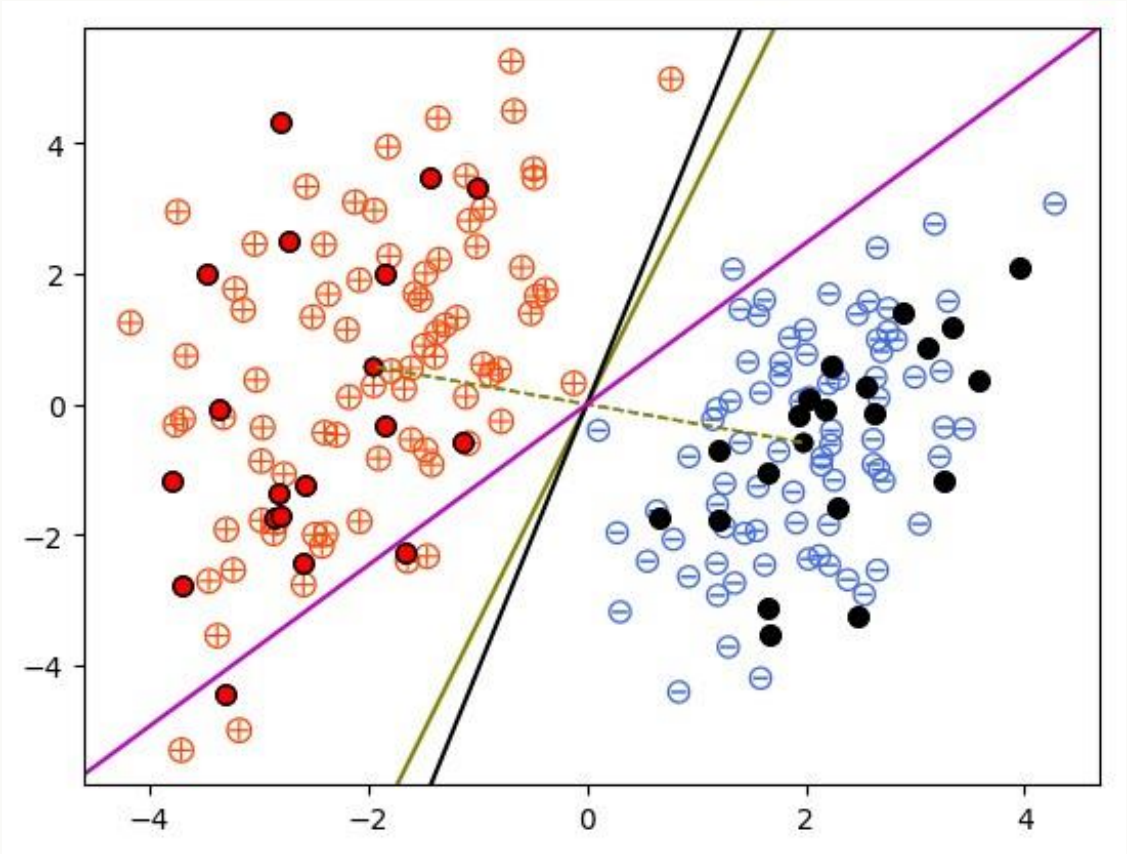
Centro de Investigación en Ciencias

Universidad Autónoma del Estado de Morelos

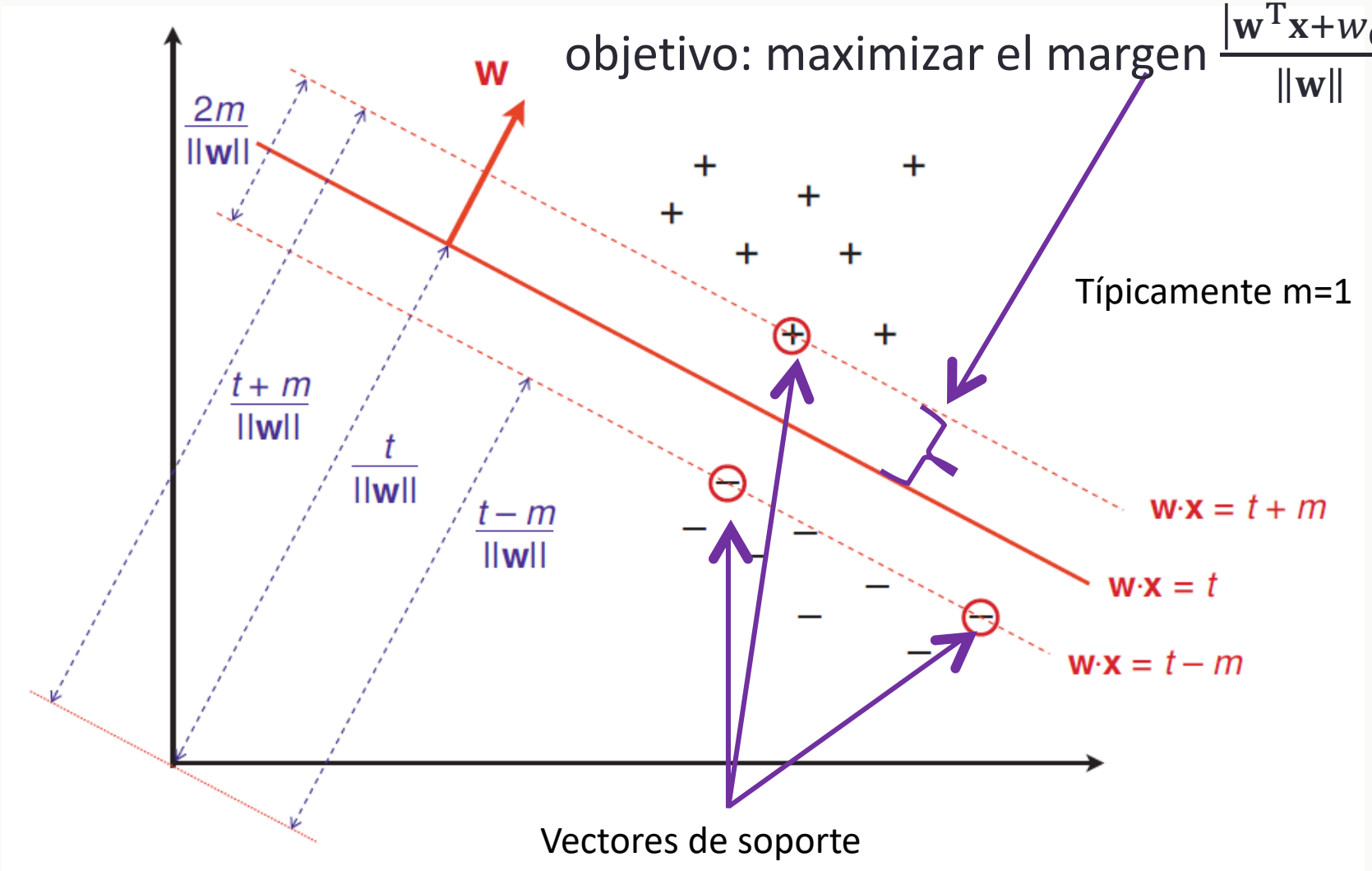
SVM DE MARGEN DURO

SVM: MODELO LINEAL ÓPTIMO





CLASIFICADOR POR VECTORES DE SOPORTE



PROBLEMA DE OPTIMIZACIÓN CON RESTRICCIONES

Maximizar el margen es equivalente a **minimizar** $\mathbf{w}$ o mejor aún $\frac{1}{2}\|\mathbf{w}\|^2$

$$\mathbf{w}^*, t^* = \underset{\mathbf{w},t}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{w}\|^2 \text{ sujeto a } y_i\left(\langle \mathbf{w}, \mathbf{x}_i \rangle - t\right) \geq 1, \qquad 1 \leq i \leq N$$

Para ello usaremos el metodo de Multiplicadores de Lagrange.

Fuentes de consulta adicionales:

- http://www-mtl.mit.edu/Courses/6.050/2004/unit9/wyatt.apr.7.pdf
- Pattern Recognition: Concepts, Methods, and Applications. J. P. Marques de Sá. Springer Science & Business Media, 2001.
- Lagrange Multipliers Tutorial in the Contextof Support Vector Machines. Baxter Tyson Smith. http://www.engr.mun.ca/~baxter/Publications/LagrangeForSVMs.pdf

El problema de optimizacion en SVM's

- Nuestro objetivo es:

$$\mathbf{w}^*, t^* = \underset{\mathbf{w},t}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{w}\|^2$$

- Sujeto a las siguientes N restricciones:

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle - t\ ) \geq 1, \qquad 1 \leq i \leq N$$

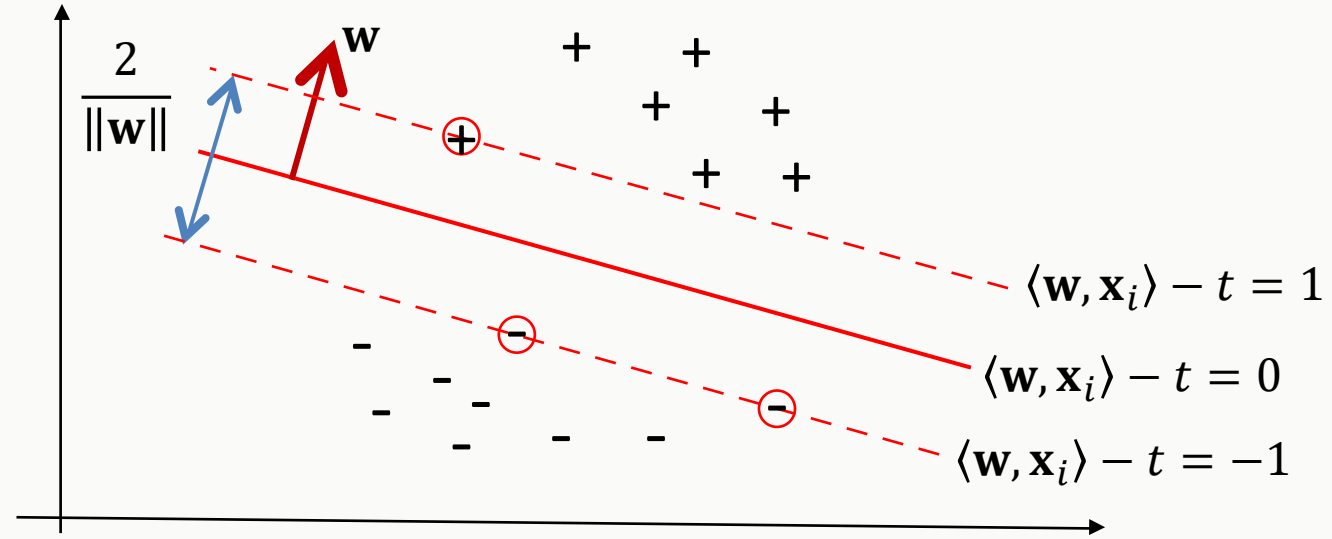
El problema de optimizacion en SVM's

- Nuestro objetivo es:

$$\mathbf{w}^*, t^* = \underset{\mathbf{w},t}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{w}\|^2$$

- Sujeto a las siguientes N restricciones:

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle - t\ ) \geq 1, \qquad 1 \leq i \leq N$$



EL PROBLEMA DE OPTIMIZACION EN SVM'S

- Definimos el lagrangiano

Minimizar **Maximizar**

$$\mathcal{L}_P(\mathbf{w}, t, \alpha_1, \cdots, \alpha_N) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{N} \alpha_i (yi\ (\langle \mathbf{w}, \mathbf{x}_i \rangle - t\ ) - \ 1)$$

- $$= \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{N} \alpha_i y_i \langle \mathbf{w}, \mathbf{x}_i \rangle + \sum_{i=1}^{N} \alpha_i y_i t + \sum_{i=1}^{N} \alpha_i$$

$$= \frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle - \left\langle \mathbf{w}, \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \right\rangle + t\left(\sum_{i=1}^{N} \alpha_i y_i\right) + \sum_{i=1}^{N} \alpha_i$$

- Para un t óptimo $\partial_t \mathcal{L}_P = 0 \Longrightarrow \sum_{i=1}^{N} \alpha_i y_i = 0$
- Para pesos óptimos $\partial_{\mathbf{w}} \mathcal{L}_P = 0 \Longrightarrow \mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i$

MODELO DUAL DE OPTIMIZACIÓN

- *Reinsertando estas expresiones en $\mathcal{L}_P$ obtenemos $\mathcal{L}_D$ el lagrangiano del problema dual:*

$$\mathcal{L}_D(\alpha_1, \cdots, \alpha_N) = -\frac{1}{2}\left\langle \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \, , \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \right\rangle + \sum_{i=1}^{N} \alpha_i$$

$$= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \, \alpha_j y_i y_j \left\langle \mathbf{x}_i, \mathbf{x}_j \right\rangle + \sum_{i=1}^{N} \alpha_i$$

MODELO DUAL DE OPTIMIZACIÓN

- *El problema de optimización dual es el siguiente:*

$$\alpha_1^*, \cdots, \alpha_N^* = \underset{\alpha_1, \cdots, \alpha_N}{\operatorname{argmax}} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \, \alpha_j y_i y_j \left\langle \mathbf{x}_i, \mathbf{x}_j \right\rangle + \sum_{i=1}^{N} \alpha_i$$

- *Sujeto a las restricciones:*

$$\alpha_i > 0 \ , \ 1 \leq i \leq N \qquad \text{y} \qquad \sum_{i=1}^{N} \alpha_i y_i = 0$$

ASPECTOS IMPORTANTES

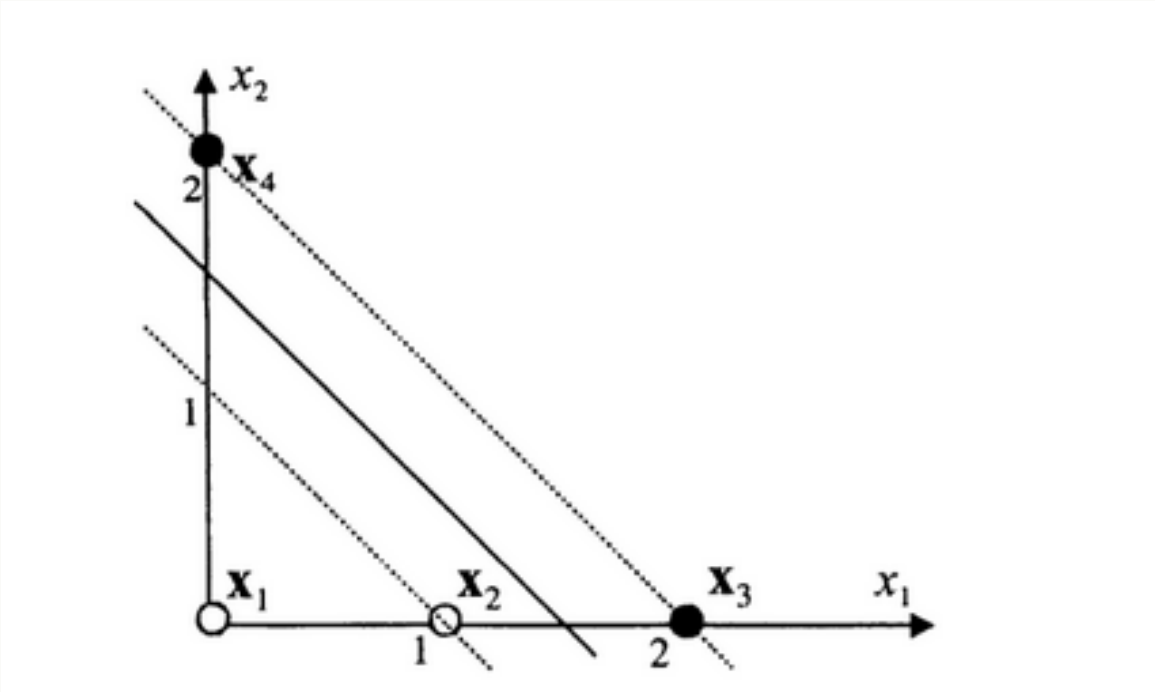
- *La forma dual del problema de optimización de las SVM ilustra dos aspectos importantes :*

$$\alpha_1^*, \cdots, \alpha_N^* = \underset{\alpha_1,\cdots,\alpha_N}{\operatorname{argmax}} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\,\alpha_j y_i y_j \left\langle \mathbf{x}_i, \mathbf{x}_j \right\rangle + \sum_{i=1}^{N}\alpha_i$$

1. Maximizar el margen es equivalente a encontrar los vectores de soporte; es decir, los puntos para los cuales los multiplicadores de Lagrange son no nulos.
2. El problema de optimización está completamente definido por el producto punto de pares de instancias de entrenamiento: las entradas de la matriz Gram.

EJEMPLO SENCILLO

- Encuentra W óptimo para este problema: X1=[0,0] X2=[1,0] para la clase (+1) y X3=[2,0] y X4=[0,2] para la clase (-1)



EJEMPLO SENCILLO

$$\mathcal{L}_D(\alpha_1, \cdots, \alpha_N) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\, \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

$$\mathcal{L}_D = (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) - \frac{1}{2}\left(\alpha_2{}^2 - 4\alpha_2\alpha_3 + 4\alpha_3{}^2 + 4\alpha_4{}^2\right)$$

Diferenciando con respecto a los α's y utilizando la restricción $\sum_{i=1}^{N} \alpha_i y_i = 0$ obtenemos:

$$\begin{cases} \alpha_1 + \alpha_2 - \alpha_3 - \alpha_4 = 0 \\ \alpha_2 - 2\alpha_3 = 1 \\ -2\alpha_2 + 4\alpha_3 = 1 \\ 4\alpha_4 = 1 \end{cases}$$

de donde: $\alpha_1 = 0,\ \alpha_2 = 1,\ \alpha_3 = \frac{3}{4},\ \alpha_4 = 1/4$

Aplicando: $\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i$, finalmente obtenemos

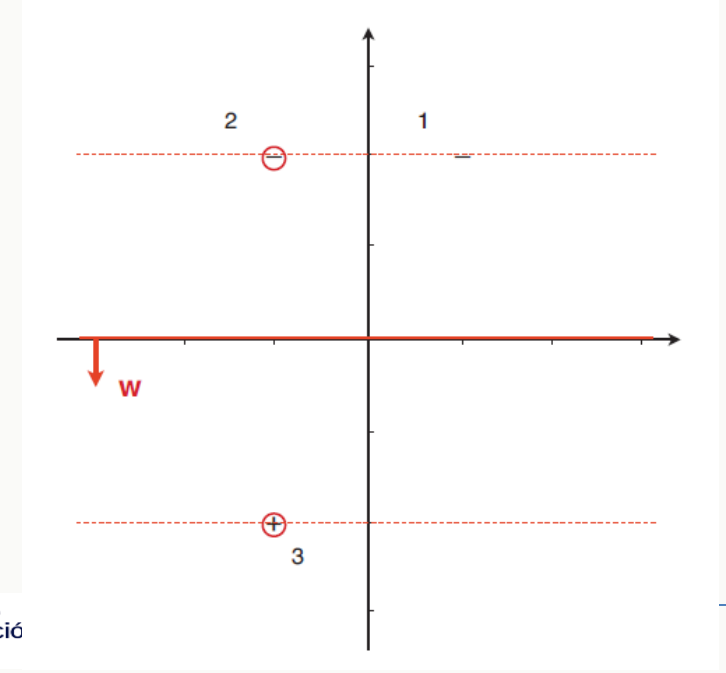
$\mathbf{w} = \begin{bmatrix} -1/2 \\ -1/2 \end{bmatrix}$, $w_0 = 3/4$ y $d(x) = 3 - 2\mathrm{x}_1 - 2\mathrm{x}_2 = 0$

EJERCICIO

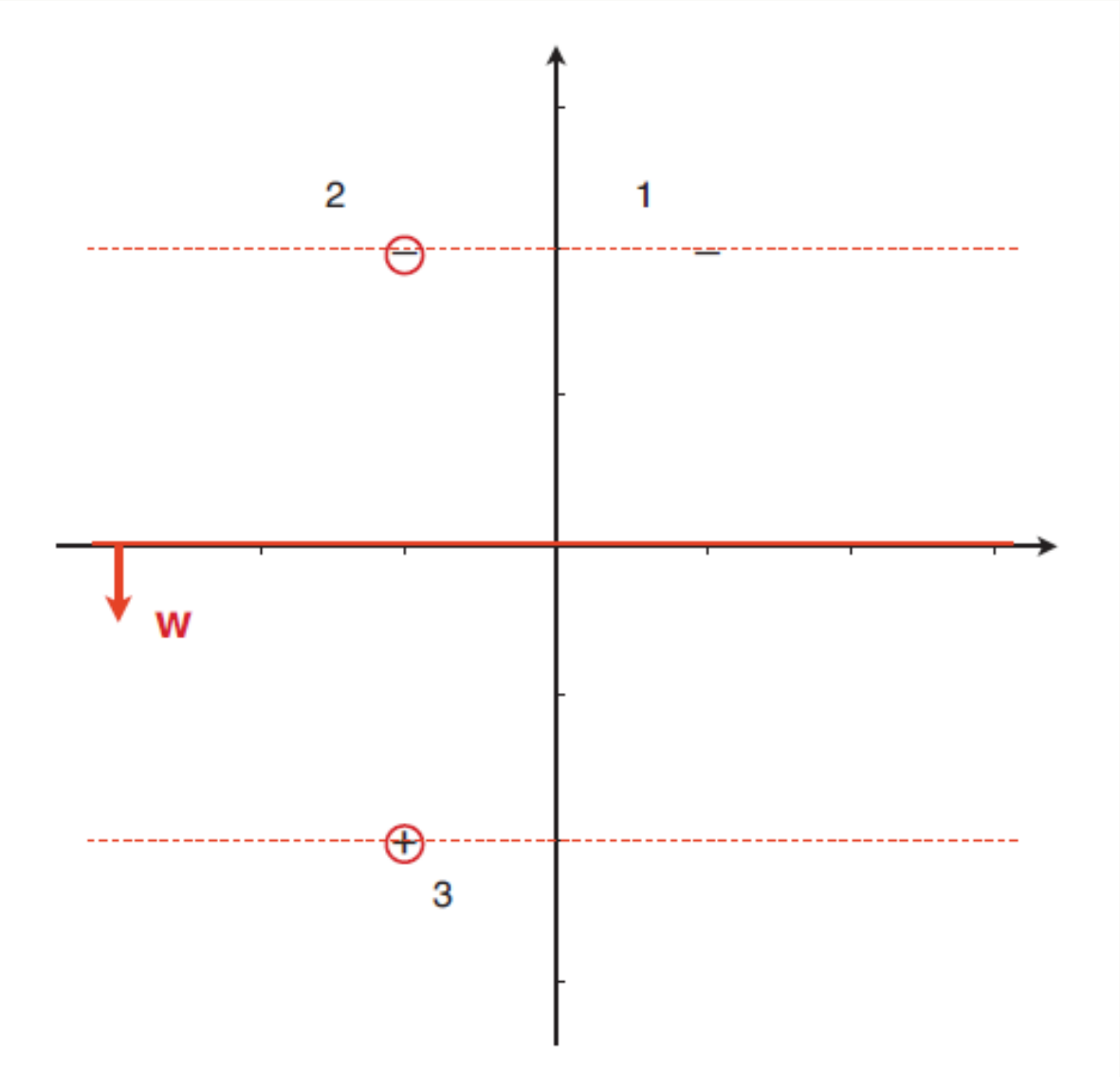
Let the data points and labels be as follows (see Figure):

$$\mathbf{X} = \begin{pmatrix} 1 & 2 \\ -1 & 2 \\ -1 & -2 \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} -1 \\ -1 \\ +1 \end{pmatrix} \quad \mathbf{X}' = \begin{pmatrix} -1 & -2 \\ 1 & -2 \\ -1 & -2 \end{pmatrix}$$

The matrix $\mathbf{X}'$ on the right incorporates the class labels; i.e., the rows are $y_i \mathbf{x}_i$.



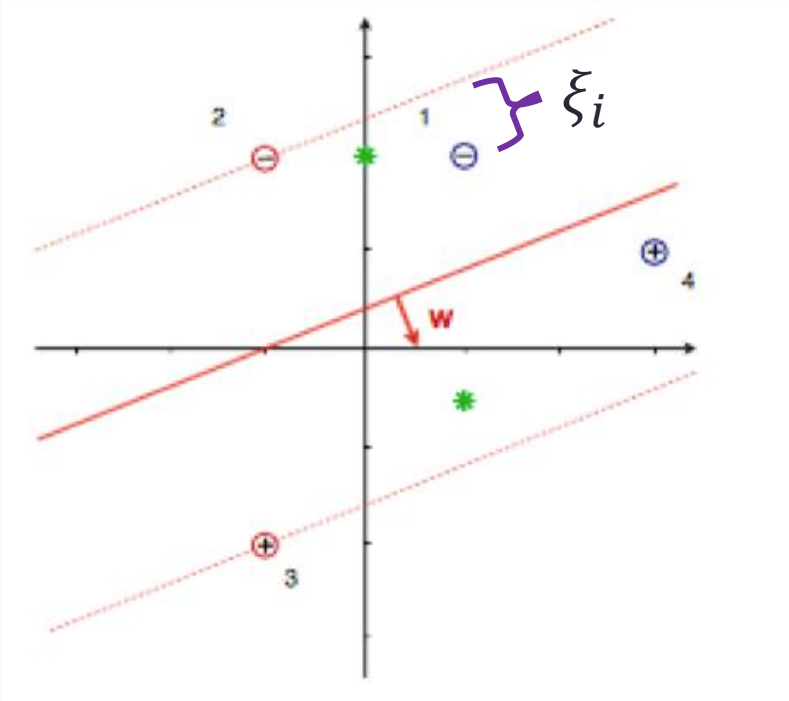
1. Encuentre la matriz Gram
2. Exprese el problema de optimización Dual
3. Encuentre los vectores de soporte


2
1
w
3

SVM DE MARGEN SUAVE

SVM con margen suave

- La SVM anterior no funciona con datos no-separables
- Introducimos variables de holgura ξ_i para cada dato de entrada, lo que les permite a algunos de ellos estar dentro del margen, o incluso del lado equivocado de la frontera de decision.



SVM con margen suave

$$\mathbf{w}^*, t^*, \xi_i^* = \underset{\mathbf{w},t,\xi_i}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{sujeto a } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle - t) \geq 1 - \xi_i \text{ y } \xi_i \geq 0, 1 \leq i \leq N$$

- C es un parámetro definido por el usuario que balancea la maximización del margen contra la minimización de las variables de holgura:
  - *un valor alto de C significa que los errores de margen son altamente costosos,*
  - *un valor pequeño de C permite más errores de margen con tal de hacer mas grande el margen.*
- Si permitimos más errores de margen necesitamos menos vectores de soporte, por lo tanto C controla la 'complejidad' de la SVM y por ello se le denomina el *parámetro de complejidad*.

SVM con margen suave

- Buscamos soluciones mediante el nuevo Lagrangiano:

$$\mathcal{L}(\mathbf{w}, t, \xi_i, \alpha_i, \beta_i) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\big(yi\,(\langle \mathbf{w}, \mathbf{x}_i\rangle - t\,) - (1-\xi_i)\big) - \sum_{i=1}^{N}\beta_i\xi_i$$

$$= \mathcal{L}(\mathbf{w}, t, \alpha_i\,) + \sum_{i=1}^{N}(C - \alpha_i - \beta_i)\xi_i$$

- La solución óptima es tal que $\partial_{\xi_i}\mathcal{L} = 0 \Longrightarrow$ el término añadido desaparece en el problema dual.
- Además, puesto que α_i y β_i son positivos, α_i no puede ser mayor a C:

$$\alpha_1^*, \cdots, \alpha_N^* = \underset{\alpha_1,\cdots,\alpha_N}{\operatorname{argmax}} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\,\alpha_j y_i y_j \left\langle \mathbf{x}_i, \mathbf{x}_j \right\rangle + \sum_{i=1}^{N}\alpha_i$$

Sujeto a las restricciones: $0 \le \alpha_i \le C\ ,\ 1 \le i \le N$ y $\sum_{i=1}^{N}\alpha_i y_i = 0$

Significado de *C* como cota superior para α_i

- En el caso óptimo, para cada ejemplo (dato de entrada) se debe cumplir

$$C - \alpha_i - \beta_i = 0$$

- Distinguimos tres casos para los ejemplos:
  1. $\alpha_i = 0$ *significa que están fuera o sobre el margen.*
  2. $0 < \alpha_i < C$ *estos son los vectores de soporte sobre el margen.*
  3. $\alpha_i = C \Longrightarrow \beta_i = 0$ *estos están sobre o dentro del margen.*

- Como todavía tenemos: $\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i$ los últimos dos casos contribuyen a expandir el margen.

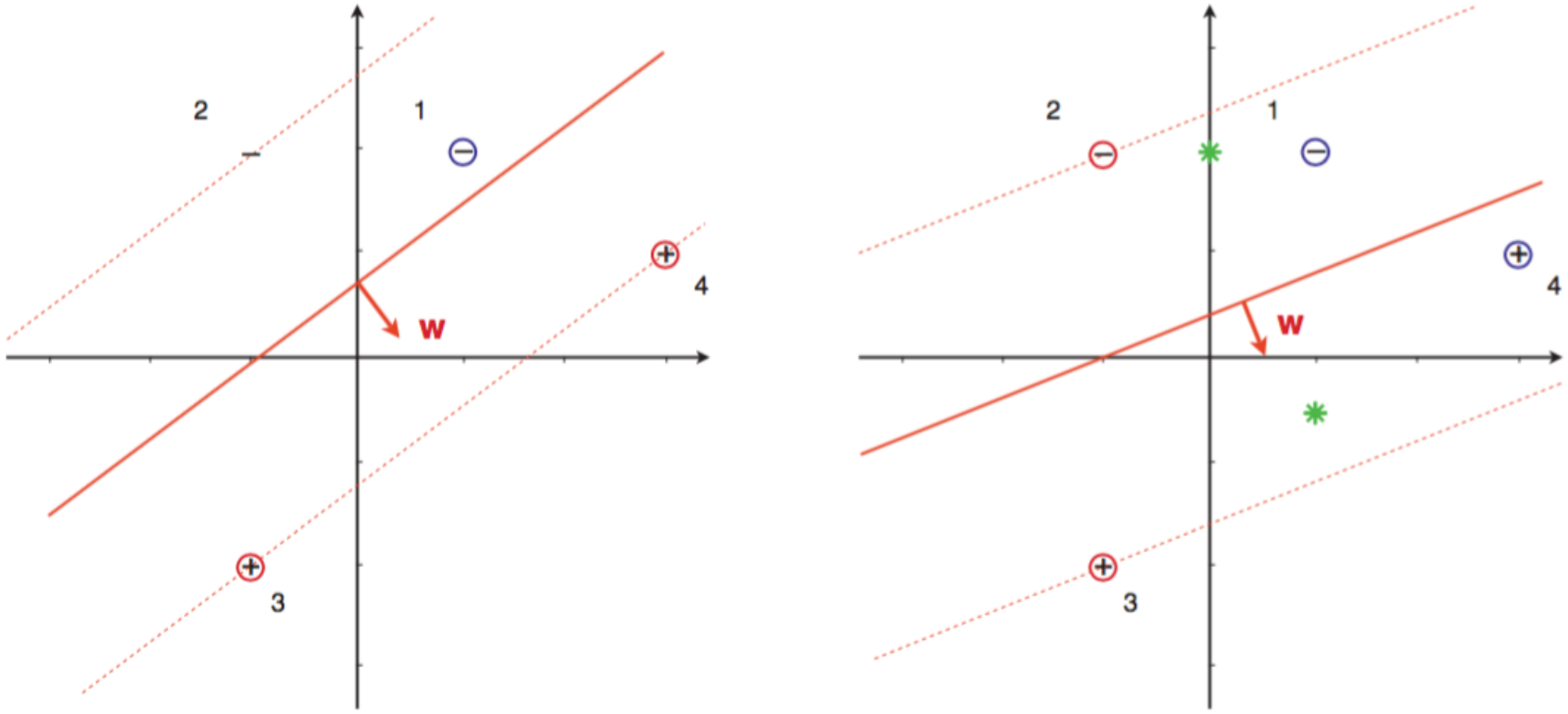


Figure 7.9. **(left)** The soft margin classifier learned with $C = 5/16$, at which point $\mathbf{x}_2$ is about to become a support vector. **(right)** The soft margin classifier learned with $C = 1/10$: all examples contribute equally to the weight vector. The asterisks denote the class means, and the decision boundary is parallel to the one learned by the basic linear classifier.

Clasificador lineal “básico”

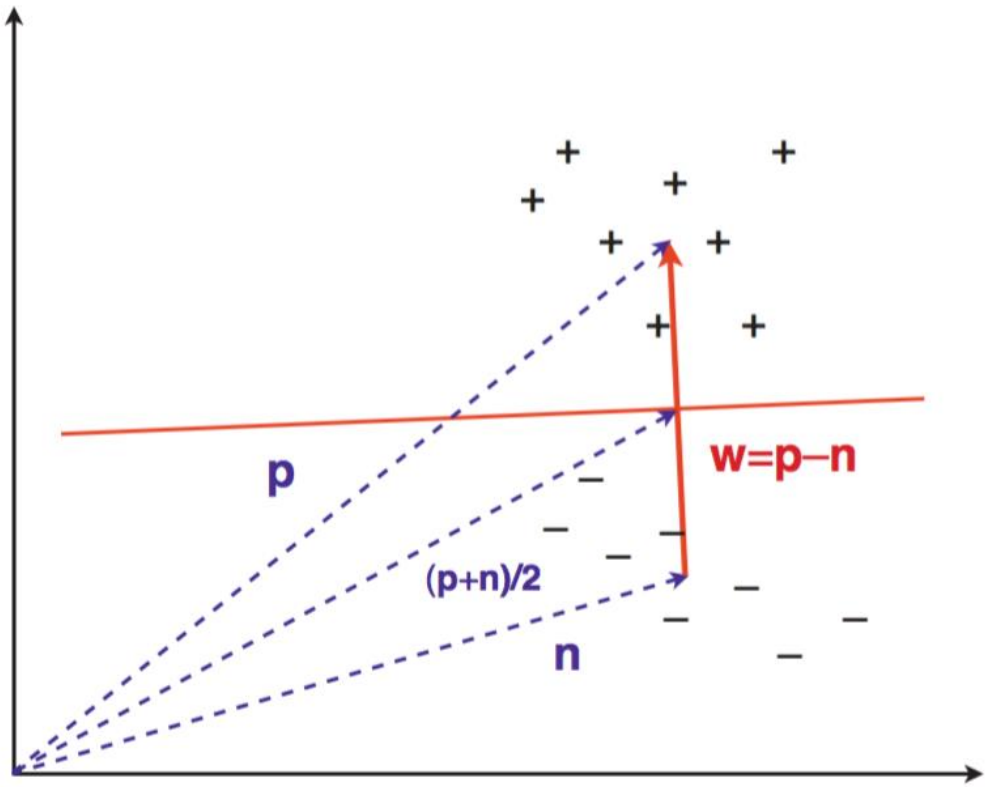
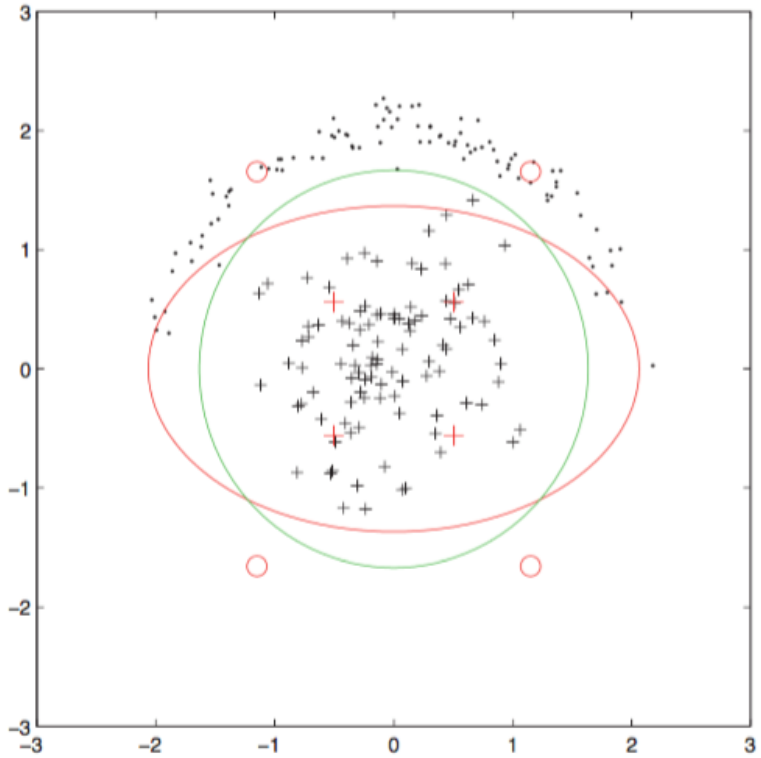


Figure 1.1. The basic linear classifier constructs a decision boundary by half-way intersecting the line between the positive and negative centres of mass. It is described by the equation $\mathbf{w} \cdot \mathbf{x} = t$, with $\mathbf{w} = \mathbf{p} - \mathbf{n}$; the decision threshold can be found by noting that $(\mathbf{p} + \mathbf{n})/2$ is on the decision boundary, and hence $t = (\mathbf{p} - \mathbf{n}) \cdot (\mathbf{p} + \mathbf{n})/2 = (||\mathbf{p}||^2 - ||\mathbf{n}||^2)/2$, where $||\mathbf{x}||$ denotes the length of vector $\mathbf{x}$.

Mas allá de la clasificación lineal: métodos de Kernel



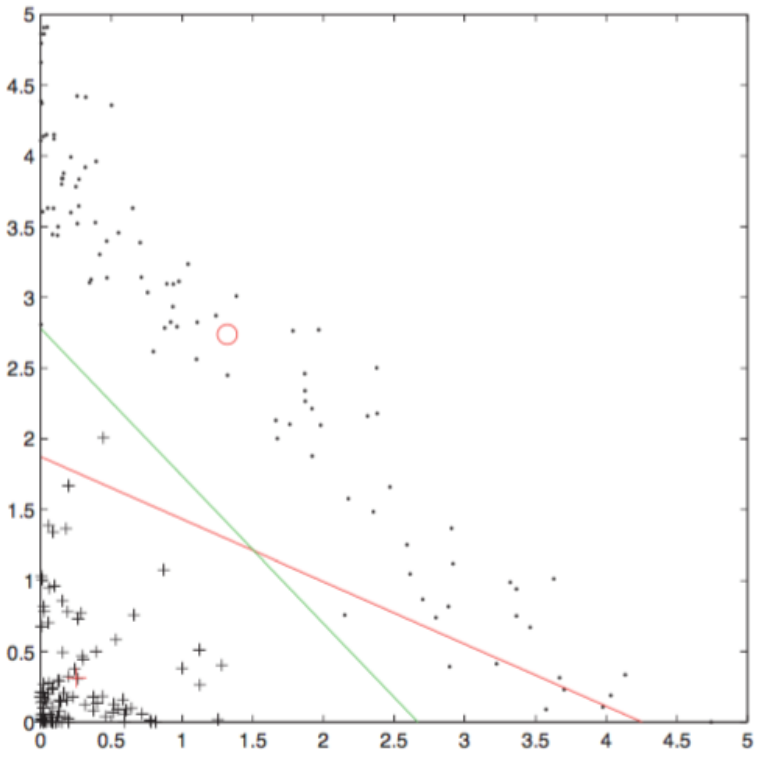


Figure 7.14. (left) Decision boundaries learned by the basic linear classifier and the perceptron using the square of the features. **(right)** Data and decision boundaries in the transformed feature space.

Kernels básicos

- Aunque nuevos kernels aparecen en la literatura, los siguientes cuatro son básicos y ampliamente utilizados:

Lineal:	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$
Polinomial:	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = (\langle \mathbf{x}_i, \mathbf{x}_j \rangle + r)^p, r \geq 0$
Gaussiano (*Radial Basis Function* – RBF):	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$
Sigmoide:	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma \langle \mathbf{x}_i, \mathbf{x}_j \rangle + r)$

donde r, p, γ son parámetros de los modelos.

SVM's con kernels

- El "truco" del kernel (*kernel trick*) es comúnmente empleado con las máquinas de vectores de soporte:

$$\alpha_1^*, \cdots, \alpha_N^* = \underset{\alpha_1,\cdots,\alpha_N}{\operatorname{argmax}} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\, \alpha_j y_i y_j\, \kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) + \sum_{i=1}^{N} \alpha_i$$

Sujeto a las restricciones: $0 \leq \alpha_i \leq C \;,\; 1 \leq i \leq N$ y $\sum_{i=1}^{N} \alpha_i y_i = 0$

- No perder de vista:
  1. *La frontera de decisión no puede representarse como un simple vector de pesos en el espacio de entrada.*
  2. *Para clasificar un nuevo dato $\mathbf{x}_i$ se necesita evaluar:*

$$y_{\mathrm{i}} \sum_{j=1}^{N} \alpha_j y_j \kappa\left(\mathbf{x}_i, \mathbf{x}_j\right)$$

Procedimiento para clasificar con SVM's

- El siguiente procedimiento se propone en:
  - *Hsu, C., Chang, C., & Lin, C. (2016). A practical guide to support vector classification. Department of Computer Science National Taiwan University, Taipei 106, Taiwan (https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf)*

1. Transforma los datos en un formato conveniente para usar SVM's
2. Realiza un escalamiento sencillo de los datos
3. Considera el kernel RBF (gaussiano)
4. Utiliza **grid-search** y **validación cruzada** para elegir la mejor combinación de C y γ para entrenar el modelo con todos los datos
5. Prueba