Support Vector Machines

Mathias Bourel

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

April 25, 2019

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

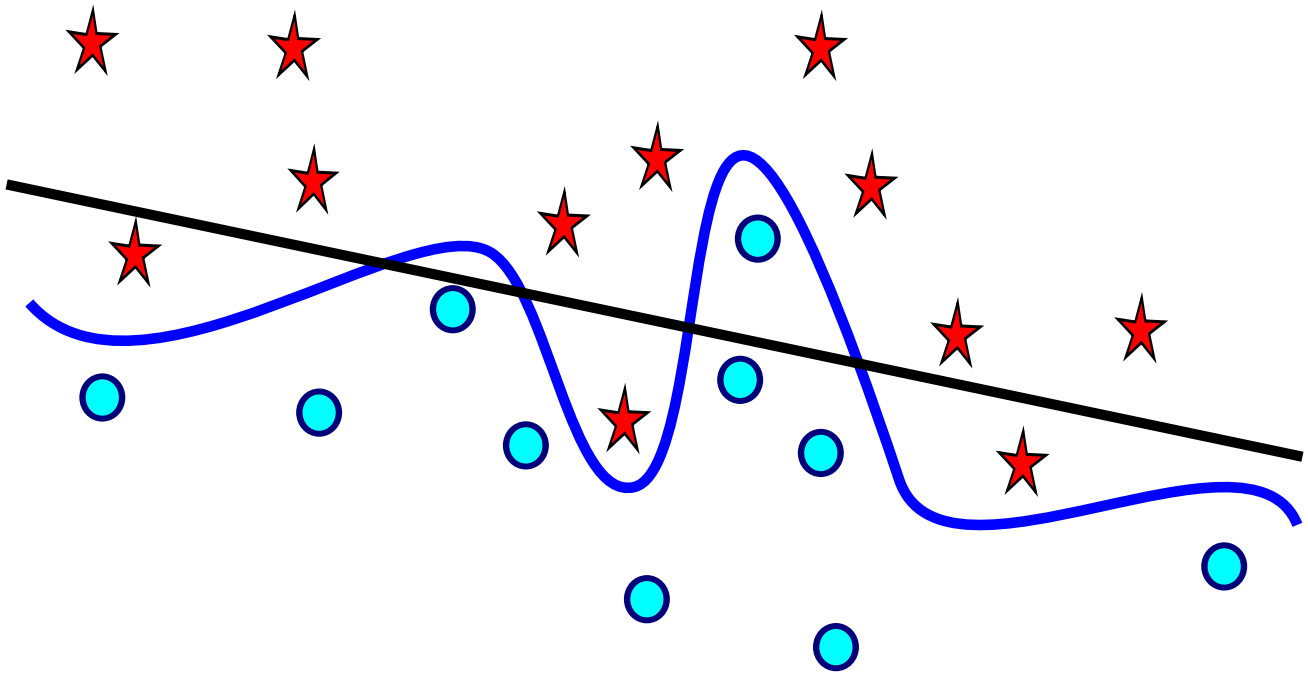


Figure: Los datos no son linealmente separables

SVM Caso 1) Linealmente separables

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

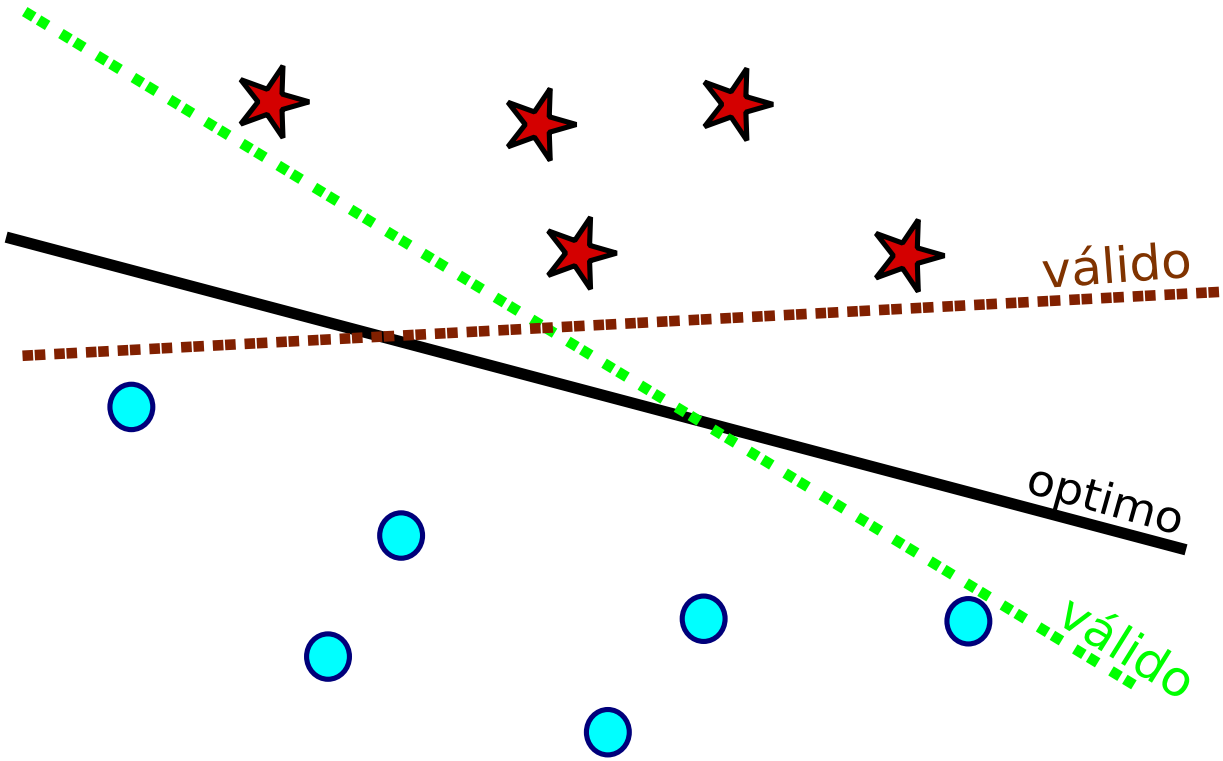


Figure: Los datos son linealmente separables, y hay infinitos hiperplanos que los separan

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 $d(\boldsymbol{x}_j, H_1) = |H(\boldsymbol{x}_j)|$.

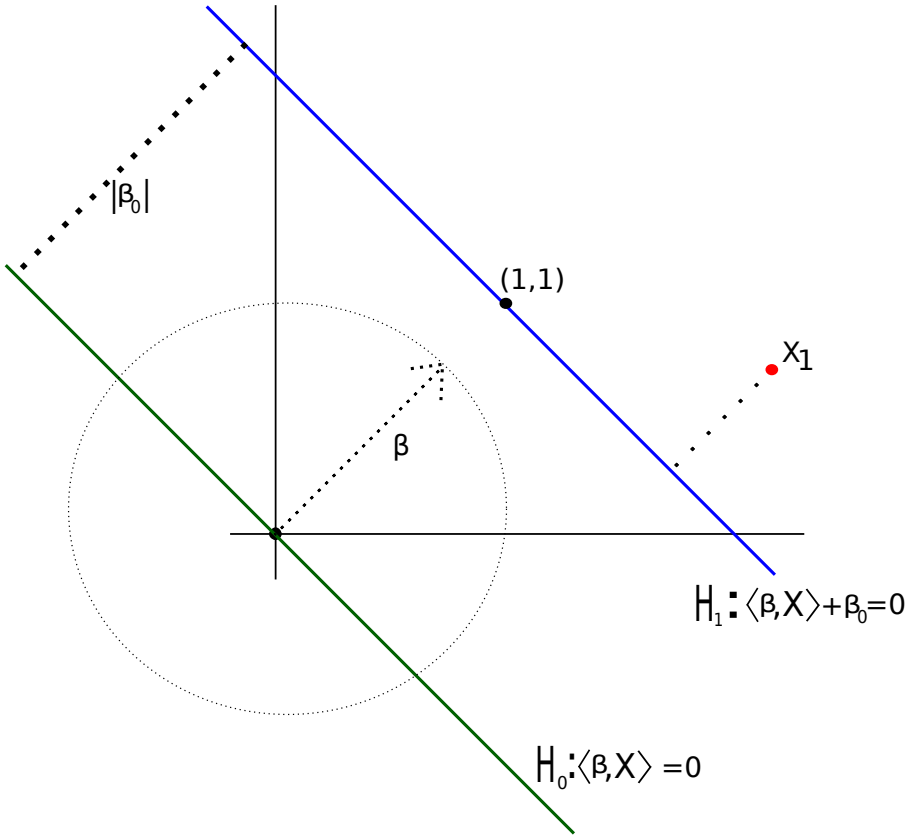


Figure: La distancia de $\boldsymbol{x}_1$ a H_1 es $|H(\boldsymbol{x}_1)|$

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 $\langle \beta, \boldsymbol{x}_i \rangle + \beta_0 \pm C = 0$.

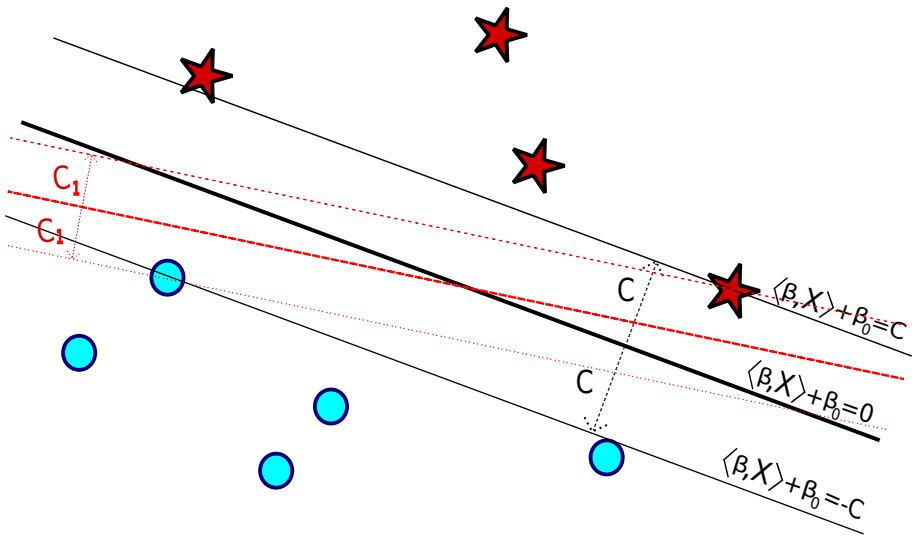


Figure: Si $y_i H(\boldsymbol{x}_i) > 0$ el dato $\boldsymbol{x}_i$ está bien clasificado

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?

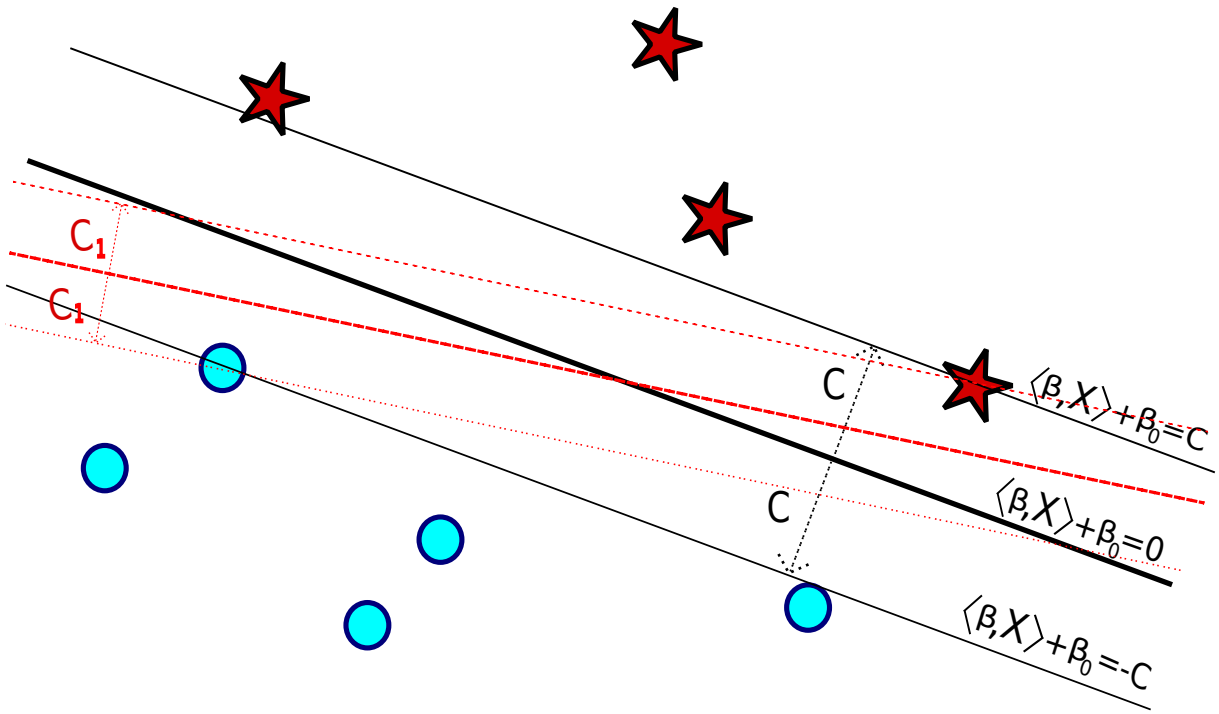


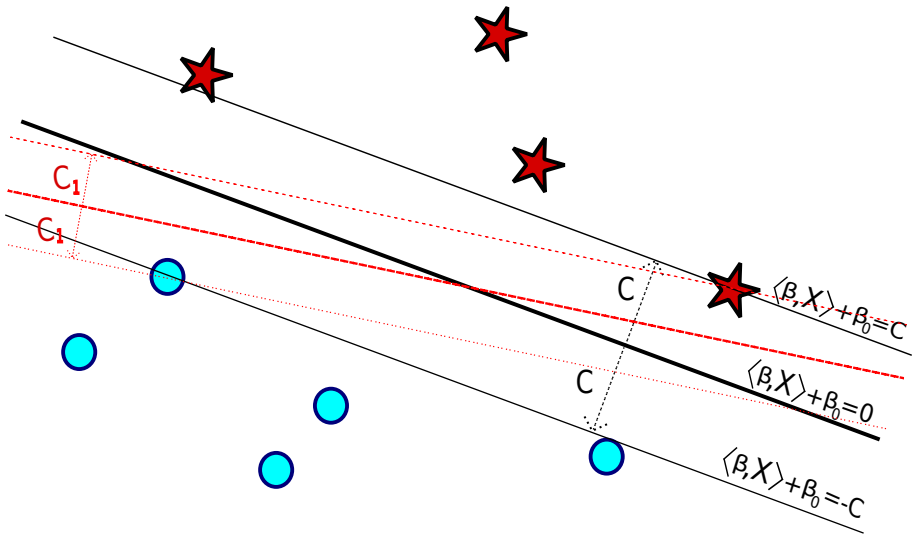
Figure: Si $y_i H(\boldsymbol{x}_i) > 0$ el dato $\boldsymbol{x}_i$ está bien clasificado

Regla de clasificación

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

SVM Caso 1) Linealmente separables

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



El margen no tiene por qué ser único

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 (llamadas condiciones de Karush-Kuhn-Tucker (KKT)) para determinar el vector 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) 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 $\alpha_i^* > 0$.

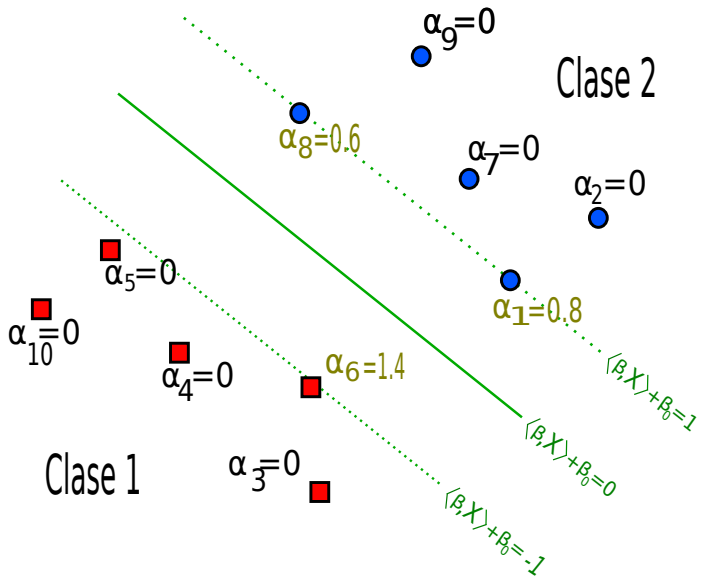


Figure: Los *support vectors* son $\boldsymbol{x}_1, \boldsymbol{x}_6, \boldsymbol{x}_8$. En la resolución del problema relajado, se prueba que $\sum_{i=1}^{n} \alpha_i y_i = 0$

Por lo tanto el clasificador final es:

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

SVM Caso 1) 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 pequeña 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-):

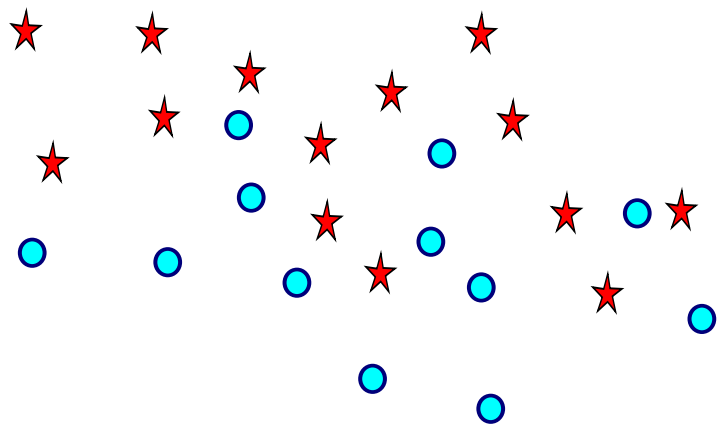


Figure: Soft Margin

- Caso no separable

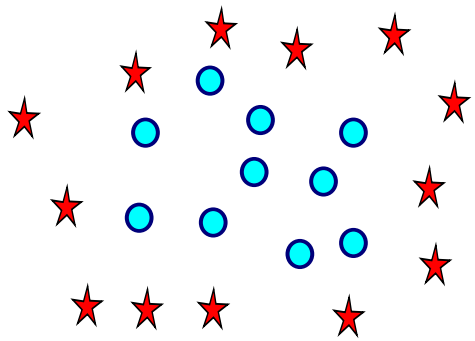


Figure: No separable

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) $\xi_i \geq 0$ $i = 1, \ldots, n$.

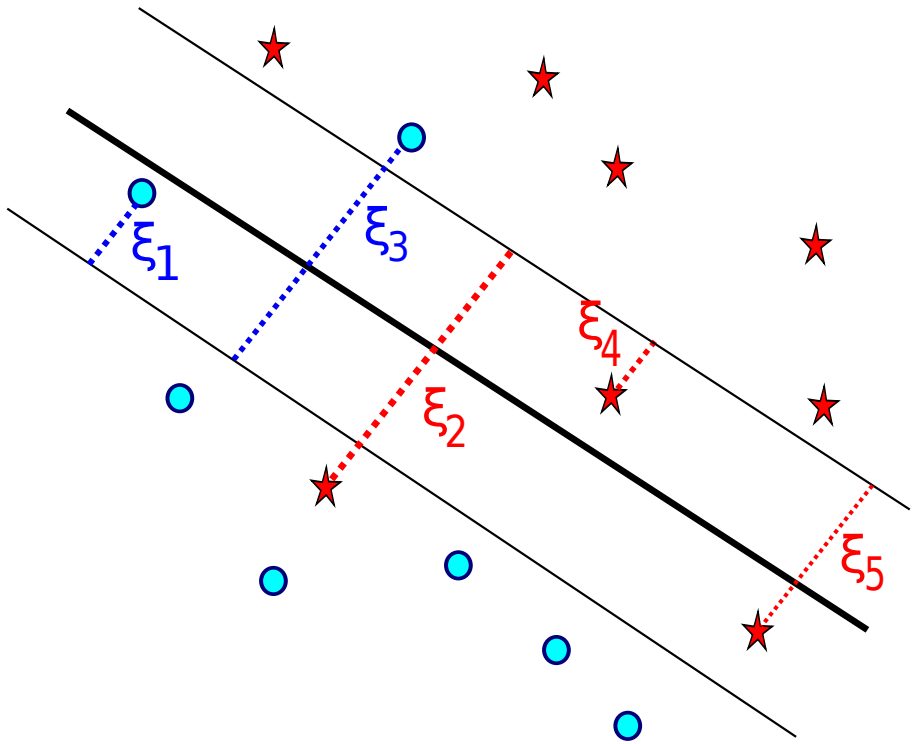


Figure: Si quitamos $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_5$ es linealmente separable

SVM Caso 2) Soft Margin

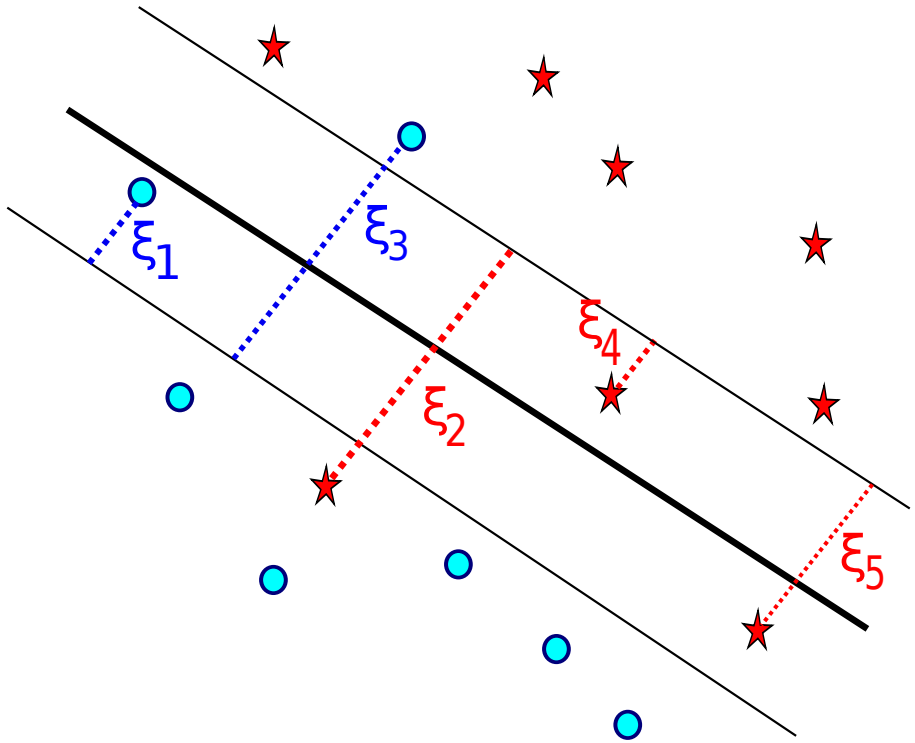


Figure: Si $\xi_i > 1$ el dato i está del otro lado del hiperplano separador $\langle \beta, \boldsymbol{x} \rangle + \beta_0 = 0$

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:

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

El clasificador final es:

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

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.

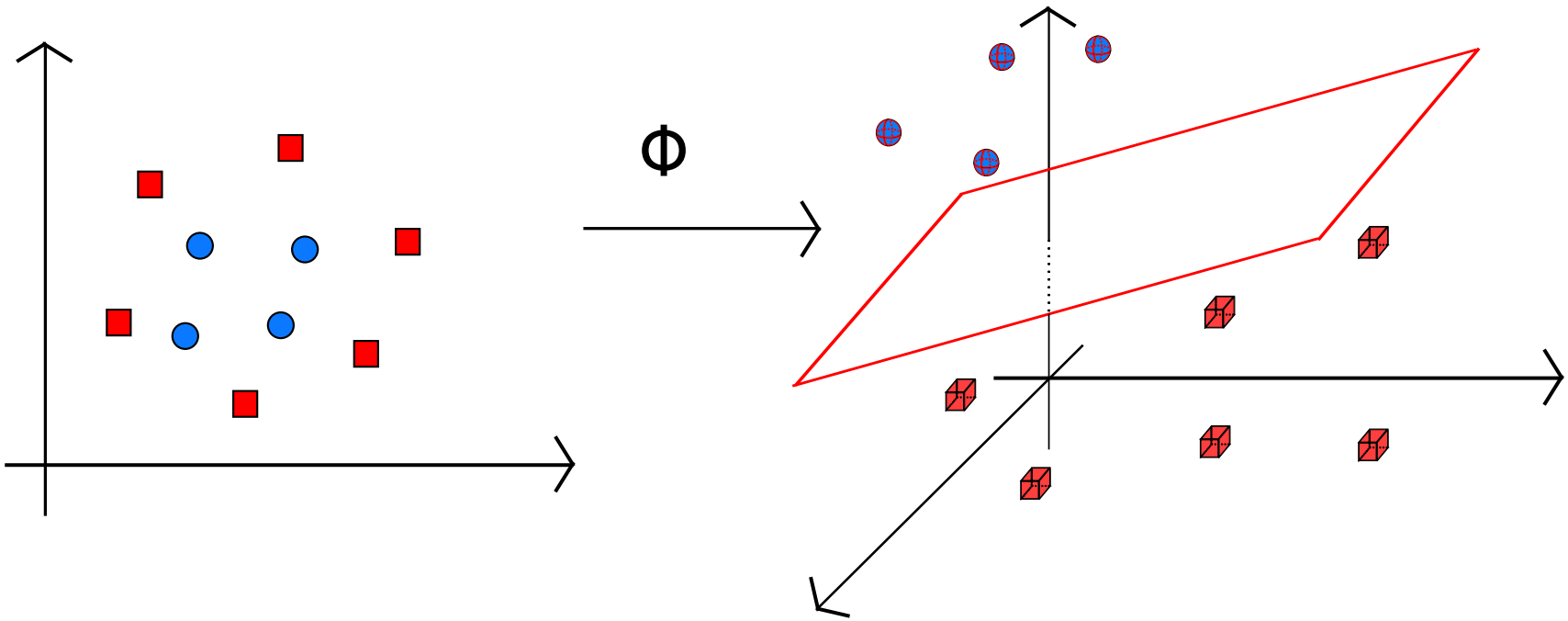


Figure: Los datos en $\mathbb{R}^2$ no son linealmente separables pero podemos tomar $\Phi : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$ de modo que en $\mathbb{R}^3$ sean linealmente separables, (existe un subespacio que los separa)

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 $k(,)$.

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 $\Phi(x) = (x^2, \sqrt{2}x, 1)$.

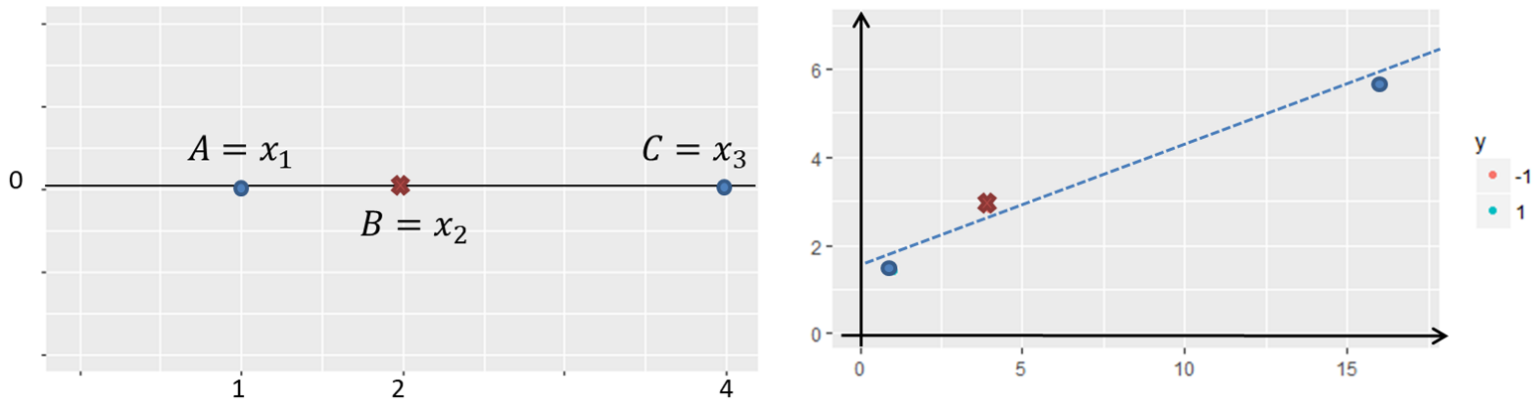


Figure: A la izquierda ejemplo de observaciones en $\mathbb{R}$ no separables. A la derecha los mismos puntos luego de mapearlos en $\mathbb{R}^3$, siendo separables en esta nueva dimensión.

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:

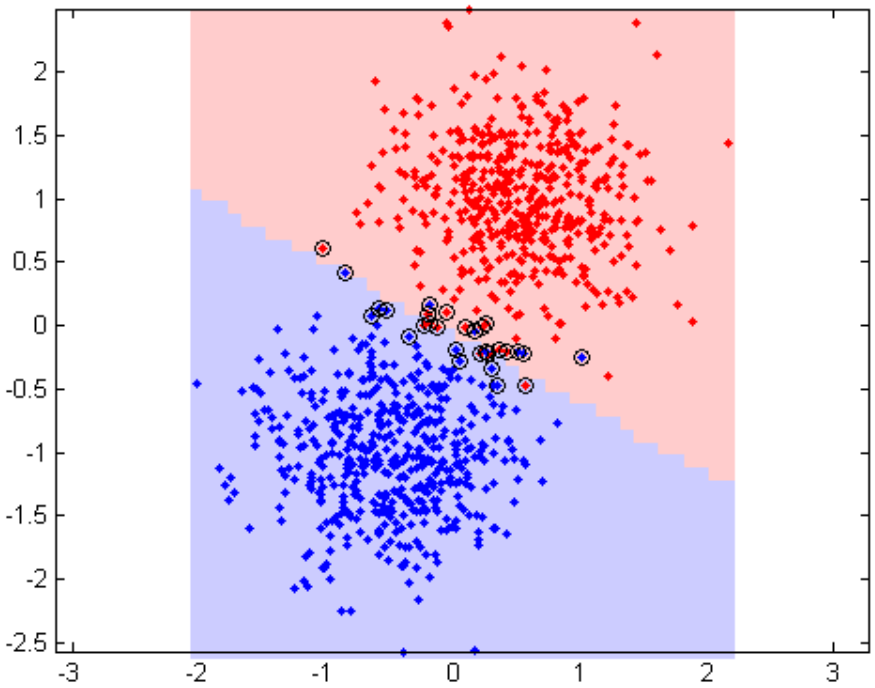
$$\langle \Phi(x_i), \Phi(x_j) \rangle = x_i^2 x_j^2 + \sqrt{2} x_i \sqrt{2} x_j = x_i^2 x_j^2 + 2 x_i x_j + 1 = (x_i x_j + 1)^2 = k(x_i, x_j)$$

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.

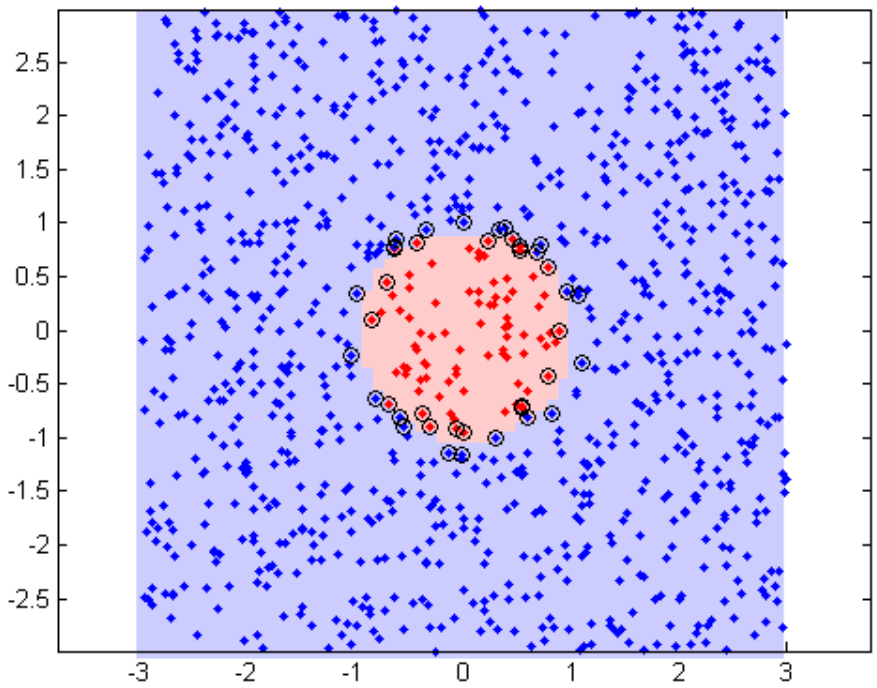
SVM Caso 3) No separable- Núcleos

- 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.
- El clasificador final es:

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



Kernel Lineal



Kernel Radial

Plan

SVM multiclass: una contra todas

Se construyen K clasificadores binarios, uno para cada clase.

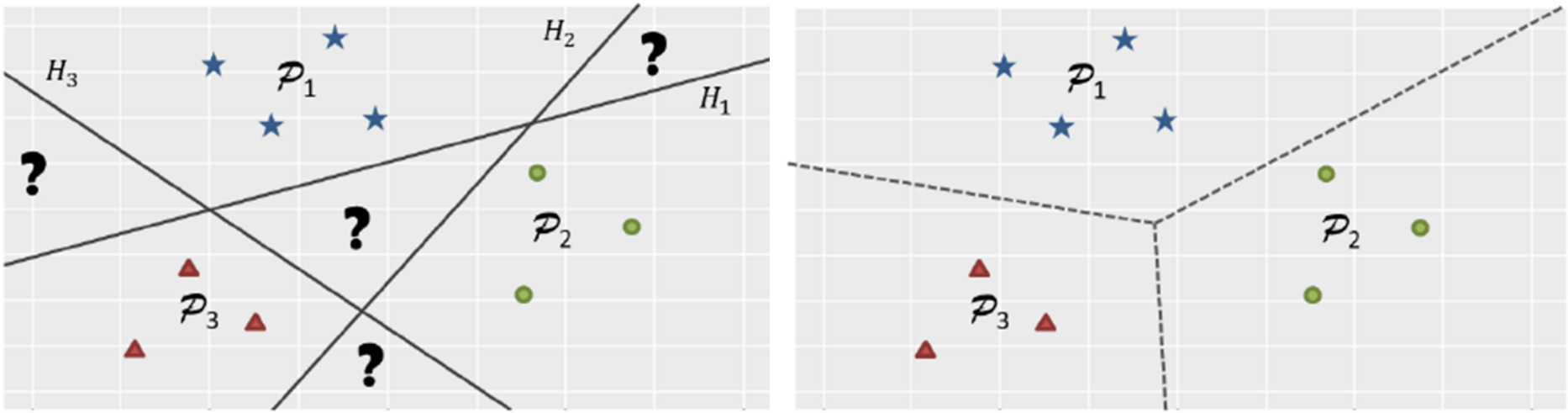


Figure: Descripción gráfica de la estrategia *one vs all*. En la izquierda se muestra un ejemplo con 3 subpoblaciones $\mathcal{P}_1$,$\mathcal{P}_2$, y $\mathcal{P}_3$, junto con los 3 hiperplanos separadores. Se marca las zonas de indeterminación con un signo de pregunta. A la derecha se muestra como 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 $\binom{K}{2}$.

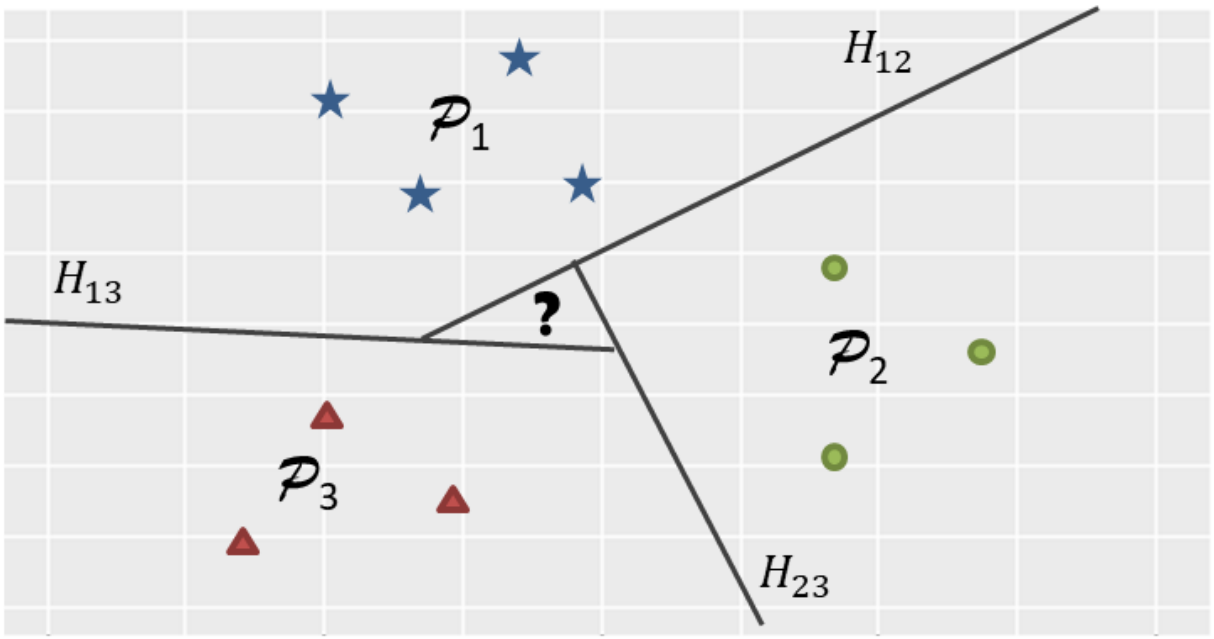


Figure: Descripción gráfica de la estrategia *one versus one*. Se muestran las 3 subpoblaciones $\mathcal{P}_1$, $\mathcal{P}_2$, y $\mathcal{P}_3$, junto con los hiperplanos separadores H_{ij} donde ij indica las poblaciones que separa dicho hiperplano. Una nueva observación $\boldsymbol{x}_0$ se etiqueta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

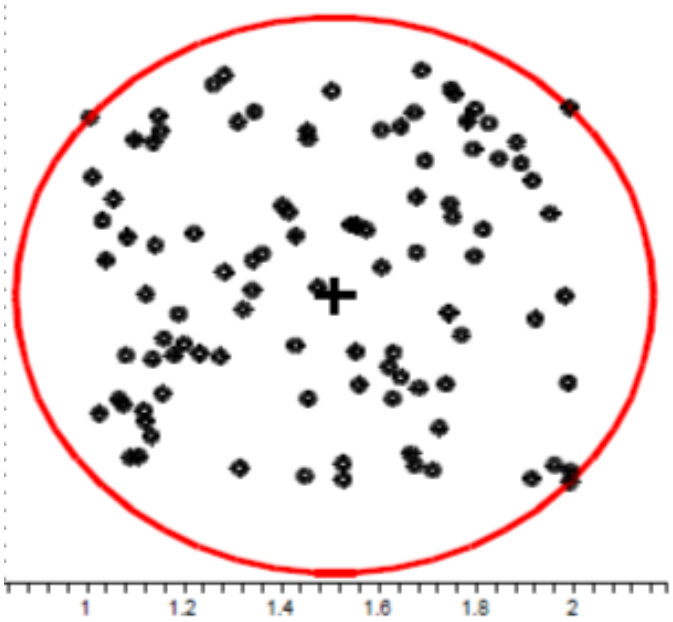
One class SVM

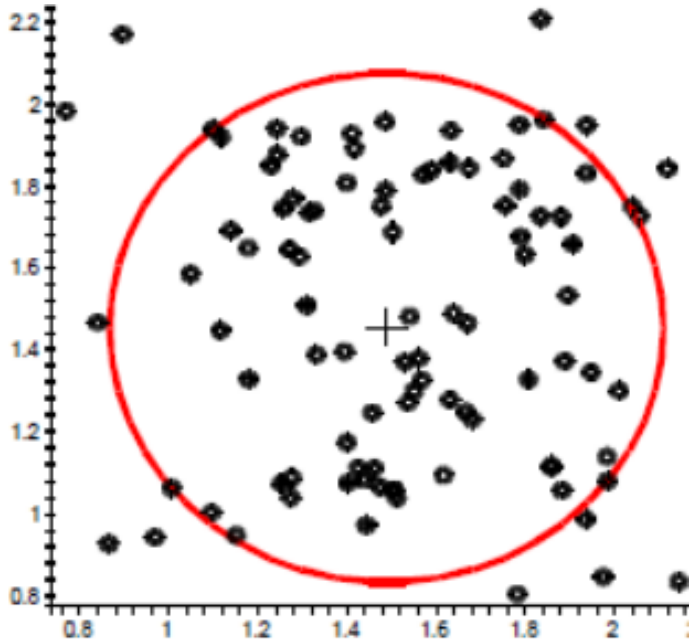
Encerrar los datos en una hiperesfera y classificar nuevos datos como normales si cae dentro de la hiperesfera o anomalo si cae afuera. También podemos suponer tolerancia al ruido.

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

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

One class SVM





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:

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

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 $\mathcal{L}$.

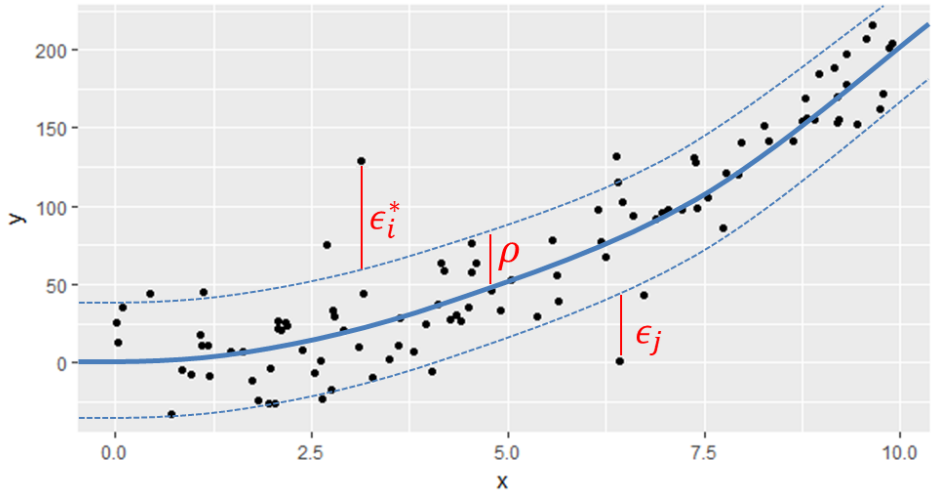


Figure: Hiperplano hallado para *SVR*. Se muestra el radio ρ del "hiper-tubo" y las variables de holgura de dos observaciones (ϵ_i^* y ϵ_j).


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.