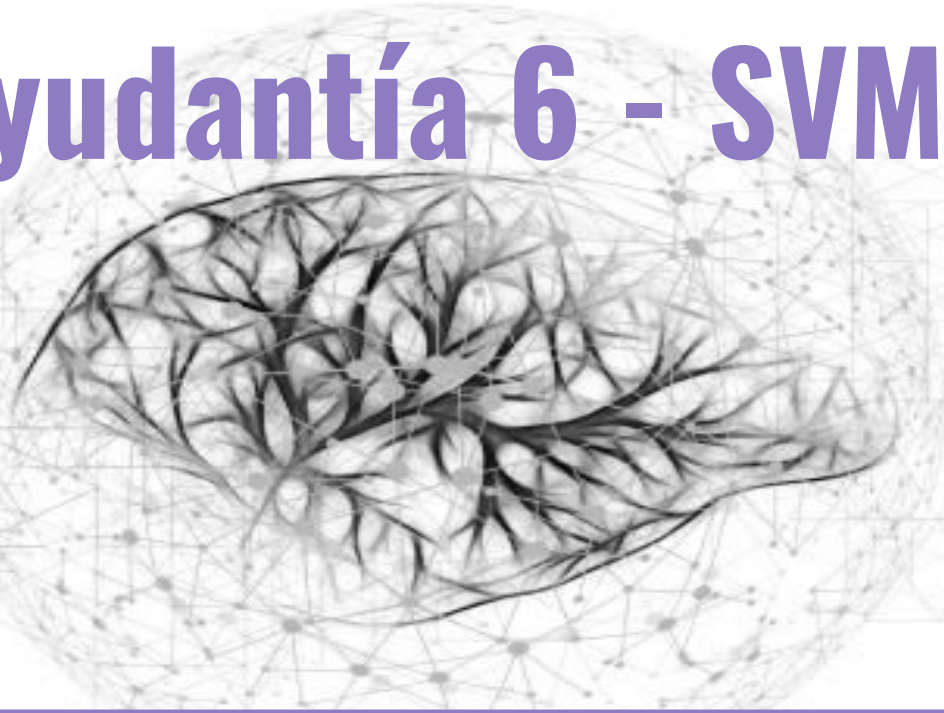
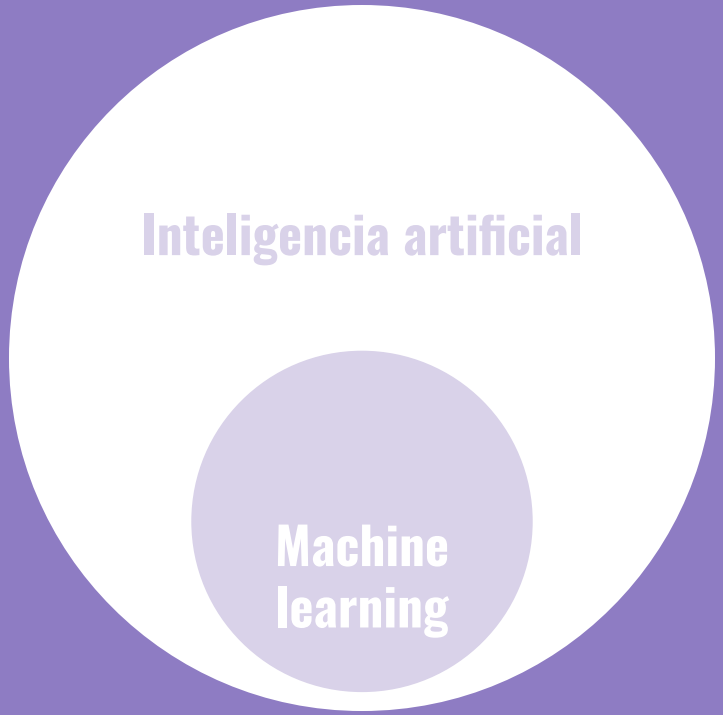
Ayudantía 6 - SVM

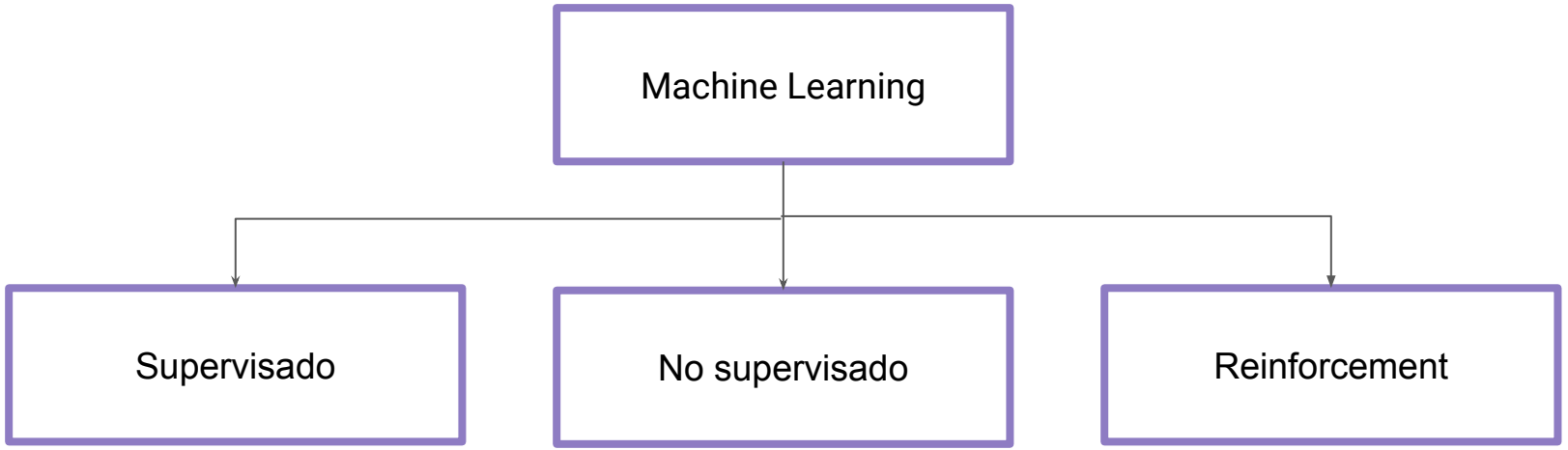


Por Sarah Everke y Florencia Valdivia

Machine Learning



Categorías

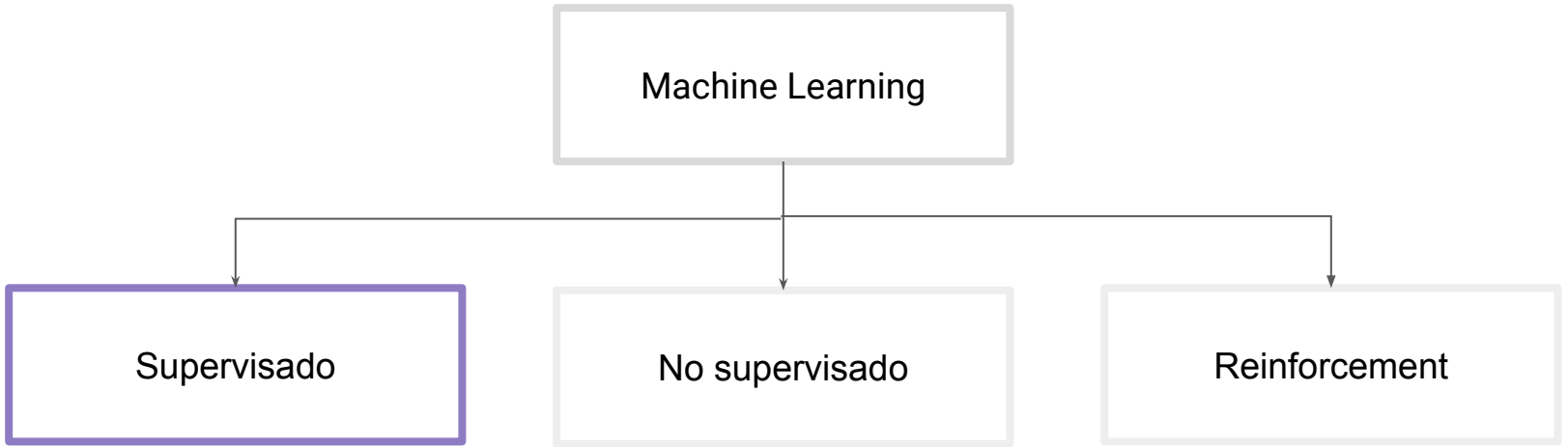


Entrenados con **datos etiquetados**

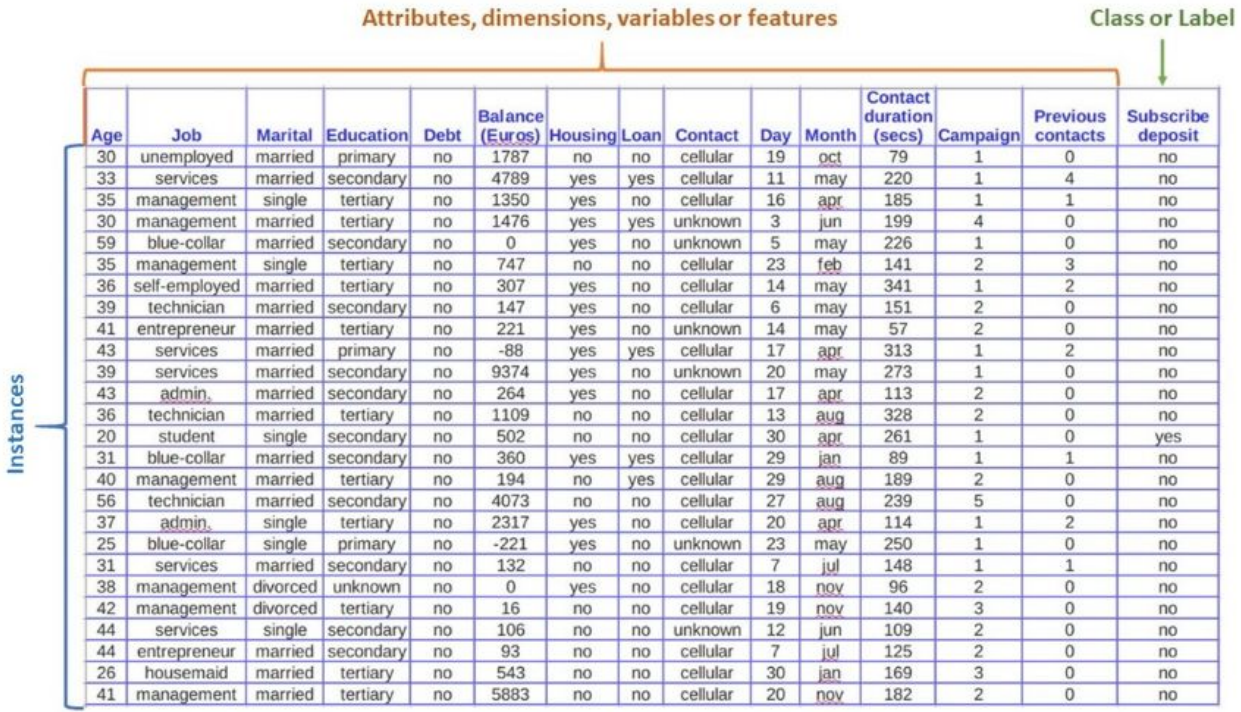
Se buscan patrones en **datos no etiquetados**

Se entrena mediante **prueba y error** para tomar la mejor acción mediante un **sistema de recompensas**

Categorías

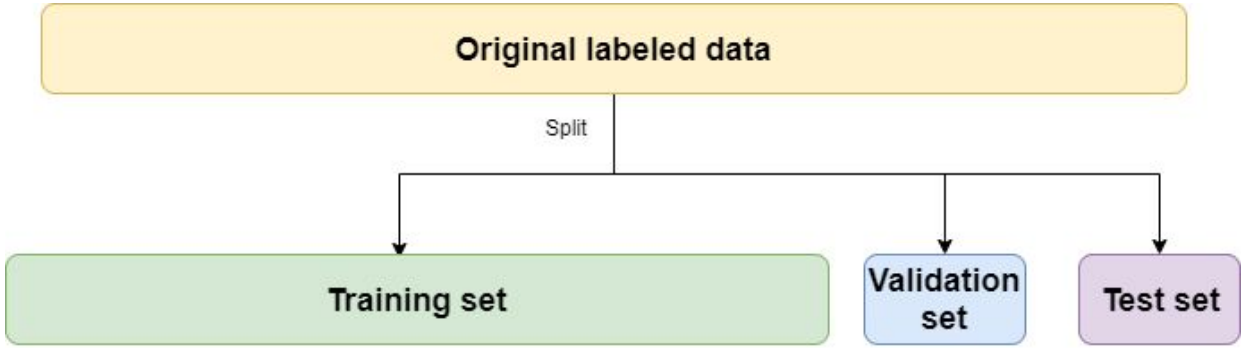

Machine Learning
Supervisado
No supervisado
Reinforcement

Set de datos



Age	Job	Marital	Education	Debt	Balance (Euros)	Housing	Loan	Contact	Day	Month	Contact duration (secs)	Campaign	Previous contacts	Subscribe deposit
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	0	no
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	4	no
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	1	no
30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	0	no
59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	0	no
35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	3	no
36	self-employed	married	tertiary	no	307	yes	no	cellular	14	may	341	1	2	no
39	technician	married	secondary	no	147	yes	no	cellular	6	may	151	2	0	no
41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14	may	57	2	0	no
43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	2	no
39	services	married	secondary	no	9374	yes	no	unknown	20	may	273	1	0	no
43	admin.	married	secondary	no	264	yes	no	cellular	17	apr	113	2	0	no
36	technician	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	0	no
20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	0	yes
31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	1	no
40	management	married	tertiary	no	194	no	yes	cellular	29	aug	189	2	0	no
56	technician	married	secondary	no	4073	no	no	cellular	27	aug	239	5	0	no
37	admin.	single	tertiary	no	2317	yes	no	cellular	20	apr	114	1	2	no
25	blue-collar	single	primary	no	-221	yes	no	unknown	23	may	250	1	0	no
31	services	married	secondary	no	132	no	no	cellular	7	jul	148	1	1	no
38	management	divorced	unknown	no	0	yes	no	cellular	18	nov	96	2	0	no
42	management	divorced	tertiary	no	16	no	no	cellular	19	nov	140	3	0	no
44	services	single	secondary	no	106	no	no	unknown	12	jun	109	2	0	no
44	entrepreneur	married	secondary	no	93	no	no	cellular	7	jul	125	2	0	no
26	housemaid	married	tertiary	no	543	no	no	cellular	30	jan	169	3	0	no
41	management	married	tertiary	no	5883	no	no	cellular	20	nov	182	2	0	no

Set de datos



Training set: utilizado para entrenar el modelo.

Testing set: utilizado para medir el rendimiento.

Validation set: utilizado para ajustar los hiperparámetros.

SVM

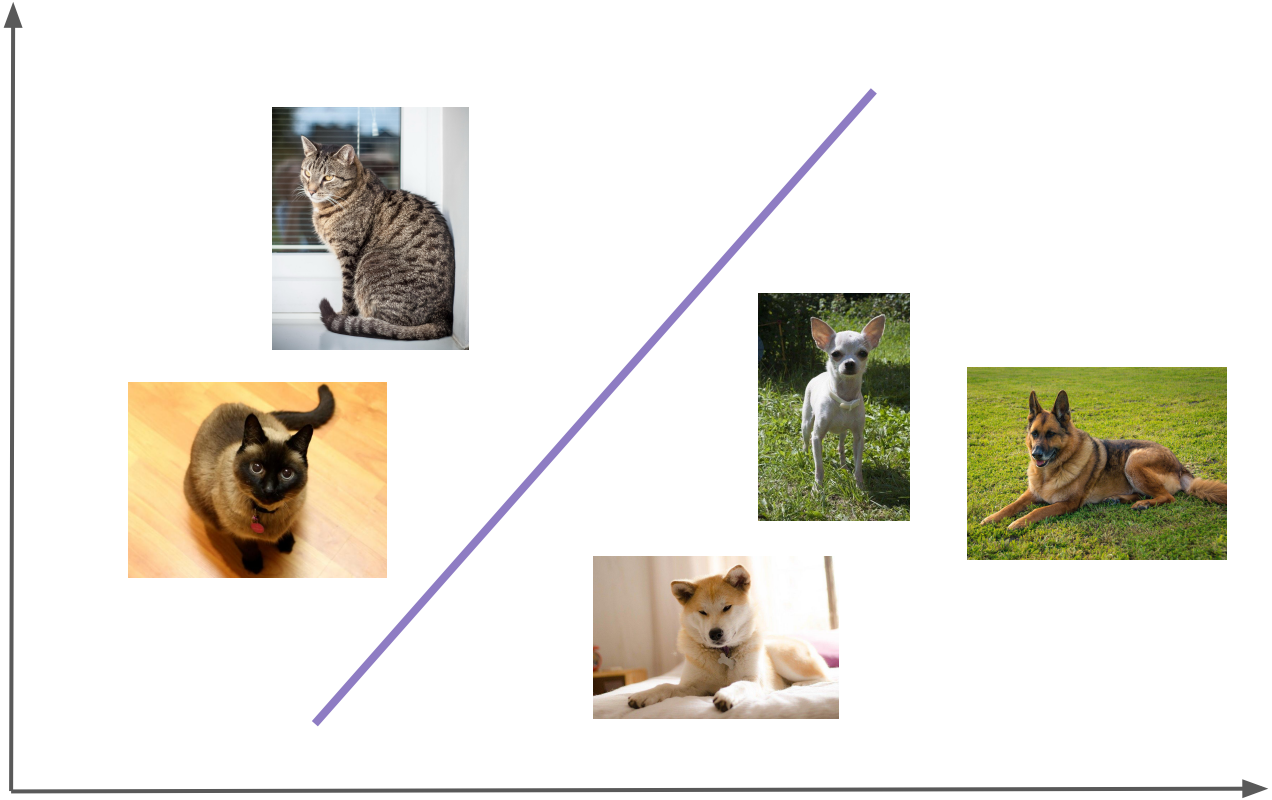
¿En qué casos lo usamos?

➔ SVM es excelente para conjuntos de datos relativamente pequeños con pocos outliers.

➔ Casos de uso: Clasificación, regresión, detección de outliers, clustering.

➔ Es buen algoritmo para casos extremos donde queremos clasificar un nuevo dato que es difícil (caso extremo).

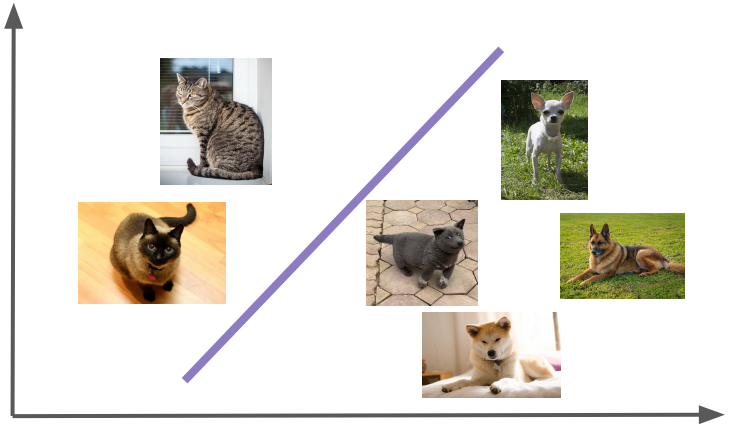
Un caso extremo...



¿  ?

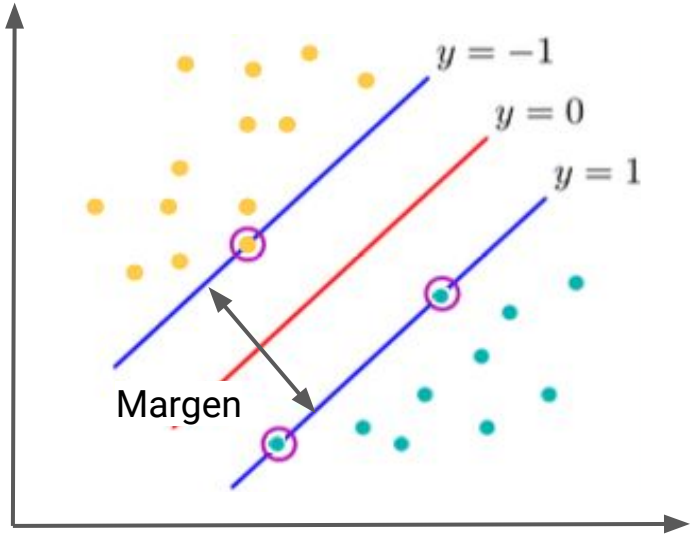
SVM

Es un algoritmo de aprendizaje automático supervisado que normalmente se utiliza para la clasificación. Dadas 2 o más clases de datos etiquetados, actúa como un clasificador discriminante, definido por un hiperplano óptimo que separa todas las clases. Los nuevos ejemplos mapeados en ese mismo espacio se pueden clasificar en función del **lado de la brecha** en el que caen.



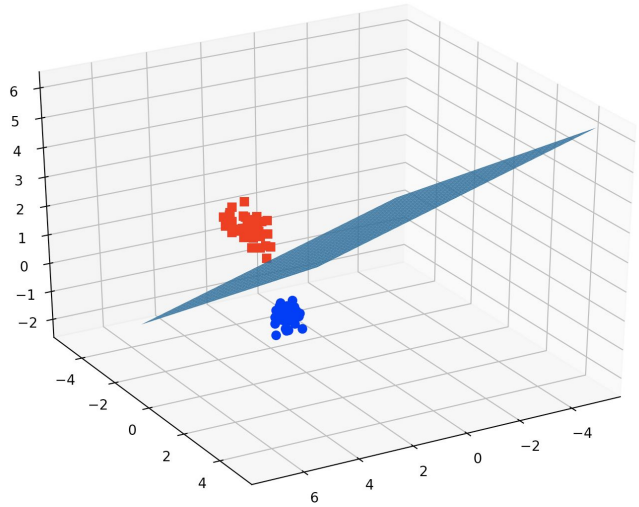
Support Vectors

Los **support vectors** son los puntos de datos más cercanos al hiperplano. Si estos puntos se eliminan de un conjunto de datos, alterarían la posición del hiperplano divisor. Pueden considerarse los elementos críticos de un conjunto de datos, son los que nos ayudan a construir nuestro SVM.



Hiperplano

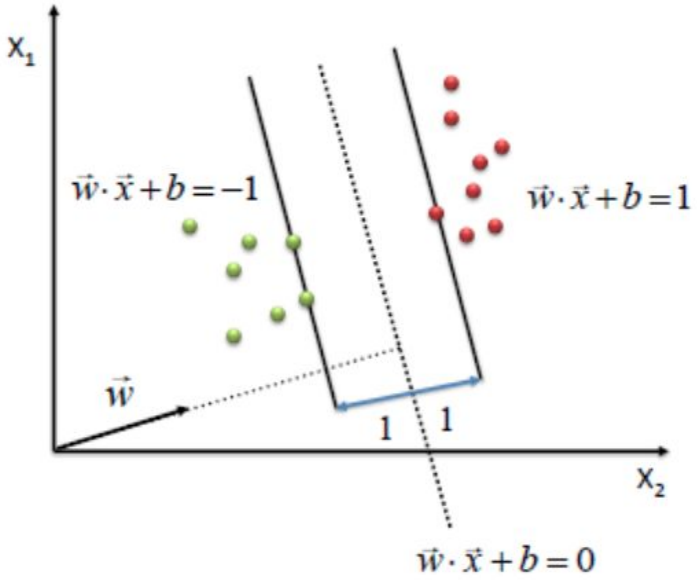
Es una **superficie de decisión lineal** que divide el espacio en dos partes, siendo un clasificador binario. Queremos encontrar la superficie de decisión que esté lo más lejos posible de cualquier punto de datos.



SVM

El hiperplano está dado por $g(x) = wx_k + c$

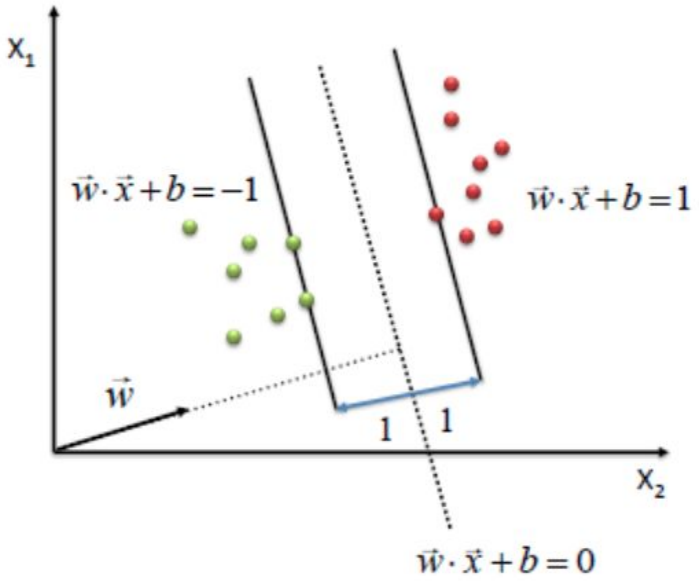
y $g(x) = wx_k + c \begin{cases} > 0, & \text{si } x_k \in C_1 \\ \leq 0, & \text{si } x_k \in C_2 \end{cases}$



SVM

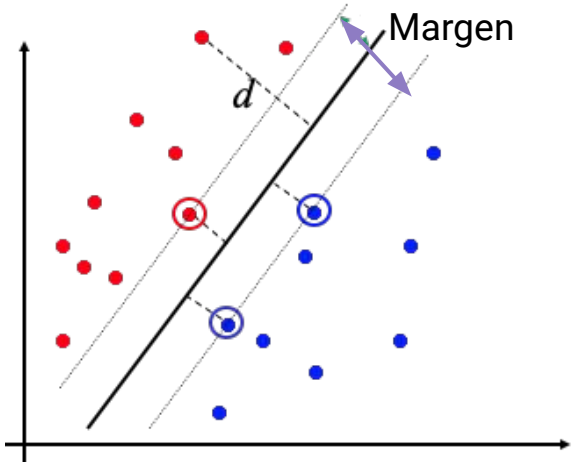
Por conveniencia se define $z_k = \pm 1$ dependiendo de la clase que pertenece x_k

y la desigualdad anterior queda como $z_k(wx_k + c) > 0,\ k = 1...n$



SVM

La distancia una instancia x_k al hiperplano está dada por $d = \frac{|g(x_k)|}{||w||}$

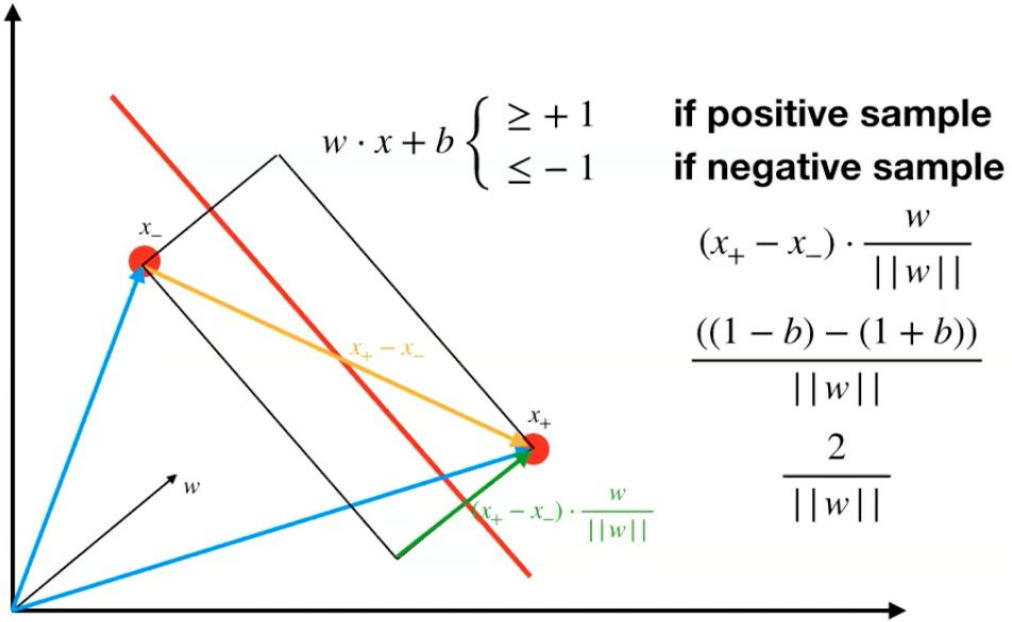


SVM

Así, el problema de optimización es

$$\underset{w,c}{\text{argmax}} \left\{ \frac{1}{||w||} \min_k \{ z_k g(x_k) \} \right\}$$

$$\text{sujeto a } z_k(wx_k + c) > 0, \; k = 1..0$$



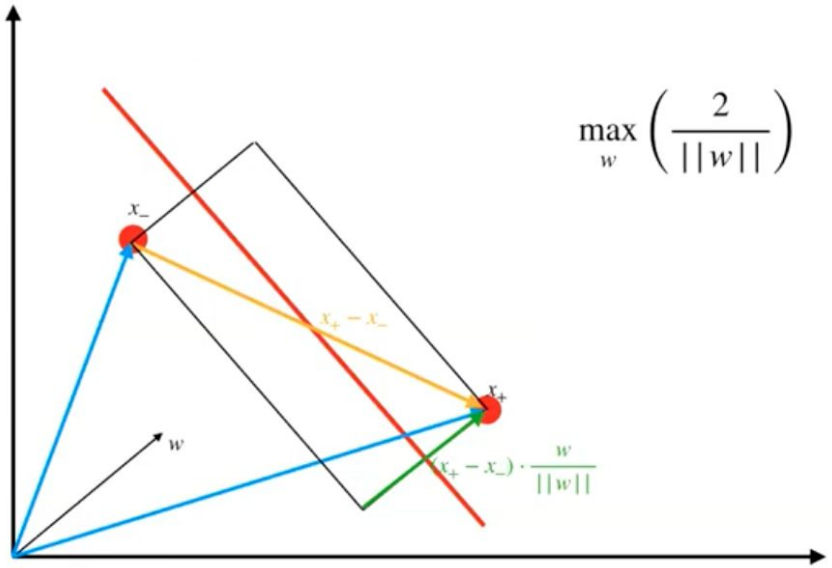
Ref. Cristóbal Eyzaguirre

SVM

Así, el problema de optimización es

$$\underset{w,c}{\text{argmax}} \left\{ \frac{1}{||w||} \min_{k} \{z_k g(x_k)\} \right\}$$

$$\text{sujeto a } z_k(wx_k + c) > 0,\ k = 1..0$$



$$\max_{w} \left(\frac{2}{||w||} \right)$$

Ref. Cristóbal Eyzaguirre

SVM

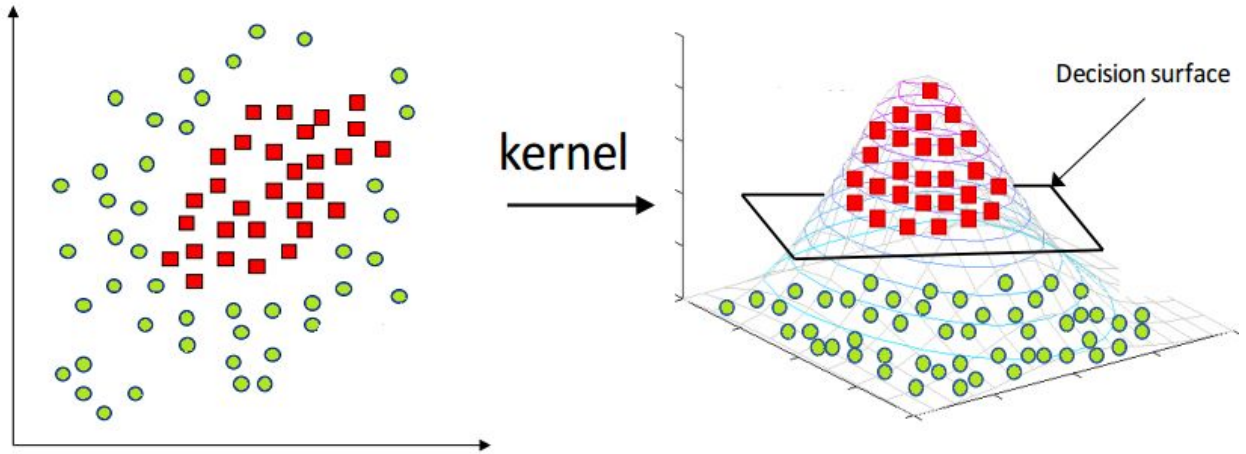
Así, el problema de optimización es

$$\underset{w,c}{\text{argmax}} \frac{1}{||w||} \quad \text{sujeto a } z_k(wx_k + c) > 0 \rightarrow \underset{w,c}{\text{argmin}} ||w|| \quad \text{sujeto a } z_k(wx_k + c) > 0 \rightarrow \underset{w,c}{\text{argmin}} \frac{1}{2}||w||^2 \quad \text{sujeto a } z_k(wx_k + c) > 0$$

Más detalles en: https://www.youtube.com/watch?v=_PwhiWxHK8o

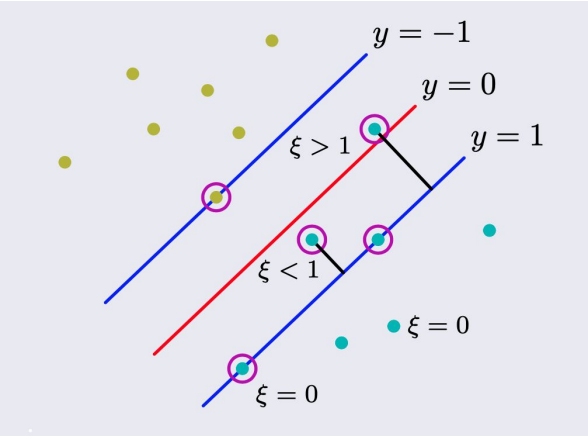
SVM: Caso no lineal

Mapear los datos a **espacio de mayor dimensionalidad** donde los datos son linealmente separables.



SVM: Caso no son separables

Se introducen **variables de slack** $\xi_k = |z_k - g(x_k)|$



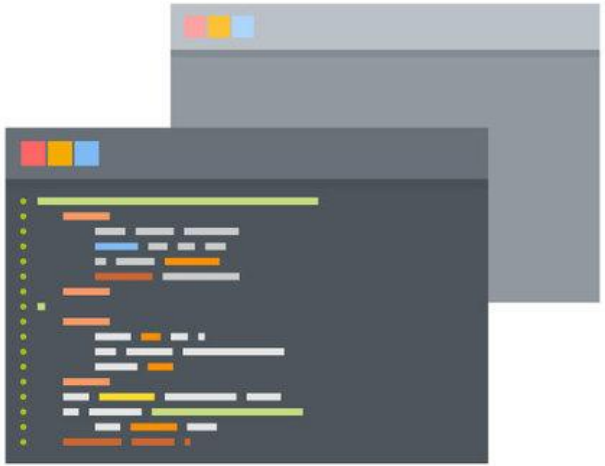
SVM: Caso no son separables

El problema de optimización queda

$$\underset{w,c}{\operatorname{argmin}} \frac{1}{2}||w||^2 + C\sum_{k=1}^{K} \xi_k$$

$$\text{sujeto a } z_k(wx_k + c) > 0,\ k = 1..0$$

Ejercicios



¿Cómo usamos SVM para más de dos clases?

¿Cómo usamos SVM para más de dos clases?