Técnicas de Aprendizaje Estadístico
Support Vector Machines

Mathias Bourel

IMERL - Facultad de Ingeniería, Universidad de la República, Uruguay

29 novembre 2018

Plan

Presentación del Problema

Datos :

- Dada la muestra de entrenamiento $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$ con $\boldsymbol{x}_i \in \mathbb{R}^d$, por ejemplo $\boldsymbol{x}_i$ son características observadas en pacientes : *fiebre*$>$ 38*, tos, dolor de cabeza, dolor articulaciones, irritación ojos, flujo nasal, etc*.
- la etiqueta $y_i \in \{-1, 1\}$ indica, por ejemplo, la presencia o ausencia de *A/H1NM1 (gripe porcina)*.

Objetivo

Construir una función de decisión que clasifique nuevos datos $f : \mathbb{R}^d \longrightarrow \{-1, 1\}$

Diagnóstico

f(nuevo paciente)

Presentación del Problema - Vapnik 1995

Buscar una frontera de decisión para clasificar nuevos ejemplos

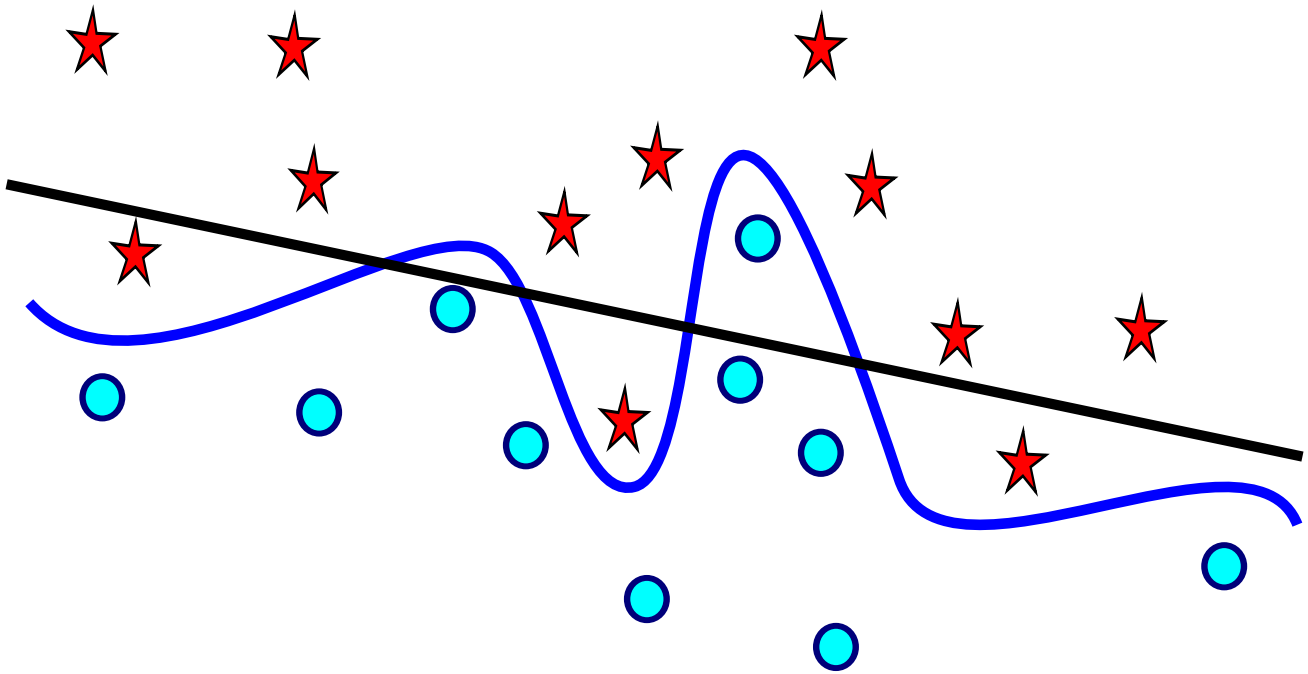


FIGURE – Los datos no son linealmente separables

SVM Caso 1) Linealmente separables

Buscamos el “mejor” hiperplano que separe los datos, es decir, que “pase” lo mas lejos posible de todos ellos.

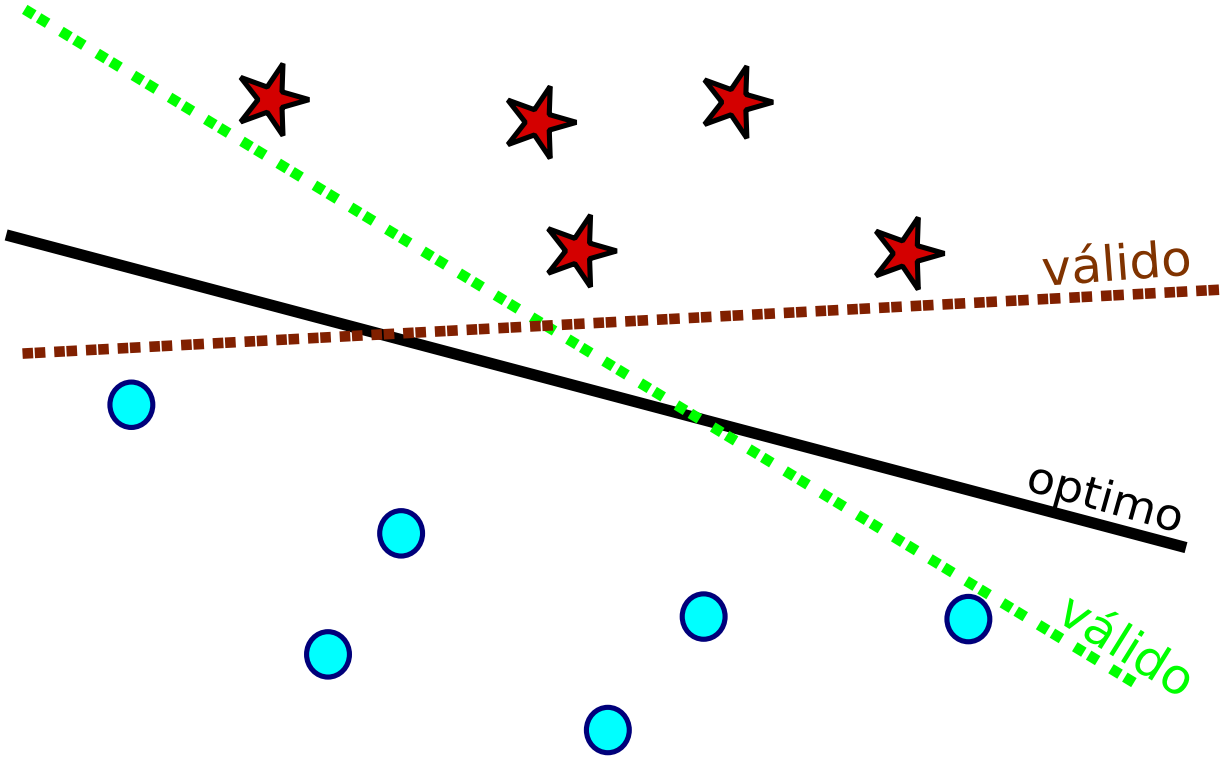


FIGURE – Los datos son linealmente separables, y hay infinitos hiperplanos que los separan

SVM Caso 1) Linealmente separables

Tenemos datos $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$, $\boldsymbol{x}_i \in \mathbb{R}^d$ e $y_i \in \{-1, 1\}$, sea $\beta \in \mathbb{R}^d$ tal que $\|\beta\| = 1$, si $H(\boldsymbol{x}) = \langle \beta, \boldsymbol{x} \rangle + \beta_0$ entonces $d(\boldsymbol{x}_j, H_1) = |H(\boldsymbol{x}_j)|$.

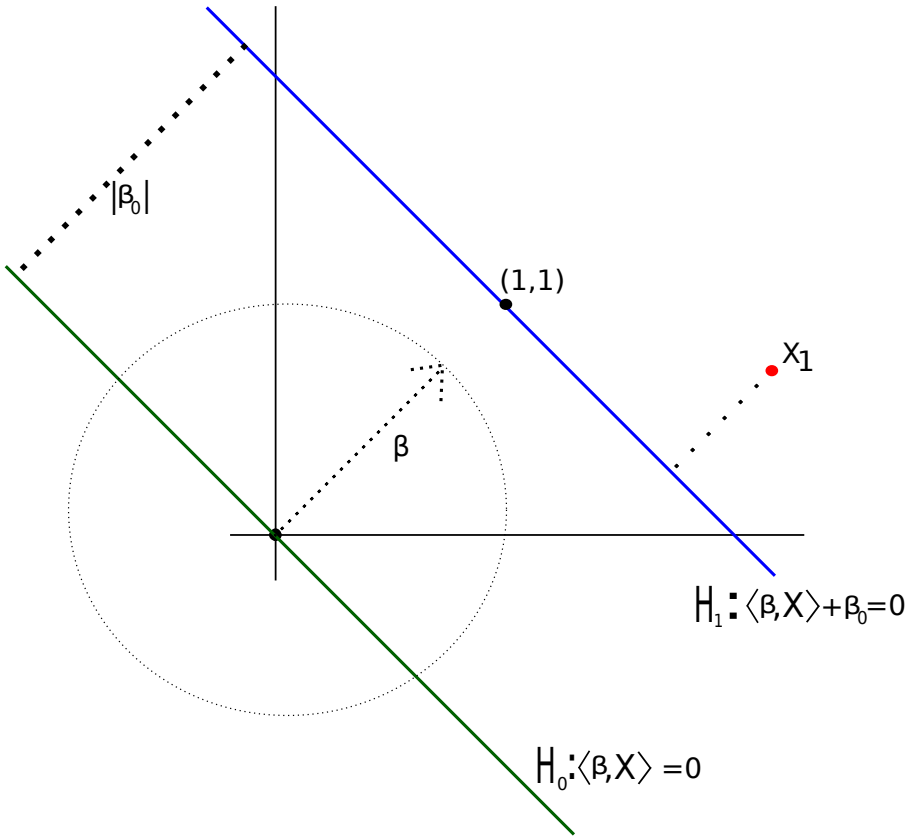


FIGURE – La distancia de $\boldsymbol{x}_1$ a H_1 es $|H(\boldsymbol{x}_1)|$

SVM Caso 1) Linealmente separables

- Una nueva observación estará bien clasificada si $y_i H(\boldsymbol{x}_i) > 0$.
- Los datos deben verificar que existe $C > 0$ tal que $\forall i$, $y_i(\langle \beta, \boldsymbol{x}_i \rangle + \beta_0) \geq C$. La igualdad anterior se da cuando el dato está en alguno de los 2 hiperplanos $\langle \beta, \boldsymbol{x}_i \rangle + \beta_0 \pm C = 0$.

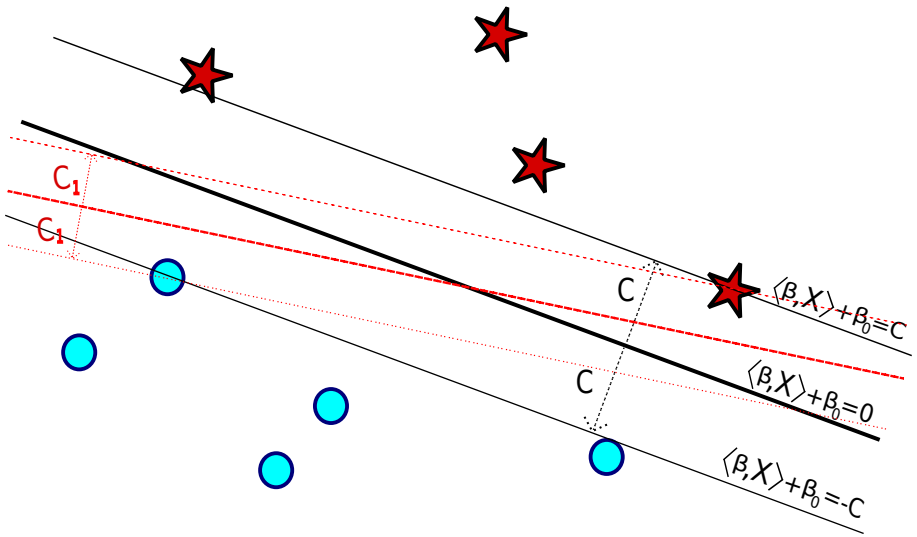


FIGURE – Si $y_i H(\boldsymbol{x}_i) > 0$ el dato $\boldsymbol{x}_i$ está bien clasificado

SVM Caso 1) Linealmente separables

- Queremos determinar β y β_0 de manera que el margen C sea lo más grande posible.
- Predicción : De qué lado del hiperplano se encuentra el nuevo dato ?

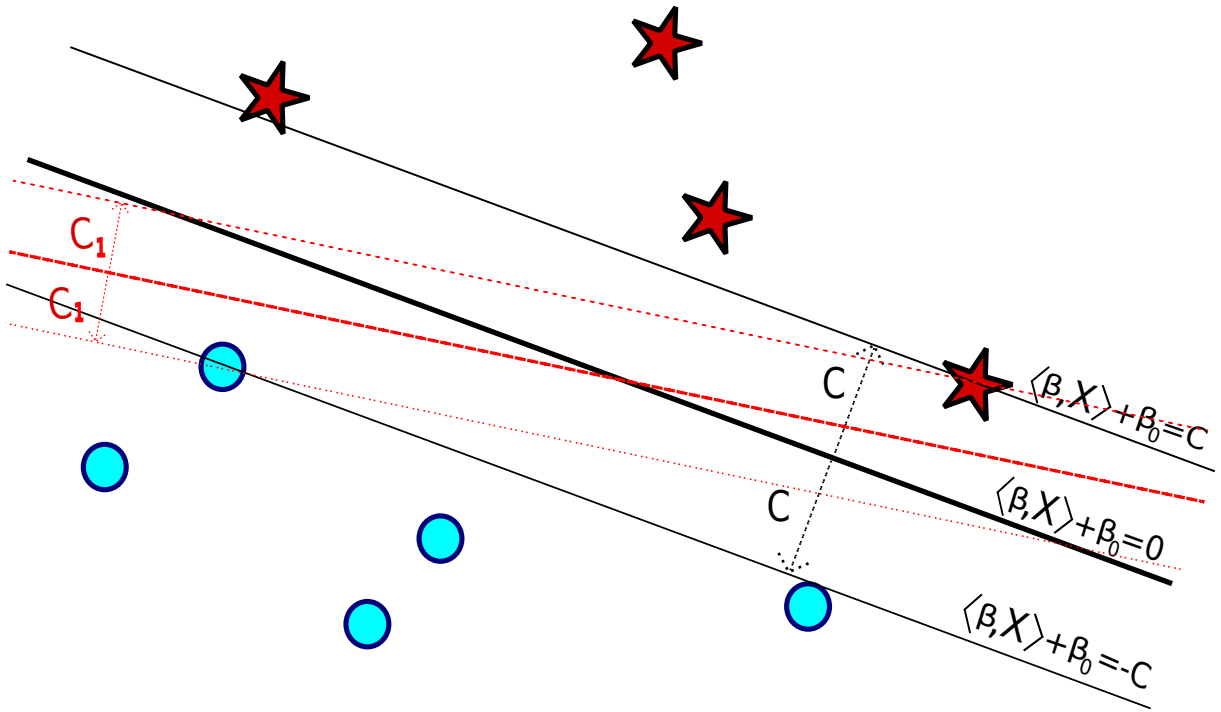


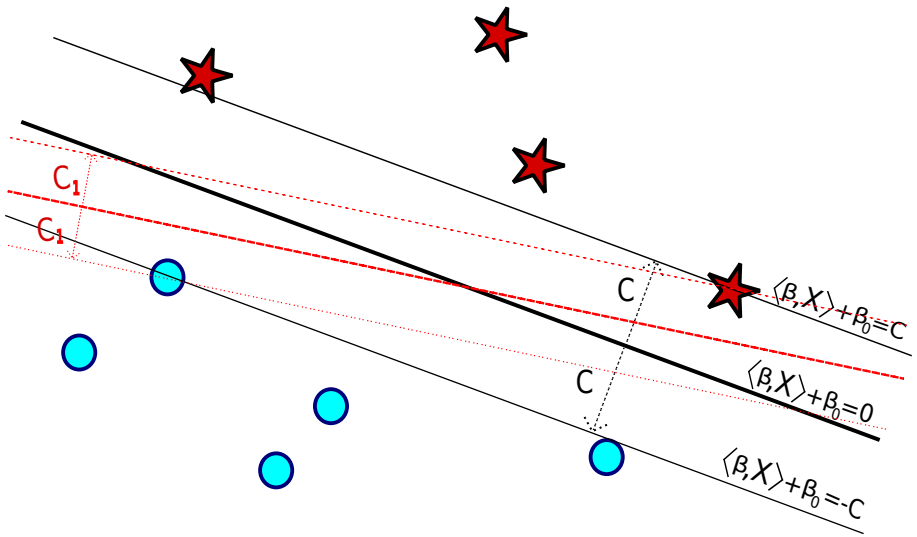
FIGURE – Si $y_i H(\boldsymbol{x}_i) > 0$ el dato $\boldsymbol{x}_i$ está bien clasificado

Regla de clasificación

$f(\boldsymbol{x}_{nue}) = sgn(\langle \beta, \boldsymbol{x}_{nue} \rangle + \beta_0)$.

SVM Caso 1) Linealmente separables

Queremos resolver
$$\begin{cases} \max\limits_{\beta,\beta_0} C(\beta,\beta_0) \\ \text{sujeto a} \\ \quad y_i(\langle \boldsymbol{x}_i,\beta\rangle + \beta_0) \geq C(\beta,\beta_0) \quad \forall i = 1,\ldots,n \\ \quad \|\beta\| = 1, \beta_0 \in \mathbb{R} \end{cases}$$



Si bien buscamos (β,β_0) que maximicen $C(\beta,\beta_0)$, éstos no tienen por qué ser únicos.

SVM Caso 1) Linealmente separables

Como

$$\left\{ \boldsymbol{x} : \mathbb{R}^d : \langle \beta, \boldsymbol{x} \rangle + \beta_0 = 0 \right\} = \left\{ \boldsymbol{x} \in \mathbb{R}^d : \left\langle \frac{\beta}{C} \, , \, \boldsymbol{x} \right\rangle + \frac{\beta_0}{C} = 0 \right\}$$

Haciendo el cambio de variable

$$\tilde{\beta} = \frac{\beta}{C}, \quad \tilde{\beta_0} = \frac{\beta_0}{C}, \quad \|\tilde{\beta}\| = \frac{1}{C}.$$

El problema equivale a $P_1 = \begin{cases} \min\limits_{\tilde{\beta}, \tilde{\beta}_0} \|\tilde{\beta}\| \\ \text{sujeto a} \\ \quad y_i(\langle \boldsymbol{x}_i, \tilde{\beta} \rangle + \tilde{\beta}_0) \geq 1 \quad \forall i = 1, \ldots, n \\ \quad \tilde{\beta}_0 \in \mathbb{R}, \tilde{\beta} \in \mathbb{R}^d \end{cases}$

Observemos que el conjunto
$S = \{(\beta, \beta_0) \in \mathbb{R}^d \times \mathbb{R} : g_i(\beta, \beta_0) \leq 0\}$ con $g_i(\beta, \beta_0) = 1 - y_i(\langle \boldsymbol{x}_i, \beta \rangle + \beta_0)$ es convexo. Esto se debe a que las funciones g_i son convexas.
(es una función que verifica : $g\big(\alpha x + (1-\alpha)y\big) \leq \alpha g(x) + (1-\alpha)g(y) \;\; \forall\, \alpha \in [0,1]$).

SVM Caso 1) Linealmente separable

Abusando de la notación si $\tilde{\beta} = \beta$ y $\tilde{\beta}_0 = \beta_0$, el problema P_1 es equivalente al problema

$$P_2 = \begin{cases} \min\limits_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 \\ \text{sujeto a} \\ \quad y_i(\langle \boldsymbol{x}_i, \beta\rangle + \beta_0) \geq 1 \quad \forall i = 1, \ldots, n \\ \quad \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^d \end{cases}$$

P_2 es un problema de optimización convexa que se resuelve considerando primero el *problema relajado* (que depende de α) :

$$P_\alpha = \begin{cases} \min\limits_{\beta,\beta_0,\alpha} \mathcal{L}(\beta, \beta_0, \alpha) := \frac{1}{2}\|\beta\|^2 - \sum\limits_{i=1}^{n} \alpha_i \big(y_i(\langle \boldsymbol{x}_i, \beta\rangle + \beta_0) - 1\big) \\ \text{sujeto a} \\ \alpha = (\alpha_1, \ldots, \alpha_n) \geq 0 \end{cases}$$

Lo que se hace es resolver P_α en función de α y obtener así $\hat{\beta}(\alpha)$ y $\hat{\beta}_0(\alpha)$ y luego imponer condiciones (llamas condiciones de Karush-Kuhn-Tucker) para determinar el α que haga que $\hat{\beta}(\alpha)$ y $\hat{\beta}_0(\alpha)$ sean soluciones de P_2. Estas condiciones son necesarias y suficientes para encontrar el óptimo.

SVM Caso 1) Linealmente separable

Para resolver P_α, como $\mathcal{L}(\beta, \beta_0, \alpha)$ es convexa en β, β_0 basta resolver
$\nabla\left(\frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{n} \alpha_i\big(y_i(\langle \boldsymbol{x}_i, \beta\rangle + \beta_0) - 1\big)\right) = \mathbf{0}$ sujeto a $\alpha \geq 0$:

$$\frac{\partial \mathcal{L}(\beta, \beta_0, \alpha)}{\partial \beta} = \beta - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i = 0 \Rightarrow \beta = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \tag{1}$$

$$\frac{\partial \mathcal{L}(\beta, \beta_0, \alpha)}{\partial \beta_0} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{2}$$

Las condiciones de $K - T$ son

a) $y_i(\langle \boldsymbol{x}_i, \beta\rangle + \beta_0) - 1 \geq 0$ (los datos están bien clasificados)

b) $\alpha_i\big(y_i(\langle \boldsymbol{x}_i, \beta\rangle + \beta_0) - 1\big) = 0 \ \forall i$ ($\alpha_i \neq 0$ sólo en los *support vectors*)

c) $\alpha \geq 0$

Dado que el problema de encontrar β, β_0, α que verifiquen (1),(2) y a),b),c) es muy difícil lo que se hace es sustituir (1) y (2) en $\mathcal{L}(\beta, \beta_0, \alpha)$. Esto nos da un problema de optimización en α (denominado problema dual).

SVM Caso 1) Linealmente separable

Si sustituimos por (1) y (2) en $\mathcal{L}(\beta, \beta_0, \alpha)$ obtenemos

$$\Phi(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \tag{3}$$

El problema dual es

$$D = \begin{cases} \max\limits_{\alpha} \Phi(\alpha) \\ \text{sujeto a} \\ \quad \sum\limits_{i=1}^{n} \alpha_i y_i = 0 \\ \quad \alpha = (\alpha_1, \ldots, \alpha_n) \geq 0 \end{cases}$$

Observemos que el problema anterior sólo depende de los productos escalares entre las observaciones.

SVM Caso 1) Resolución del problema

Si $\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*)$ es la solución del problema, los vectores soporte son aquellos $\boldsymbol{x}_i$ tales que $\alpha_i^* > 0$.

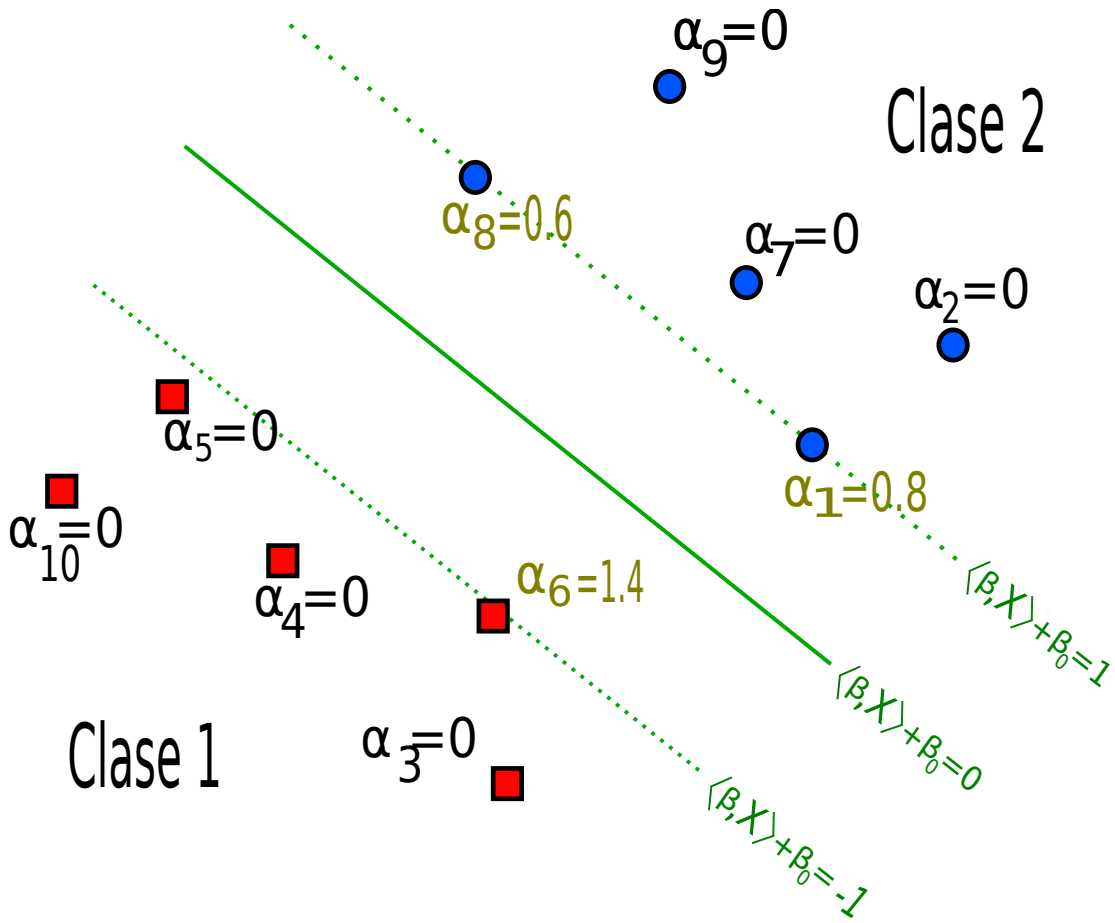


FIGURE – Los *support vectors* son $\boldsymbol{x}_1$, $\boldsymbol{x}_6$, $\boldsymbol{x}_8$

SVM Caso 1) Resolución del problema

Sea el conjunto de vectores soportes $\mathcal{SV} = \{i \in \{1, ..., n\} : y_i H(\boldsymbol{x}_i) = 1\}$ y $N_{\mathcal{SV}} = |\mathcal{SV}|$ la cantidad de vectores soportes. Entonces :

$$\beta^* = \sum_{i=1}^{n} \alpha_i^* y_i \boldsymbol{x}_i = \sum_{i=1}^{n_S} \alpha_i^* y_i \boldsymbol{x}_i.$$

Aún no se ha obtenido el valor de β_0^*. Si $(\boldsymbol{x}_{vs}, y_{vs}) \in \mathcal{SV}$ se tiene :

$$y_{vs} \left(\sum_{i \in \mathcal{SV}} \alpha_i y_i \langle \boldsymbol{x}_i, \boldsymbol{x}_{vs} \rangle + \beta_0 \right) = 1$$

Se multiplica por y_{vs} y se realiza la sumatoria :

$$\sum_{vs \in \mathcal{SV}} y_{vs}^2 \left(\sum_{i \in \mathcal{SV}} \alpha_i y_i \langle \boldsymbol{x}_i, \boldsymbol{x}_{vs} \rangle + \beta_0 \right) = \sum_{vs \in \mathcal{SV}} y_{vs}$$

$$\beta_0^* = \frac{1}{N_{\mathcal{SV}}} \sum_{vs \in \mathcal{SV}} \left(y_{vs} - \sum_{i \in \mathcal{SV}} \alpha_i y_i \langle \boldsymbol{x}_i, \boldsymbol{x}_{vs} \rangle \right)$$

Por lo tanto el clasificador final es :

$$\boxed{f(\boldsymbol{x}) = \text{signo}\left(H(\boldsymbol{x})\right) = \text{signo}\left(\sum_{i \in \mathcal{SV}} \alpha_i^* y_i \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + \beta_0^* \right)}$$

SVM Caso 1) Resolución del problema

Obsérvese que $\boldsymbol{\beta}^*$ y β_0^* dependen únicamente de pocas observaciones de $\mathcal{L}$, los vectores de soporte, y no de todas las observaciones consideradas inicialmente.

Como :$||\boldsymbol{\beta}^*||^2 = \langle \boldsymbol{\beta}^*, \boldsymbol{\beta}^* \rangle = \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = \sum_{j \in \mathcal{SV}} \alpha_j \, y_j \underbrace{\sum_{i \in \mathcal{SV}} \alpha_i y_i \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}_{1 - y_j \beta_0^*}$ pues

$y_j \left(\sum_{i \in \mathcal{SV}} \alpha_i y_i \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + \beta_0^* \right) = 1$, entonces :

$$||\boldsymbol{\beta}^*||^2 = \sum_{j \in \mathcal{SV}} \alpha_j \left(1 - y_j \beta_0^*\right) = \sum_{j \in \mathcal{SV}} \alpha_j - \beta_0^* \underbrace{\sum_{j \in \mathcal{SV}} \alpha_j y_j}_{0} = \sum_{j \in \mathcal{SV}} \alpha_j$$

y queda definido el margen a partir de los α_i asociados a los vectores de soporte :

$$\boxed{C^* = \frac{1}{||\boldsymbol{\beta}^*||} = \left(\frac{1}{\sum_{j \in \mathcal{SV}} \alpha_j} \right)^{1/2}}$$

SVM Caso 1) Conclusiones para el caso linealmente separable

- Encontrar el hiperplano óptimo que haga que el margen entre los datos sea máximo.
- Es un problema de optimización convexo.
- La solución solamente depende de los vectores de soporte : todos los demás datos pueden ser “olvidados”.
- La cantidad de vectores de soporte puede ser muy pequea en relación a la cantidad de datos.
- La solución depende únicamente de los productos internos entre las observaciones.

Plan

SVM Caso 2) Soft Margin

- Tolerancia en el margen (soft margin -margen blando-) :

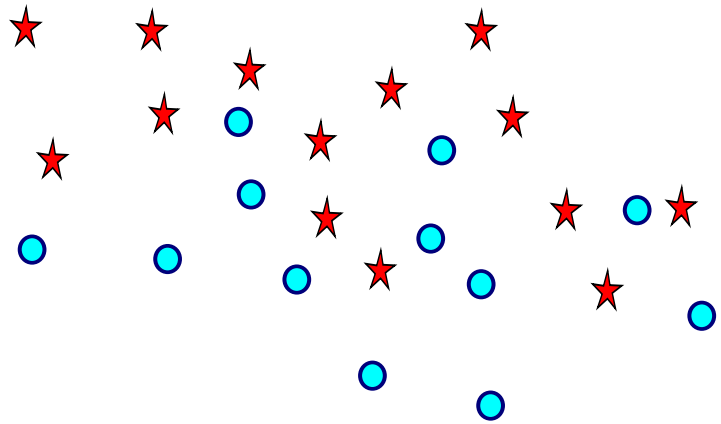


FIGURE – Soft Margin

- Caso no separable

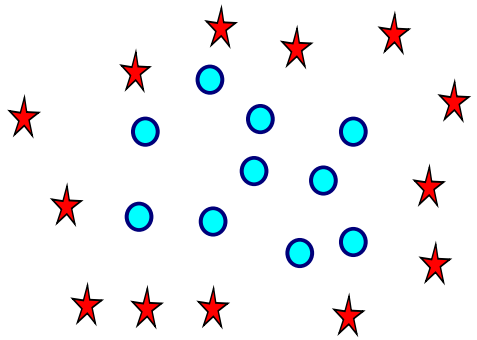


FIGURE – No separable

SVM Caso 2) Soft Margin

Idea

Para la mayoría de los datos hay margen, pero algunos cruzan la frontera.

De todos modos queremos encontrar un hiperplano “separador”. Introducimos variables de holgura (slacks) $\xi_i \geq 0$ $i = 1, \ldots, n$.

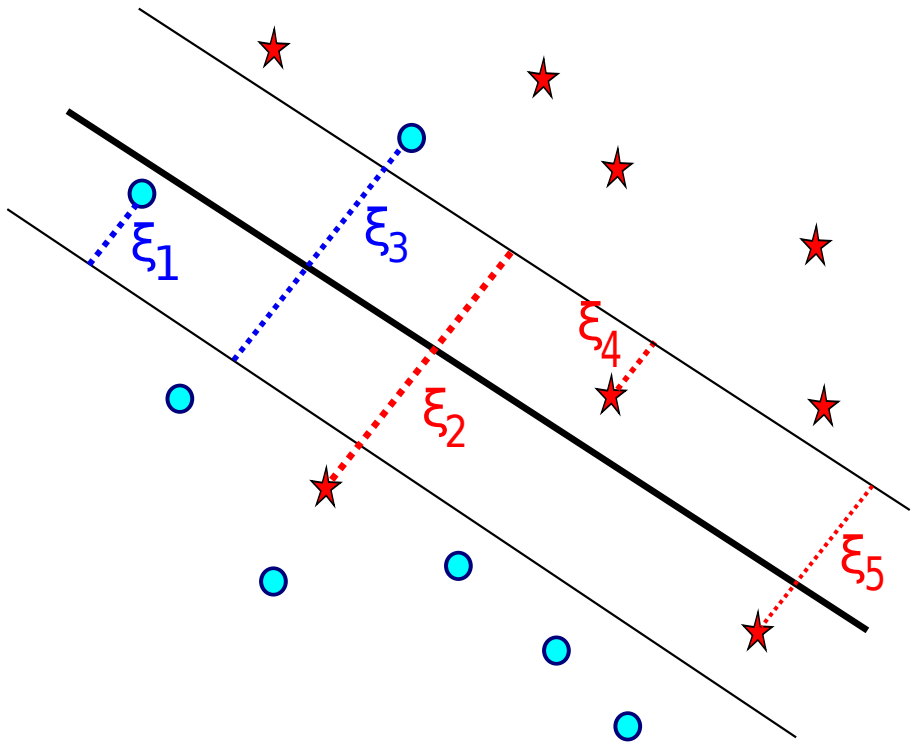


FIGURE – Si quitamos $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_5$ es linealmente separable

SVM Caso 2) Soft Margin

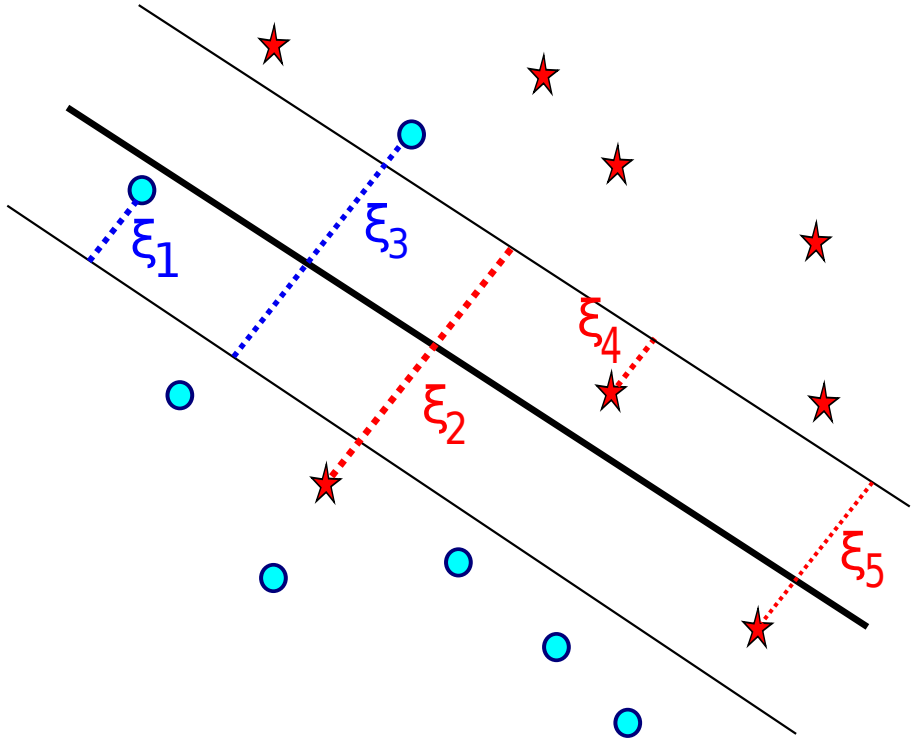


FIGURE – Si $\xi_i > 1$ el dato i está del otro lado del hiperplano separador $\langle \beta, \boldsymbol{x} \rangle + \beta_0 = 0$

Queremos que

$$y_i \left(\langle \beta, \boldsymbol{x}_i \rangle + \beta_0 \right) \geq 1 - \xi_i \quad \forall i$$

$$\xi_i \geq 0 \quad \forall i$$

SVM Caso 2) Soft Margin

Las variables de holgura son una medida de desviación con respecto a la condición inicial :

- Si $0 \leq \xi \leq 1$ el dato está del lado correcto del hiperplano pero en la región del margen.
- Si $\xi > 1$ el dato está del lado equivocado del hiperplano.

Introducimos en el problema un factor de penalización γ.

Si γ es grande estamos penalizando más los errores (permitimos pocos) y por lo tanto el margen es más chico, mientras que si γ es chico el margen es más grande (permitimos más errores).

El problema consiste ahora en resolver :

$$\begin{cases} \min\limits_{\beta,\beta_0,\xi} \ \dfrac{1}{2}\|\beta\|^2 + \gamma \sum\limits_{i=1}^{n} \xi_i \quad (*) \\ \text{sujeto a} \\ \quad y_i\left(\langle \beta, \boldsymbol{x}_i \rangle + \beta_0\right) \geq 1 - \xi_i \quad i = 1, \ldots, n \\ \quad \xi_i \geq 0 \quad i = 1, \ldots, n \end{cases}$$

Lo dicho anteriormente puede verse de forma geométrica. Si γ es variable de la función en $(*)$ tenemos una función lineal de γ y como $\sum_{i=1}^{n} \epsilon_i \geq 0$, dicha función es creciente.

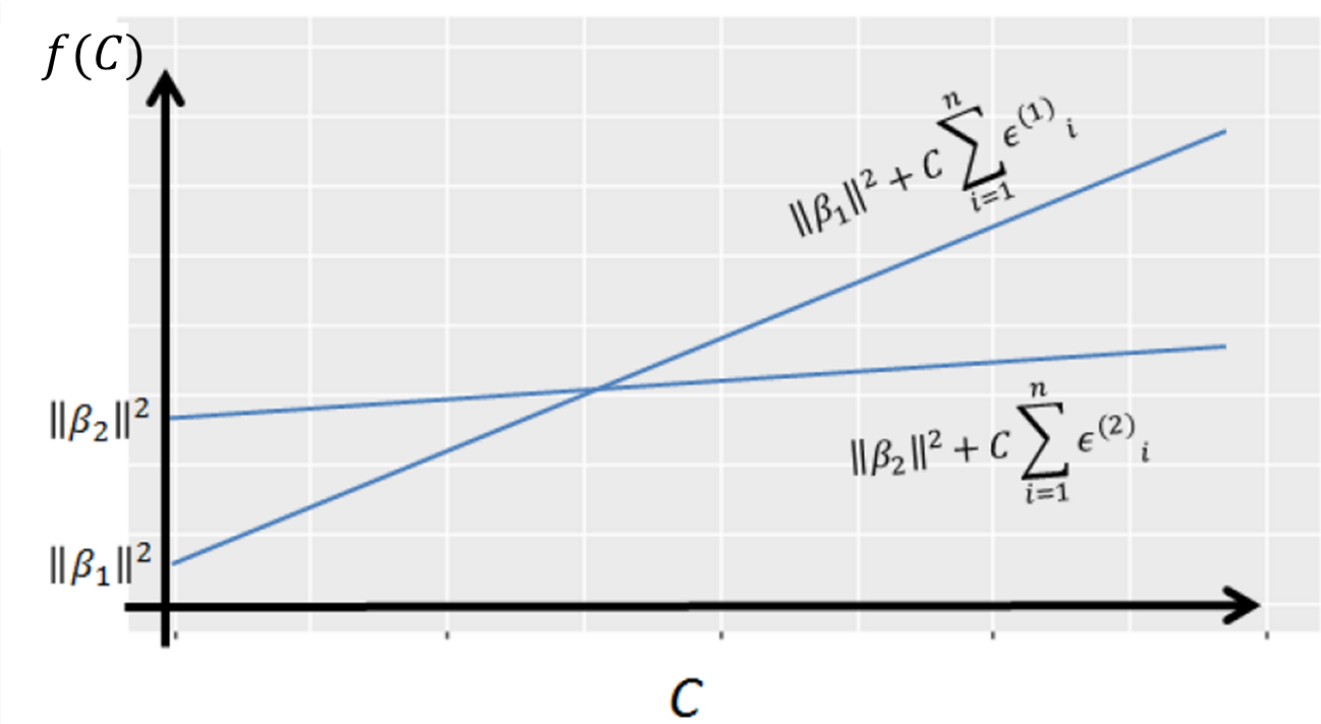


FIGURE – Relación entre el valor del costo y la amplitud del margen.

Si γ es grande se toma la recta 2, y en este caso $\sum_{i=1}^{n} \epsilon_i^{(2)} < \sum_{i=1}^{n} \epsilon_i^{(1)}$ (ya que la recta 2 tiene menor pendiente que la recta 1) y a su vez $\| \beta_2 \|^2 > \| \beta_1 \|^2$. Esto significa que el margen correspondiente a la recta 2 tendrá una menor amplitud que el margen correspondiente a la recta 1 ($\| \beta \|$ es el inverso del margen), y por ende se tolerarán menos errores conllevando a un posible sobreajuste. Sucede lo contrario si se toma un γ pequeño.

SVM Caso 2) Soft Margin

Si definimos

$$\mathcal{Q}(\beta, \beta_0, \xi, \alpha, \varphi) := \frac{1}{2}\|\beta\|^2 + \gamma \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \big(y_i(\langle \boldsymbol{x}_i, \beta\rangle + \beta_0) - 1 + \xi_i\big) - \sum_{i=1}^{n} \varphi_i \xi_i \tag{4}$$

El problema relajado es

$$P_{\alpha,\varphi} = \begin{cases} \min\limits_{\beta, \beta_0, \xi, \alpha, \varphi} \mathcal{Q}(\beta, \beta_0, \xi, \alpha, \varphi) \\ \text{sujeto a} \\ \quad \alpha, \varphi, \xi \geq 0 \end{cases}$$

Las condiciones KKT son

a) $\alpha_i \big(y_i(\langle \beta, \boldsymbol{x}_i\rangle + \beta_0) - 1 + \xi_i\big) = 0 \quad i = 1, \ldots, n$

b) $\varphi_i \xi_i = 0 \quad i = 1, \ldots, n$

c) $\alpha_i \geq 0, \quad \varphi_i \geq 0, \quad \xi_i \geq 0 \quad i = 1, \ldots, n$

SVM Caso 2) Soft Margin

Para encontrar el problema dual planteamos

$$\frac{\partial \mathcal{Q}}{\partial \beta} = 0 \quad \frac{\partial \mathcal{Q}}{\partial \beta_0} = 0 \quad \frac{\partial \mathcal{Q}}{\partial \xi} = 0$$

y obtenemos

$$\beta = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \qquad \sum_{i=1}^{n} \alpha_i y_i = 0 \qquad \alpha_i + \varphi_i = \gamma \quad \forall i = 1, \ldots, n \tag{5}$$

Sustituimos (4) en (3) y obtenemos el problema dual :

$$D = \begin{cases} \max\limits_{\alpha} \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\limits_{i,j} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \\ \text{sujeto a} \\ \quad \sum\limits_{i=1}^{n} y_i \alpha_i = 0 \\ \quad 0 \leq \alpha_i \leq \gamma \quad i = 1, \ldots, n \end{cases}$$

Plan

SVM Caso 3) No separable- Núcleos

Idea

Enviar a través de una función Φ (no necesariamente lineal) los datos $\boldsymbol{x}_i \in \mathbb{R}^d$ a un espacio de dimensión mayor, posiblemente infinita (espacio de característica, *feature space*) donde los datos son linealmente separables o con un poco de ruido.

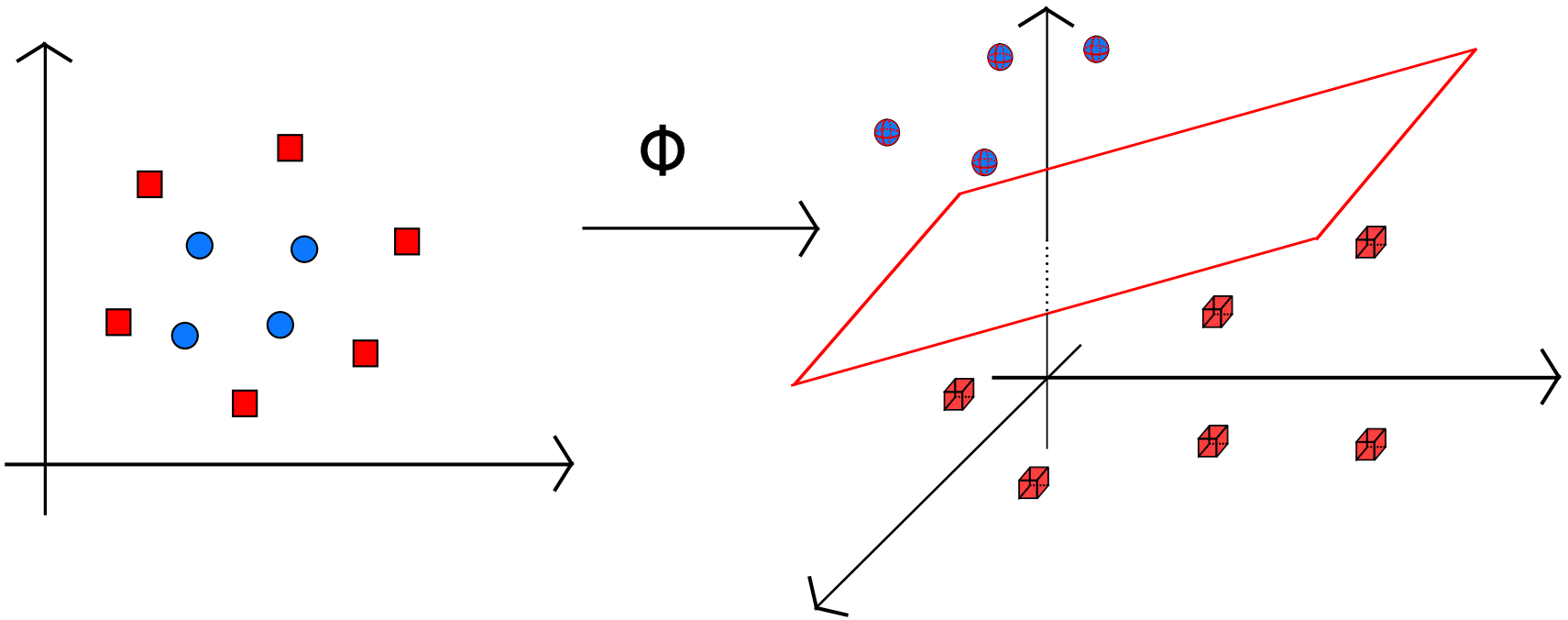


FIGURE – Los datos en $\mathbb{R}^2$ no son linealmente separables pero podemos tomar $\Phi : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$ de modo que en $\mathbb{R}^3$ sean linealmente separables, (existe un subespacio que los separa)

SVM Caso 3) No separable- Núcleos

- https ://www.youtube.com/watch ?v=3liCbRZPrZA
- Es importante observar que al resolver el problema de optimización que planteamos anteriormente, sólo intervienen los productos escalares entre los datos para encontrar β y β_0.
- Definimos $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle$
- Trabajar en el espacio de característica se reduce al caso lineal sustituyendo los $\langle , \rangle$ por $k(,)$.
- Lo que haremos será dar una función simétrica $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ definida positiva :

$$\forall n \quad \sum_{i,j=1}^{n} h_i h_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0 \quad \forall \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, h \in \mathbb{R}^d$$

en lugar del mapa Φ, veamos que efectivamente existe tal Φ.

SVM Caso 3) No separable- Núcleos

Theorem 1

*(**Teorema de Mercer**) Dada* $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *definida positiva entonces existe un espacio de Hilbert* $(H, \langle , \rangle_H)$ *y una función* $\Phi : \mathbb{R}^d \to H$ *tal que*

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle_H \quad \forall \boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^d$$

Ejemplos de Núcleos

- Lineal : $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle$
- Polinomial : $k(\boldsymbol{x}, \boldsymbol{x}') = \left(c_1 + c_2 \langle \boldsymbol{x}, \boldsymbol{x}' \rangle\right)^d$
- Gaussiano (radial) : $k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\sigma \|\boldsymbol{x} - \boldsymbol{x}'\|^2\right)$
- Laplace (radial) : $k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\sigma \|\boldsymbol{x} - \boldsymbol{x}'\|\right)$
- ANOVA (radial) $k(\boldsymbol{x}, \boldsymbol{x}') = \left(\sum_{k=1}^{d} \exp\left(-\sigma (x_k - x_k^{'})^2\right)\right)^d$
- Otros : Bessel, Splines.

Ejemplo

Sean los siguientes puntos en $\mathbb{R}$, $A = x_1 = 1$, $B = x_2 = 2$, $C = x_3 = 4$. Claramente este conjunto no es separable por un punto. Se considera la función $\Phi : \mathbb{R} \to \mathbb{R}^3$ tal que $\Phi(x) = (x^2, \sqrt{2}x, 1)$.

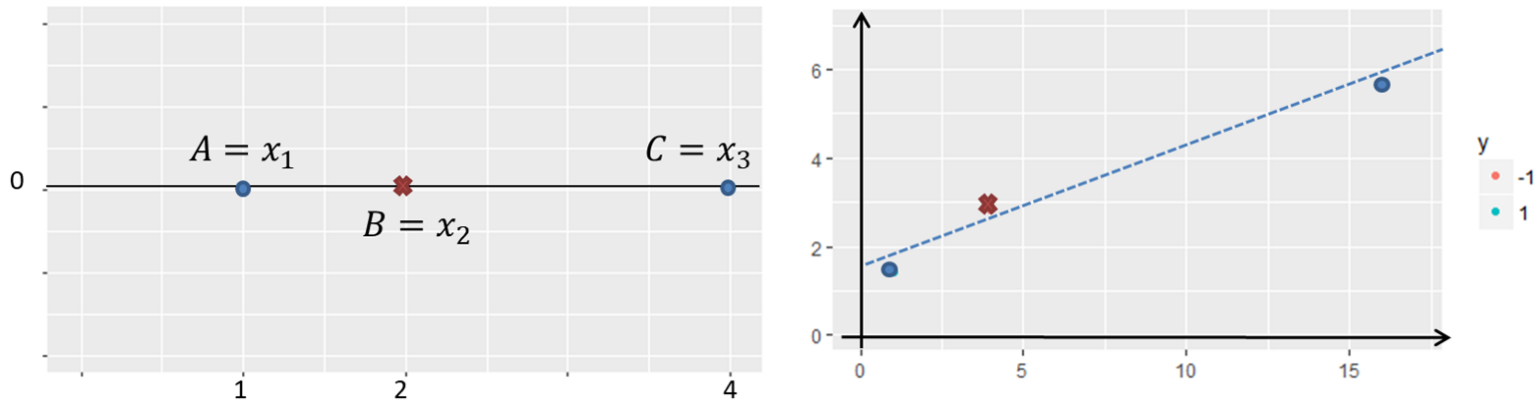


FIGURE – A la izquierda ejemplo de observaciones en $\mathbb{R}$ no separables. A la derecha los mismos puntos luego de mapearlos en $\mathbb{R}^3$, siendo separables en esta nueva dimensión.

Ejemplo

Evaluando los puntos en la función Φ :

$$\begin{cases} \Phi(x_1) = (1, \sqrt{2}, 1) \\ \Phi(x_2) = (4, 2\sqrt{2}, 1) \\ \Phi(x_3) = (16, 4\sqrt{2}, 1) \end{cases}$$

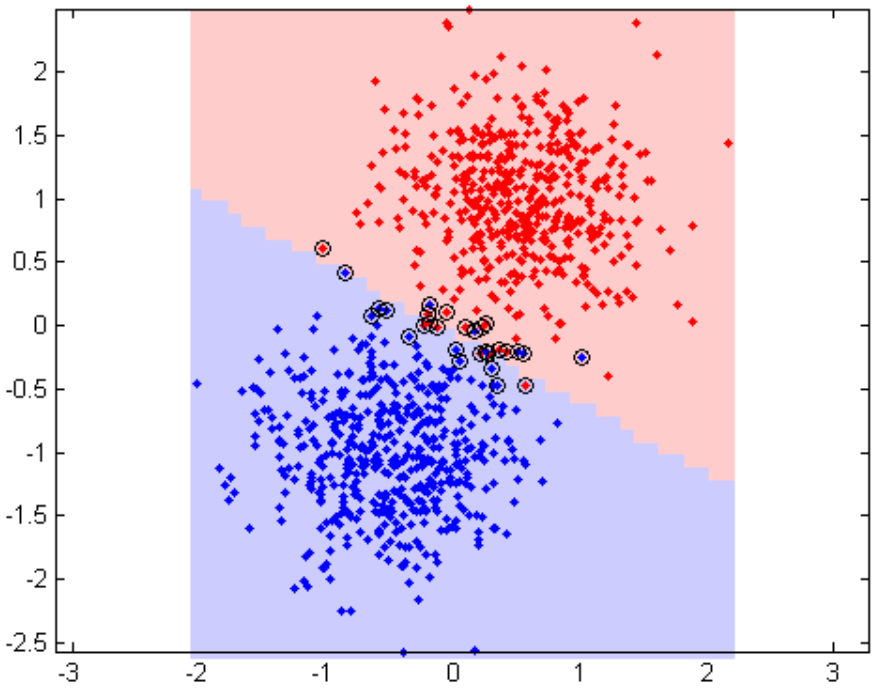
El producto interno quedará de la forma :

$$\langle \Phi(x_i), \Phi(x_j) \rangle = x_i^2 x_j^2 + \sqrt{2} x_i \sqrt{2} x_j = x_i^2 x_j^2 + 2 x_i x_j + 1 = (x_i x_j + 1)^2 = k(x_i, x_j)$$

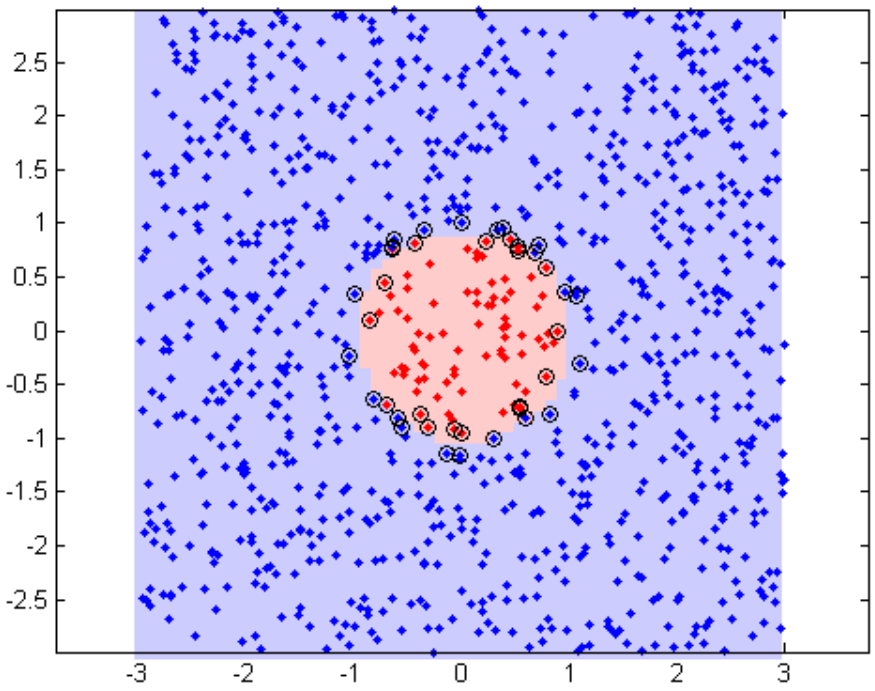
siendo k el kernel polinomial de grado 2. De esta forma se logra separar las observaciones que en el espacio original no eran separables mapeando las mismas en un espacio de dimensión mayor.

SVM Caso 3) No separable- Núcleos

- El Kernel Gaussiano y de Laplace se usan generalmente cuando no se tiene información a priori de los datos.
- El Kernel lineal se usa para *large sparse data*, ejemplo analisis de textos, cada palabra es un dato.
- Los Kernels polinomiales son usados en procesamiento de imagenes. ANOVA y Splines se desempean bien en problemas de regressión.
- Validación cruzada para elegir el kernel : Tomar subconjuntos de $n - p$ datos para construir el modelo y evaluarlo en los p restantes.



Kernel Lineal



Kernel Radial

SVM Caso 3) No separable- Núcleos

De forma análoga a como hicimos en el caso lineal, el problema dual es

$$D = \begin{cases} \max\limits_{\alpha} \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\limits_{i,j} \alpha_i \alpha_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ \text{sujeto a} \\ \quad \sum_{i=1}^{n} y_i \alpha_i = 0 \\ \quad 0 \leq \alpha_i \leq \gamma \quad i = 1, \ldots, n \end{cases}$$

y las condiciones de *KKT* son

a) $\alpha_i \left\{ y_i \left(\sum_{i=1}^{n} y_j \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \beta_0 \right) - 1 - \xi_i \right\} = 0 \quad i = 1, \ldots, n$

b) $(\gamma - \alpha_i)\xi_i = 0 \quad i = 1, \ldots, n$

c) $\alpha_i \geq 0, \quad \xi_i \geq 0 \quad i = 1, \ldots, n$

Clasificador

$$f(\boldsymbol{x}_{nue}) = sgn \left(\sum_{i=1}^{n} \alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}_{nue}) + \beta_0 \right)$$

SVM Caso 3) No separable- Núcleos

Ejemplo : $\{(x_i, y_i)\}_{i=1,\ldots,5} = \{(1,1),(2,1),(4,-1),(5,-1),(6,1)\}$

- $k(x, x') = (1 + xx')^2$
- Factor de penalización : $\gamma = 100$.

Resolviendo el problema dual se encuentran

$$\alpha_1 = 0, \quad \alpha_2 = 2.5, \quad \alpha_3 = 0, \quad \alpha_4 = 7.333, \quad \alpha_5 = 4.833$$

por lo tanto los vectores soporte son x_2, x_4 y x_5. La función h es

$$h(x) = 2.5 \times 1 \times (1 + 2x)^2 + 7.333 \times -1 \times (1 + 5x)^2 + 4.833 \times 1 \times (1 + 6x)^2 + \beta_0$$

Como x_2 es un vector soporte debe verificarse que $h(2) = 1$ (análogamente $h(5) = -1$) obtenemos $\beta_0 = 9$.

Clasificador

$$f(x) = sgn\big(0.666x^2 - 5.333x + 9\big)$$

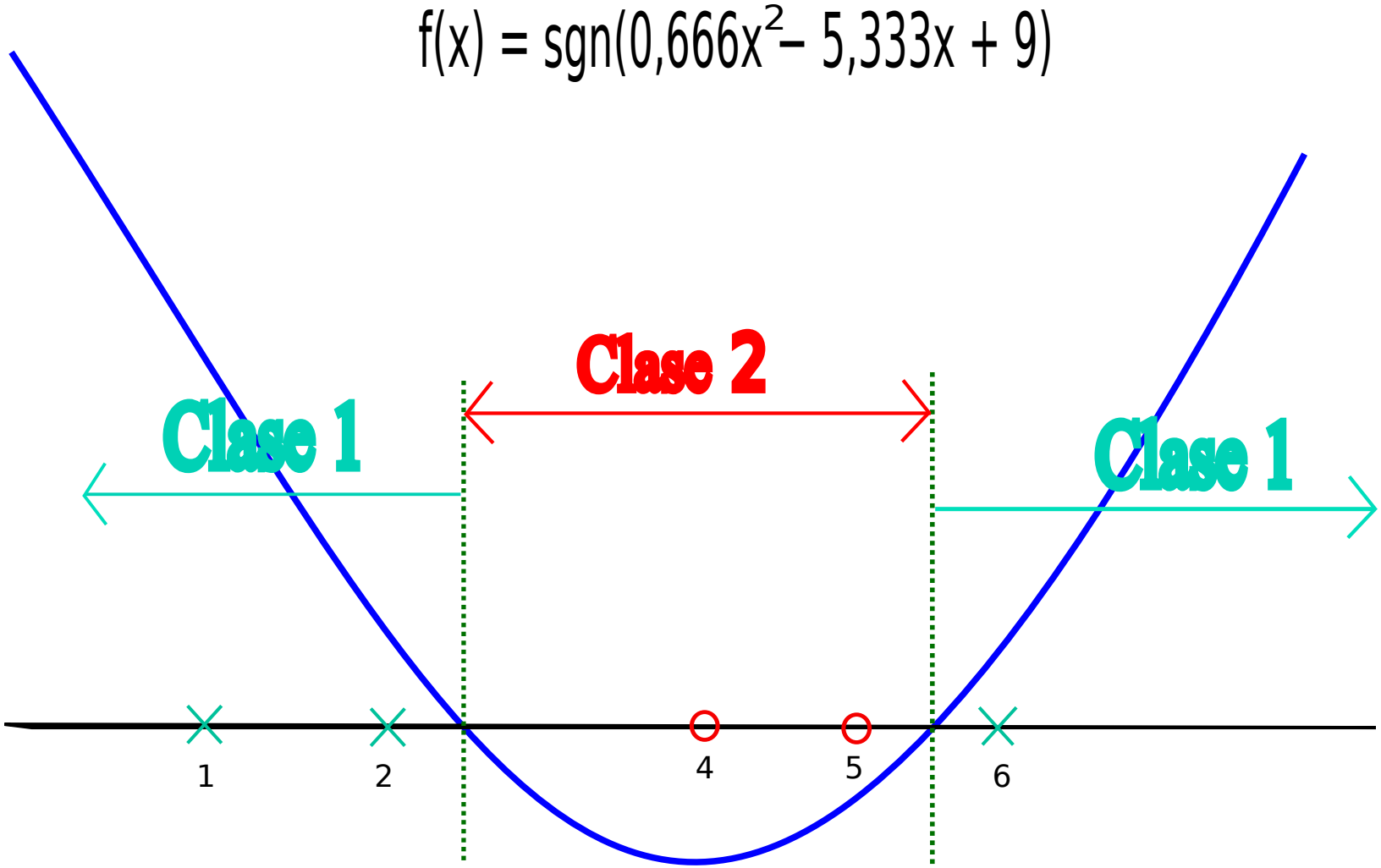


FIGURE – Los datos en $\mathbb{R}$ se separan considerándolos en $\mathbb{R}^2$

Escenario	Problema primal a optimizar
Linealmente separable	$\begin{cases} \min\limits_{\boldsymbol{\beta}} \quad \frac{1}{2} \parallel \boldsymbol{\beta} \parallel^2 \\ \text{s.a} \quad y_i H(\boldsymbol{x}_i) - 1 \geq 0 \end{cases}$
Cuasi-separable	$\begin{cases} \min\limits_{\boldsymbol{\beta},\epsilon} \quad \frac{1}{2} \parallel \boldsymbol{\beta} \parallel^2 + C \sum\limits_{i=1}^{n} \epsilon_i \\ \text{s.a} \quad y_i H(\boldsymbol{x}_i) + \epsilon_i - 1 \geq 0 \\ \qquad \epsilon_i \geq 0 \quad i = 1, 2, ..., n \end{cases}$
No separable	$\begin{cases} \min\limits_{\boldsymbol{\beta},\epsilon} \quad \frac{1}{2} \parallel \boldsymbol{\beta} \parallel^2 + C \sum\limits_{i=1}^{n} \epsilon_i \\ \text{s.a} \quad y_i H(\boldsymbol{x}_i) + \epsilon_i - 1 \geq 0 \\ \qquad \epsilon_i \geq 0 \quad i = 1, 2, ..., n \end{cases}$

TABLE – Comparación de los problemas primales para los distintos escenarios.

Escenario	Problema dual a optimizar
Linealmente separable	$\begin{cases} \max\limits_{\boldsymbol{\alpha}} \quad \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \\ \text{s.a} \quad \sum\limits_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$
Cuasi-separable	$\begin{cases} \text{máx} \quad \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \\ \text{s.a} \quad \sum\limits_{i=1}^{n} \alpha_i y_i = 0 \\ \qquad 0 \leq \alpha_i \leq C \end{cases}$
No separable	$\begin{cases} \text{máx} \quad \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ \text{s.a} \quad \sum\limits_{i=1}^{n} \alpha_i y_i = 0 \\ \qquad 0 \leq \alpha_i \leq C \end{cases}$

TABLE – Comparación de los problemas duales para los distintos escenarios.

Escenario	Clasificador
Linealmente separable	$$\text{signo}\left(H(\boldsymbol{x})\right) = \text{signo}\left(\sum_{i\in\mathcal{SV}} \alpha_i^* y_i \langle \boldsymbol{x}_i, \boldsymbol{x}\rangle + \beta_0^*\right)$$
Cuasi-separable	$$\text{signo}\left(H(\boldsymbol{x})\right) = \text{signo}\left(\sum_{i\in\mathcal{SV}} \alpha_i^* y_i \langle \boldsymbol{x}_i, \boldsymbol{x}\rangle + \beta_0^*\right)$$
No separable	$$\text{signo}\left(H(\boldsymbol{x})\right) = \text{signo}\left(\sum_{i\in\mathcal{SV}} \alpha_i^* y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + \beta_0^*\right)$$

TABLE – Comparación de los clasificadores para los distintos escenarios.

Escenario	Vectores de soporte
Linealmente separable	Observaciones de cada clase más cercanas al hiperplano separador. Aquellas que tienen asociado un $\alpha_i > 0$.
Cuasi-separable	Observaciones que tienen asociado un valor de $\alpha_i > 0$ y se encuentran : • sobre los hiperplanos canónicos ($0 < \alpha_i < C$ y $\epsilon_i = 0$), • dentro del margen, estén bien o mal clasificadas ($0 < \epsilon_i < 2$ y $\alpha_i = C$), • fuera del margen y mal clasificadas ($\epsilon_i \geq 2$ y $\alpha_i = C$).
No separable	Observaciones que tienen asociado un valor de $\alpha_i > 0$ y se encuentran : • sobre los hiperplanos canónicos ($0 < \alpha_i < C$ y $\epsilon_i = 0$), • dentro del margen, estén bien o mal clasificadas ($0 < \epsilon_i < 2$ y $\alpha_i = C$), • fuera del margen y mal clasificadas ($\epsilon_i \geq 2$ y $\alpha_i = C$).

TABLE – Comparación de las condiciones necesarias para las observaciones que son vectores soporte en los distintos escenarios.

Plan

SVM multiclass : una contra todas

Se construyen K clasificadores binarios, uno para cada clase.

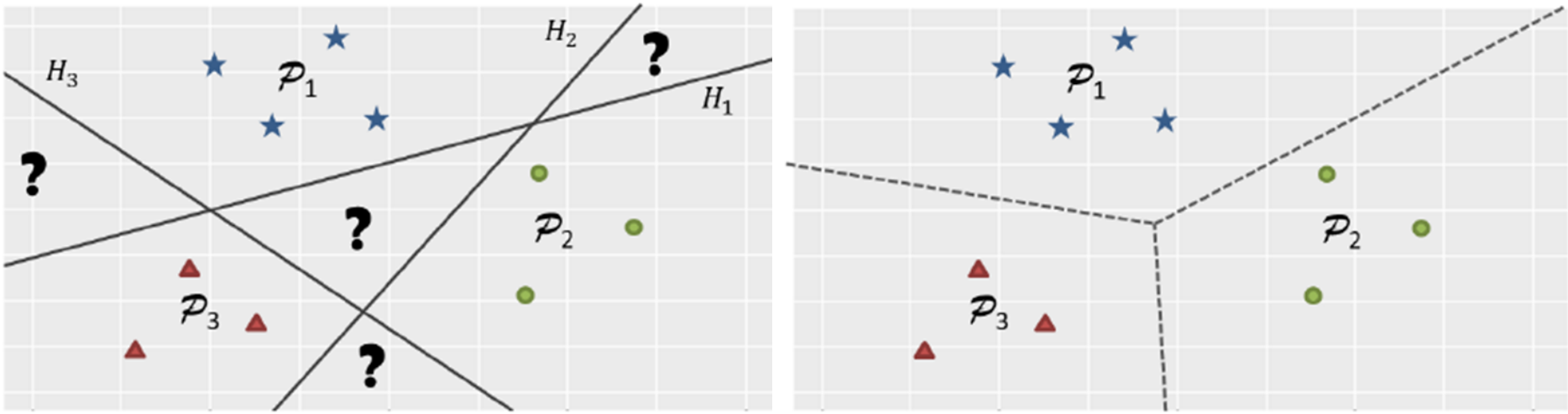


FIGURE – Descripción gráfica de la estrategia *one vs all*. En la izquierda se muestra un ejemplo con 3 subpoblaciones $\mathcal{P}_1$,$\mathcal{P}_2$, y $\mathcal{P}_3$, junto con los 3 hiperplanos separadores. Se marca las zonas de indeterminación con un signo de pregunta. A la derecha se muestra como queda la frontera de decisión al tomar el criterio de *winner takes all*.

Una nueva observación $\boldsymbol{x}_0$ se clasificará en una de las K poblaciones de la manera siguiente : si $\mathbb{P}^{(g_k)}(\boldsymbol{x}_0)$ denota la probabilidad a posteriori de que $\boldsymbol{x}_0$ pertenezca a la población $\mathcal{P}_k$, se asignará a $\boldsymbol{x}_0$ la clase k para la cual $\mathbb{P}^{(g_k)}(\boldsymbol{x}_0)$ es máxima.

SVM multiclass : una contra una

Un clasificador binario g_k para cada par de clases, en total $\binom{K}{2}$.

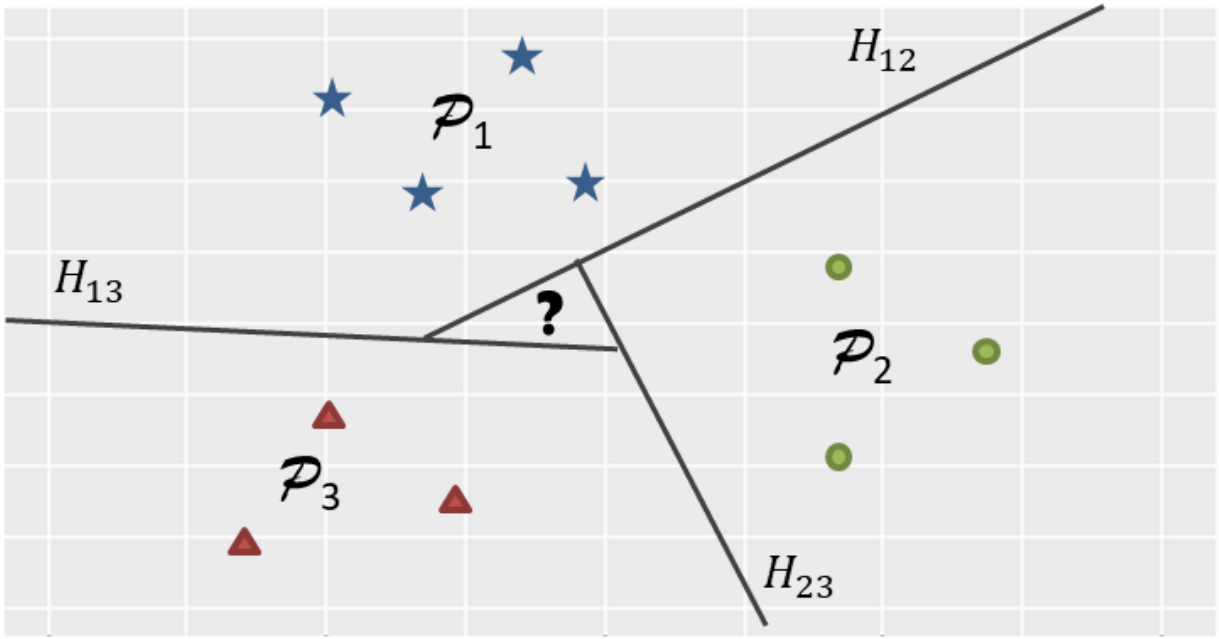


FIGURE – Descripción gráfica de la estrategia *one versus one*. Se muestran las 3 subpoblaciones $\mathcal{P}_1$, $\mathcal{P}_2$, y $\mathcal{P}_3$, junto con los hiperplanos separadores H_{ij} donde *ij* indica las poblaciones que separa dicho hiperplano. Una nueva observación $\boldsymbol{x}_0$ se etiqueta según la clasificación más frecuente.

Una nueva observación $\boldsymbol{x}_0$ se clasifica utilizando cada uno de los clasificadores binarios g_k construidos y se asigna $\boldsymbol{x}_0$ a aquella clase más frecuente. En caso de empate se asigna la clase de forma aleatoria.
Es la estrategia que usa el paquete e1071 para SVM.

One class SVM

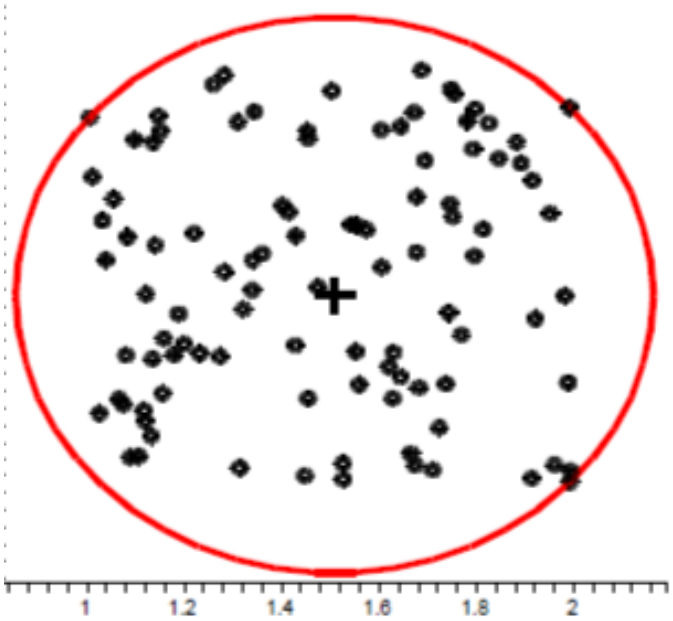
Primer encare : Tax and Duin (Support Vector Data Description, SVDD)

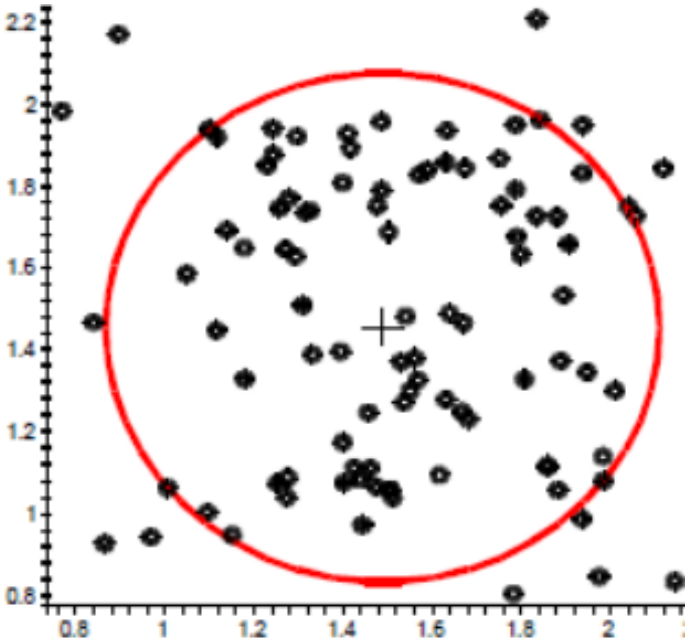
Encerrar los datos en una hiperesfera y classificar nuevos datos como normales si cae dentro de la hiperesfera o anomalo si cae afuera. También podemos suponer tolerancia al ruido.

Buscar una hiperesfera de centro a (combinación lineal de vectores de soporte) y radio R que resuelvan el problema :

$$(P_1)\begin{cases} \min\limits_{R,\boldsymbol{a}} R^2 + C\sum\limits_{i=1}^{n}\xi_i \\ \text{sujeto a} \\ \quad ||\Phi(\boldsymbol{x}_i) - \boldsymbol{a}||^2 \leq R^2 + \xi_i & \forall i = 1,\ldots,n \\ \quad \xi_i \geq 0 & \forall i = 1,\ldots,n \end{cases}$$

One class SVM





One class ν-SVM

Segundo encare : Scholkopf.

Separar en el espacio de características los puntos del origen de manera máxima pero permitir que νn de estos puntos estén entre el origen y el hiperplano.

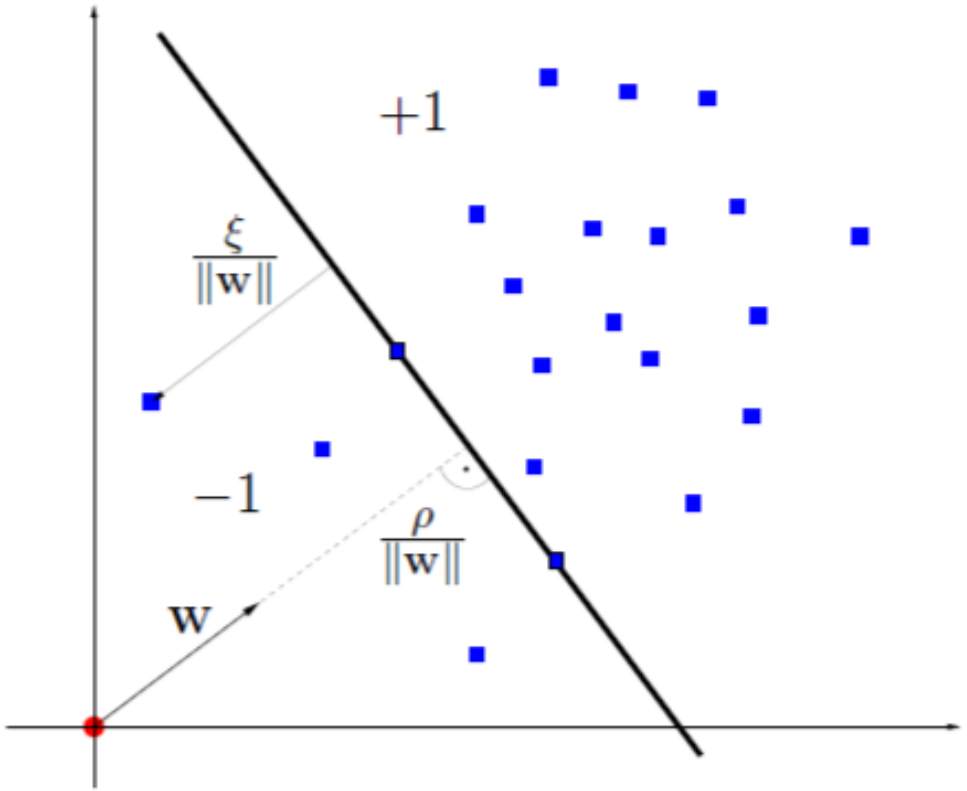
Busco un hiperplano de manera a separar los datos del origen (buscando el margen maximal).

$$(P_2)\begin{cases} \min\limits_{w,\xi_i,\rho} \frac{1}{2}||w||^2 + \dfrac{1}{\nu n}\sum\limits_{i=1}^{n}\xi_i - \rho \\ \text{sujeto a} \\ \quad w'\Phi(\boldsymbol{x}_i) \geq \rho - \xi_i & \forall i = 1,\ldots,n \\ \quad \xi_i \geq 0 & \forall i = 1,\ldots,n \end{cases}$$

El clasificador final es

$$f(\boldsymbol{x}) = sgn(w'\Phi(\boldsymbol{x}) - \rho) = sgn\left(\sum_{i=1}^{n}\alpha_i k(\boldsymbol{x},\boldsymbol{x}_i) - \rho\right)$$

One class SVM



One class ν-SVM

El parametro $\nu \in (0, 1)$, elegido por el usuario, juega el mismo papel que C, e impone la fracción de datos que le vamos a dar tolerancia a estar mal clasificados. Está probado que es una cota superior del error de aprendizaje y es una cota inferior de la fracción de vectores de soporte. El problema (P_2) es equivalente a

$$(P_2) \begin{cases} \max\limits_{\xi_i, \rho} \rho - \dfrac{1}{\nu n} \sum\limits_{i=1}^{n} \xi_i \\ \text{sujeto a} \\ \quad w' \phi(\boldsymbol{x}_i) \geq \rho - \xi_i & \forall i = 1, \ldots, n \\ \quad \xi_i \geq 0 & \forall i = 1, \ldots, n \\ \quad ||w_i|| = 1 & \forall i = 1, \ldots, n \end{cases}$$

Minimización Riesgo Estructural

A diferencia de la mayoría de los métodos de clasificación que buscan minimizar el error empírico, *SVM* busca minimizar el riesgo estructural (*SRM*). Con probabilidad $1 - \eta$, Vapnik y Chervonenkis prueban :

$$\text{Error de generalización} \leq \text{Error empírico} + f(n, h, \eta) \tag{6}$$

donde $f(n, h, \eta)$ es una función que depende de la cantidad de observaciones n, la complejidad de la clase de funciones discriminantes h (es la dimensión de Vapnik Chervonenkis) y de η. El principio *SRM* sugiere minimizar el lado derecho de la desigualdad (6) respecto de h en lugar del error empírico exclusivamente.

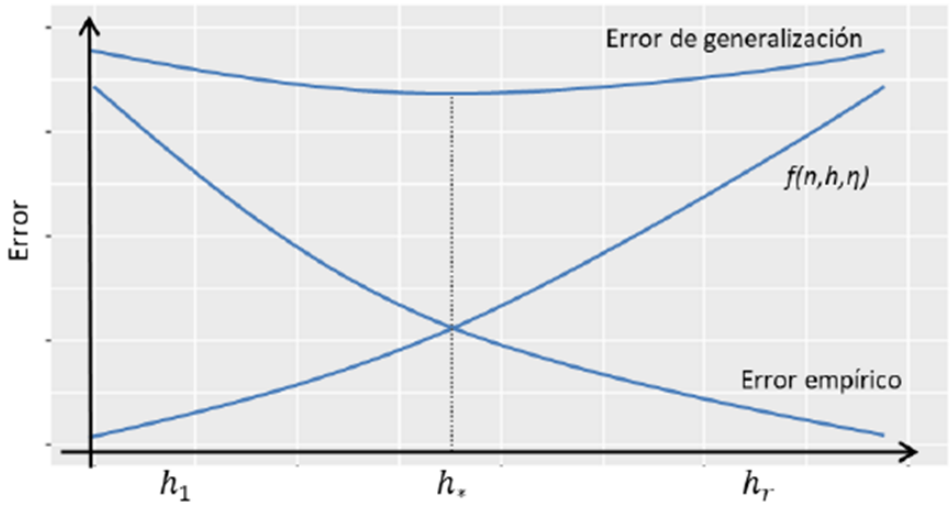


FIGURE – Minimización del riesgo estructural (*SRM*) : representación gráfica de la evolución de los errores al aumentar la complejidad (h) (es la dimensión de Vapnik Chervonenkis) del modelo. Se observa que en h_* se encuentra el óptimo.

Plan

SVM y probabilidades

Nos basamos en el hecho que la función de discriminación ideal

$$\ln\left(\frac{\mathbb{P}(Y=1|X=\boldsymbol{x})}{\mathbb{P}(Y=-1|X=\boldsymbol{x})}\right)$$

debería tener igual signo que la función de discriminación $H(\boldsymbol{x})+b$. Escribimos :

SVM y probabilidades

Nos basamos en el hecho que la función de discriminación ideal

$$\ln\left(\frac{\mathbb{P}(Y=1|X=\boldsymbol{x})}{\mathbb{P}(Y=-1|X=\boldsymbol{x})}\right)$$

debería tener igual signo que la función de discriminación $H(\boldsymbol{x})+b$. Escribimos :

$$\ln\left(\frac{\mathbb{P}(Y=1|X=\boldsymbol{x})}{\mathbb{P}(Y=-1|X=\boldsymbol{x})}\right)=AH(\boldsymbol{x})+B$$

y obtenemos una estimación de la probabilidad a posteriori

$$\mathbb{P}(Y=1|X=\boldsymbol{x})=1-\frac{1}{1+exp(AH(\boldsymbol{x})+B)}$$

donde A y B están estimados por el método de máxima verosimilitud.

Plan

SVM y Regresión (SVR)

En *SVR* se busca seleccionar el "hiperplano" regresor que mejor se ajuste al conjunto de datos de entrenamiento. En este caso no se busca clasificar a una observación en alguna población, si no encontrar un "hiperplano" que minimice la distancia global entre $H(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle + \beta_0$ y los $y_1, y_2, ..., y_n$. Se quiere encontrar $\boldsymbol{\beta}^*$ y β_0^* tales que :

$$(\boldsymbol{\beta}, \beta_0) = \underset{\boldsymbol{\beta}, \beta_0}{argmin} \sum_{i=1}^{n} \left| y_i - \left\langle \boldsymbol{\beta}, \Phi(\boldsymbol{x}_i) \right\rangle + \beta_0 \right|$$

con Φ una función kernel. Para ello se utiliza una estrategia parecida al problema de clasificación utilizando las funciones kernel para encontrar un "hiperplano" en un "hiper-tubo" de radio ρ que contenga a las observaciones de $\mathcal{L}$.

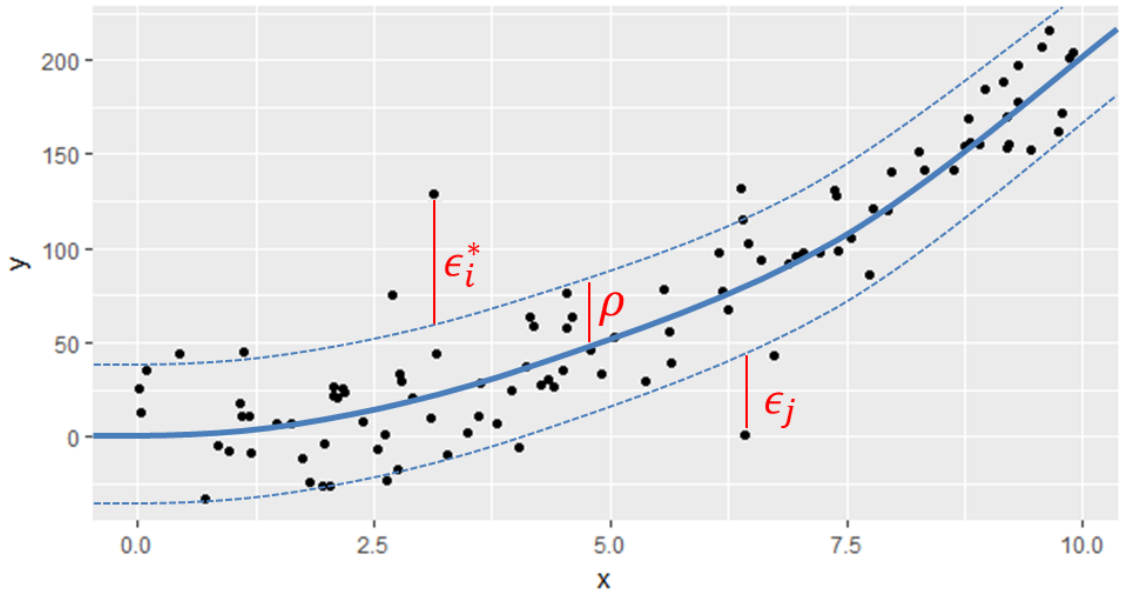


FIGURE – Hiperplano hallado para *SVR*. Se muestra el radio ρ del "hiper-tubo" y las variables de holgura de dos observaciones (ϵ_i^* y ϵ_j).

A diferencia que las subsecciones anteriores, se definen dos variables de holgura ϵ_i y ϵ_i^*, que miden la distancia de una observación que está fuera del "hiper-tubo" respecto al hiperplano canónico más cercano. El problema de optimización es el siguiente :

$$\begin{cases} \text{mín} & \frac{1}{2} \parallel \boldsymbol{\beta} \parallel^2 + C \sum_{i=1}^{n} \left(\epsilon_i + \epsilon_i^*\right) \\ \\ \text{s.a} & y_i - H(\boldsymbol{x}_i) \leq \rho + \epsilon_i^* \\ & H(\boldsymbol{x}_i) - y_i \leq \rho + \epsilon_i \\ & \epsilon_i, \epsilon_i^* \geq 0 \quad i = 1, 2, ..., n \end{cases} \tag{7}$$

Para obtener el "hiperplano" solo se consideraran las observaciones denominadas vectores de soporte, en este caso son aquellas que se encuentran a distancia menor que ρ del hiperplano separador.

Plan

Ejemplo en R

Las librerias mas comunes en R para svm son `e1071` y `kernlab`.

```
 library(e1071)
>names(bcdata)
```

```
 [1]  Samplecodenumber      ClumpThickness          UniformityofCellSize    UniformityofCellShape
 [5]  MarginalAdhesion      SingleEpithelialCellSize BareNuclei             BlandChromatin
 [9]  NormalNucleoli        Mitoses                 Class
```

```
databcall <- subset(bcdata,select=c(-Samplecodenumber,-Class))
classesbcall <- subset(bcdata,select=Class)
databctrain <- databcall[1:400,]
classesbctrain <- classesbcall[1:400,]
databctest <- databcall[401:699,]
classesbctest <- classesbcall[401:699,]

model<-svm(databctrain,classesbctrain,kernel="polynomial",degree=3,cost=10)

pred <- predict(model, databctest)
```

Ejemplo en R

```
>table(pred,t(classesbctest))
```

	benign	malignant
benign	227	4
malignant	2	66

```
>summary(model)

Parameters:

SVM-Type: C-classification
SVM-Kernel: polynomial
cost: 10
degree: 3
gamma: 0.1111111
coef.0: 0
Number of Support Vectors: 66
( 30 36 )
Number of Classes: 2
Levels:

benign malignant
```


Carmona, E., “**Tutorial sobre Máquinas de Vectores de Soporte (SVM)**", Universidad Nacional de Educación a Distancia (UNED), Madrid España, Julio 2014.


Cavallero, M., Paolillo, G, “**Support Vector Machines y comparación con otras técnicas de clasificación supervisada**”, Monografía de grado, Licenciatura en Estadística, Udelar, 2018.


Hastie, T., Tibshirani, R., Friedman, J.,“**The Elements of Statistical Learning, Data Mining, Inference and Prediction**", Springer, 2008.


James, G., Witten, D., Hastie, T., Tibshirani, R., “**An Introduction to Statistical Learning with applications in R**", Springer, 2013.


Platt, John C., “**Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods**", Microsoft Research, Marzo 1999.


Tax, D. Duin, R. “**Support Vector Data Description**”, Kluwer Academic Publishers, 2004.