EGADE Business School
Tecnológico de Monterrey

Aplicaciones de analítica de datos a los negocios II

PROF: JUAN C. BUSTAMANTE

JUCBUSTAM@TEC.MX

Normas para la conexión síncrona:

1. La clases tiene un back-up garantizado (Grabación disponible en la nube de Zoom).
2. Ingresar a la clase con la cámara del equipo de computo encendida.
3. La **cámara deberá permanecer encendida a lo largo de la clase**.
4. Al ingresar a la clase deben silenciar el micrófono del equipo de computo.
5. Levantar la mano es una opción cuando se quiere preguntar algo durante la sesión de clase, pero les recomiendo que mejor hagamos uso intensivo del chat del canal general para hacer preguntas**.**
6. En caso de necesitar hacer una pregunta, puede interrumpir la clase sin problema, activando el micrófono de vuestro equipo de computo, luego de la pregunta desactívelo nuevamente**.**
7. Para una buena clase online es indispensable **el debate, así que foméntelo!!!**.
8. Toda la información se gestiona en CANVAS LMS.

EGADE Business School
Tecnológico de Monterrey

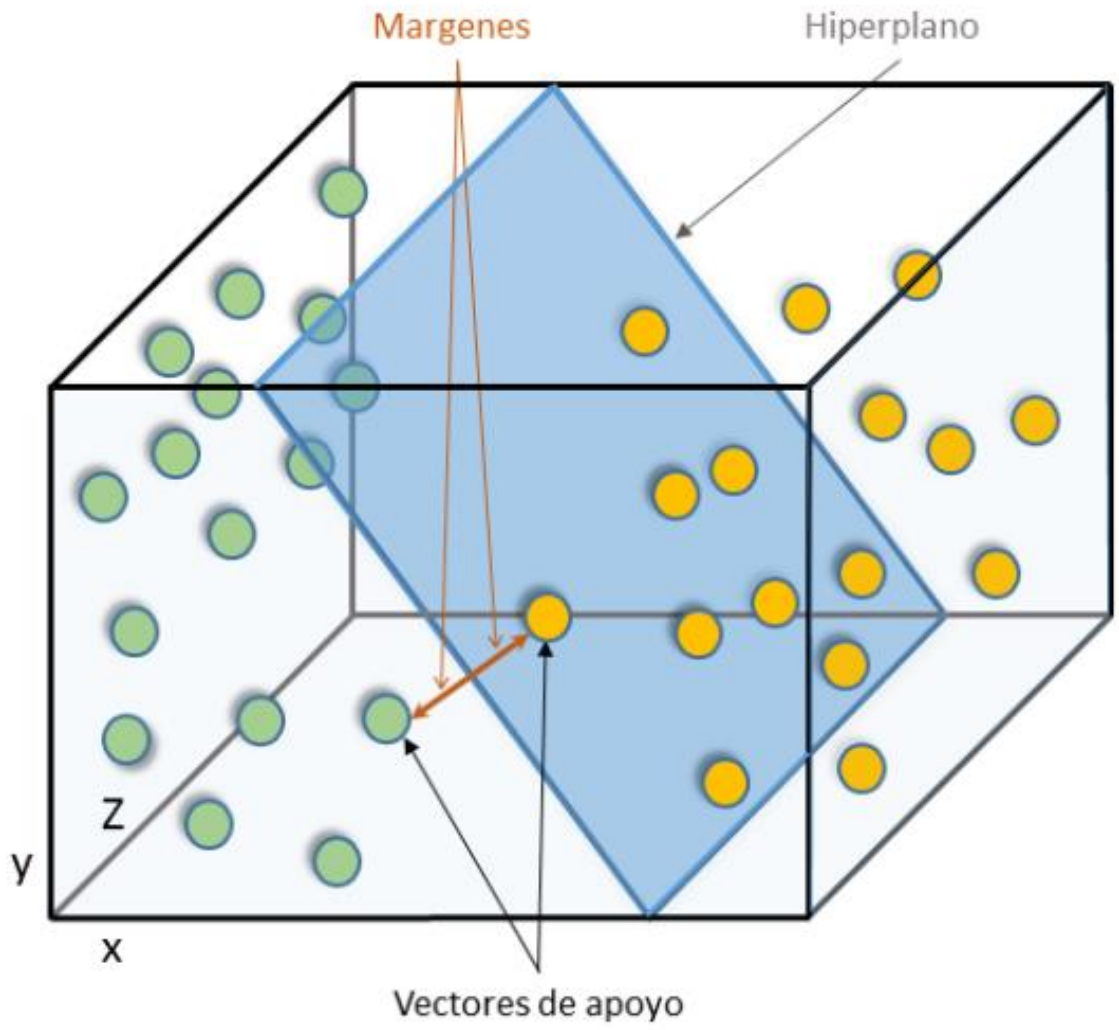
Cronograma de trabajo:

Sesiones	Contenidos	Actividad		Fecha
1	Información general del curso	Utility of classification algorithms		Martes 18/04
2	Algoritmo de regresión logística	Ejecutar script	**Solución caso:** Retention modelling at Scholastic Travel Company (A) and (B)	Martes 25/04
3	Algoritmo Naïve Bayes	Ejecutar script		Martes 02/05
4	Algoritmo k-nearest-neighbors (KNN)	Ejecutar script		Martes 09/05
5	Algoritmo Support vector machine	Ejecutar script		Martes 16/05
6	Algoritmo Decision Trees	Ejecutar script		Martes 23/05
7	Algoritmo Random Forest	Ejecutar script		Martes 30/05
8	Modelo RFM	Ejecutar script	**Solución caso**: CD Now	Martes 06/06
9	Modelo valor de vida del cliente (I)	Ejecutar script		Martes 13/06
10	Modelo valor de vida del cliente (II)	Ejecutar script		Martes 20/06
11	Análisis de series de tiempo	Ejecutar script		Martes 27/06
12	**Proyecto final**	Presentación en equipos		Martes 04/07
	Evaluación final			

1

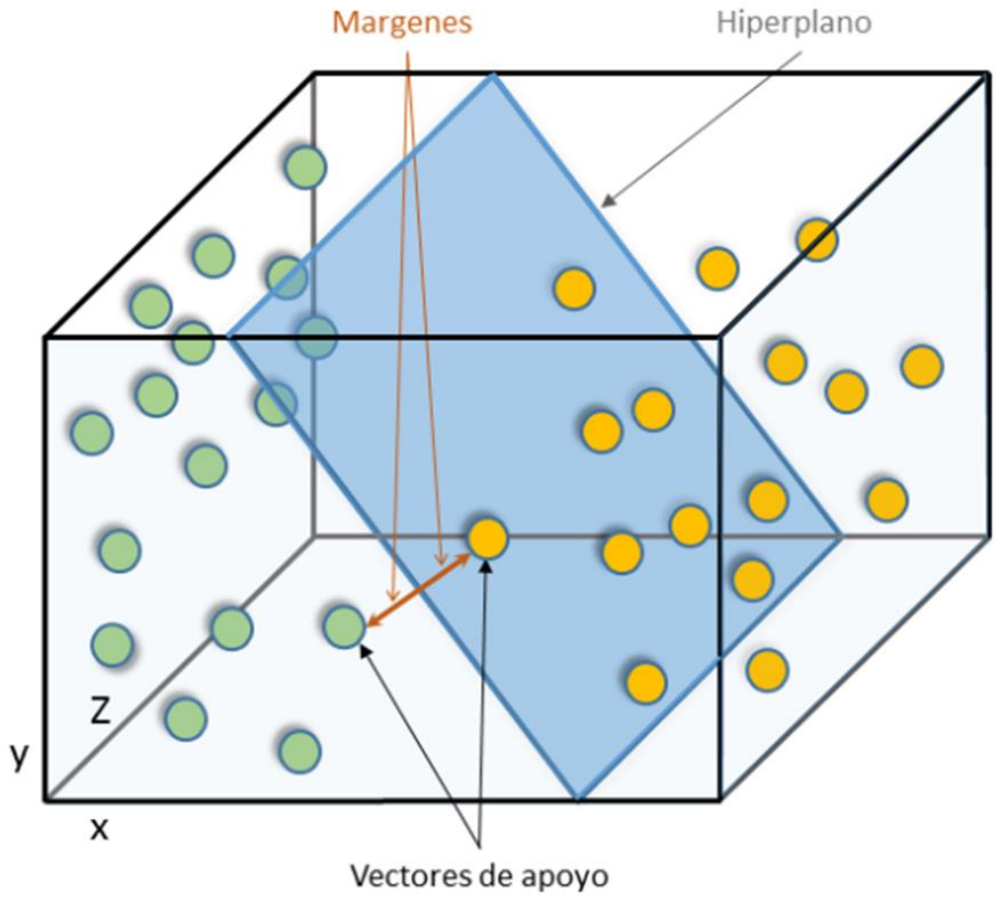
Maquinas de soporte vectorial (SVM)

Maquinas de soporte vectorial (SVM)



Maquinas de soporte vectorial (SVM)

Una máquina SVM es un algoritmo de **aprendizaje** automático supervisado que puede emplearse para fines de clasificación y regresión de dos grupos de datos. Comprende la máquina de soporte vectorial para la clasificación (**SVC**) y la máquina de soporte vectorial para la regresión (**SVR**). Dado un conjunto de ejemplos de entrenamiento, marcados como pertenecientes a una de dos categorías, el algoritmo de entrenamiento SVM construye un modelo que predice si un nuevo ejemplo cae en una categoría u otra.



SVM PARA CLASIFICACIÓN BINARIA

Dado un conjunto separable de ejemplos $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, donde $\boldsymbol{x}_i \in \mathbb{R}^d$ e $y_i \in \{+1, -1\}$, se puede definir un hiperplano de separación (ver fig. 1a) como una función lineal que es capaz de separar dicho conjunto sin error:

$$D(\boldsymbol{x}) = (w_1 x_1 + \ldots + w_d x_d) + b = < \boldsymbol{w}, \boldsymbol{x} > + b$$

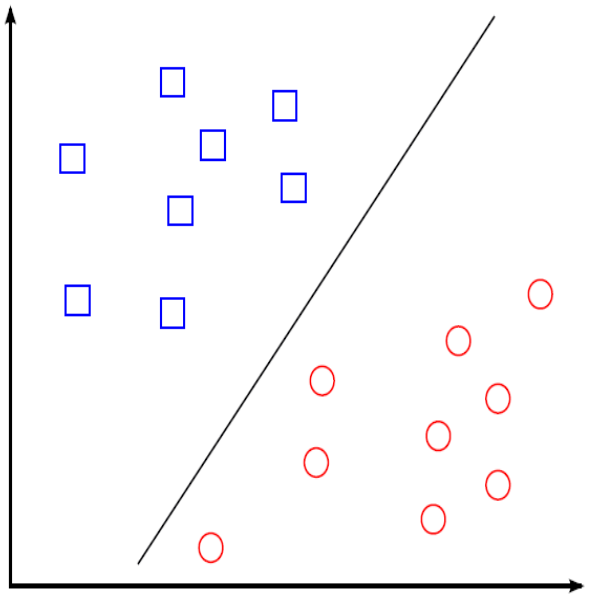
donde $\boldsymbol{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$ y el operador $< \boldsymbol{x}_1, \boldsymbol{x}_2 >$ representa el producto escalar de los vectores $\boldsymbol{x}_1$ y $\boldsymbol{x}_2$. El hiperplano de separación cumplirá las siguientes restricciones para todo $\boldsymbol{x}_i$ perteneciente al conjunto de ejemplos:

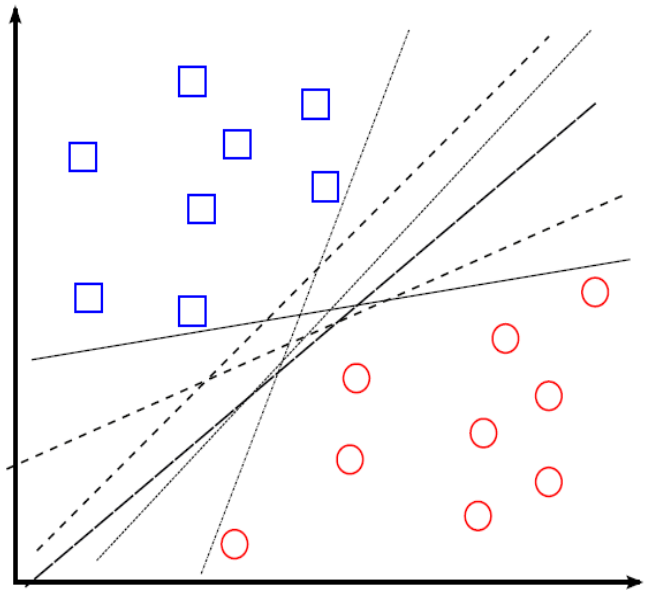
$$\begin{aligned} < \boldsymbol{w}, \boldsymbol{x}_i > + b \geq 0 \quad & \text{si} \quad y_i = +1 \\ < \boldsymbol{w}, \boldsymbol{x}_i > + b \leq 0 \quad & \text{si} \quad y_i = -1, \; i = 1, \ldots, n \end{aligned}$$

o también:

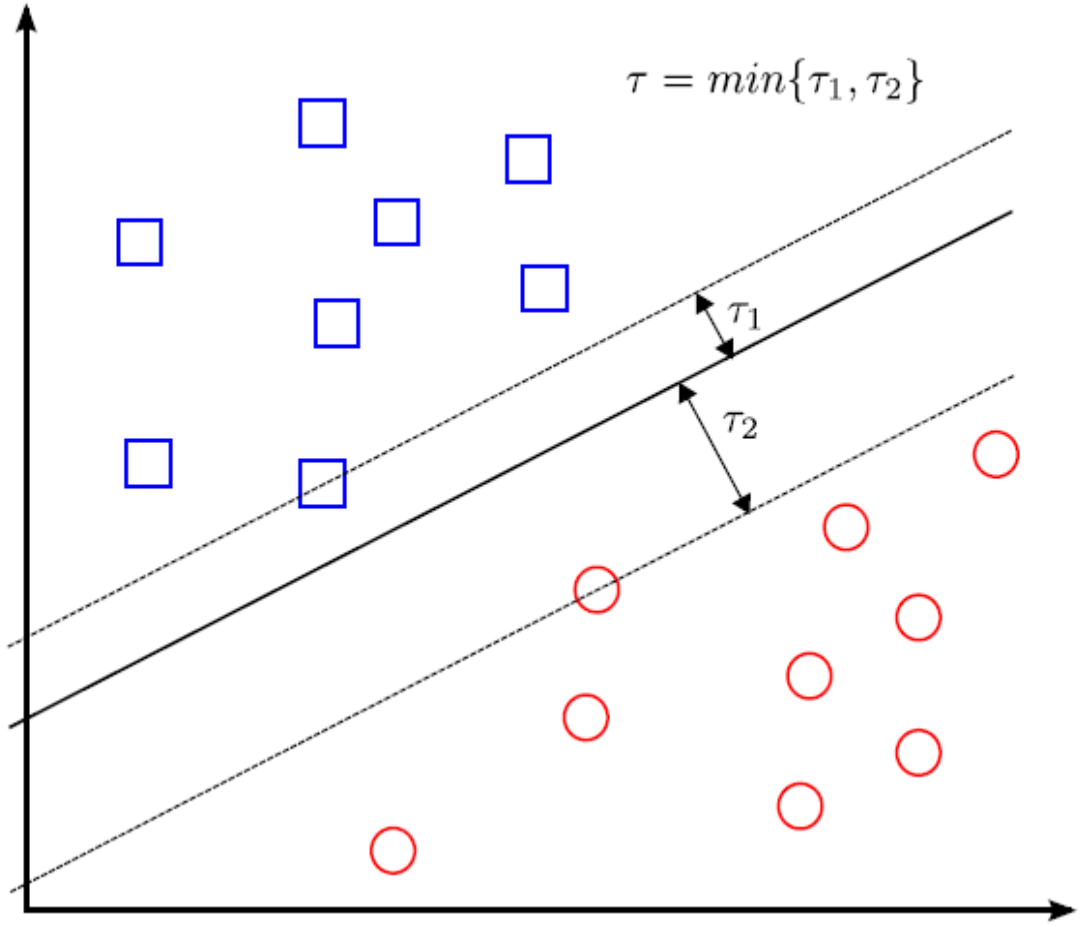
$$y_i \left(< \boldsymbol{w}, \boldsymbol{x}_i > + b \right) \geq 0, \qquad i = 1, \ldots, n$$

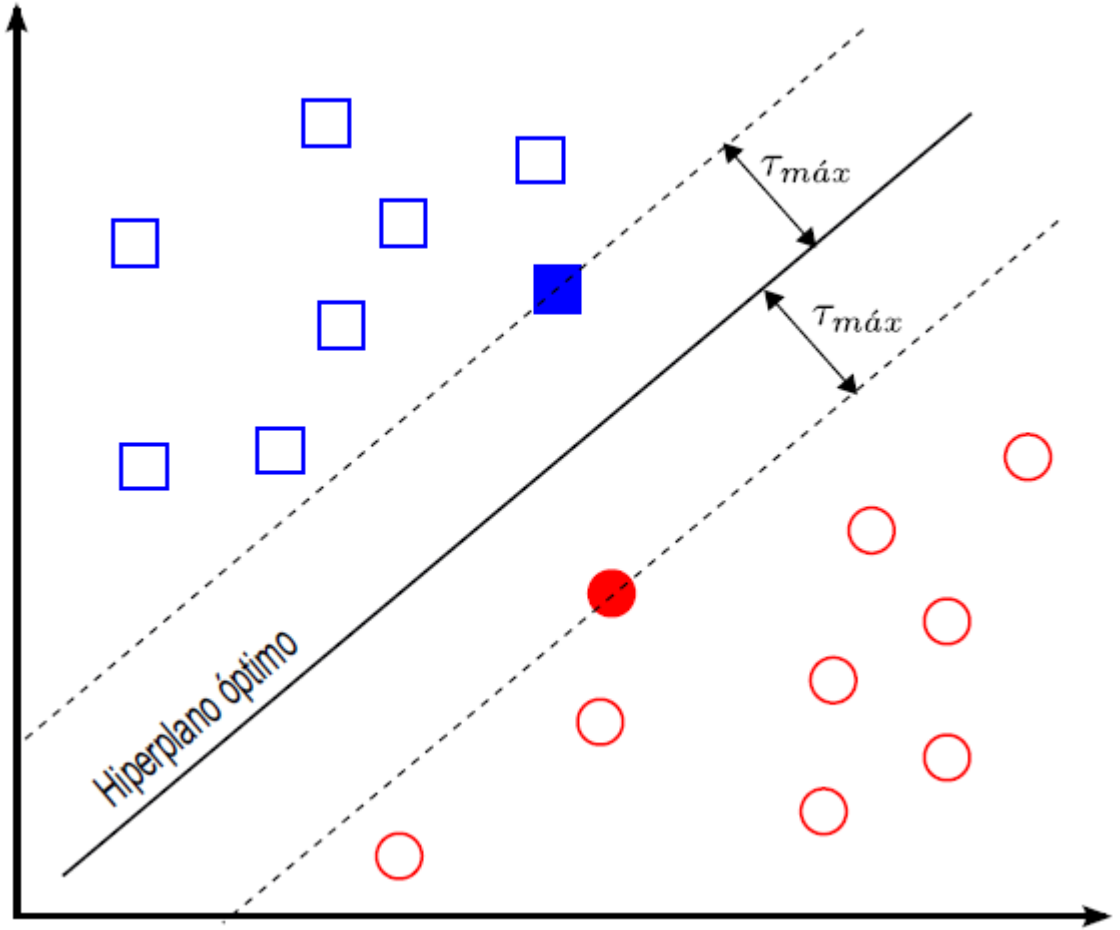
SVM (HIPERPLANOS DE SEPARACIÓN)





SVM (MARGEN DE UN HIPERPLANO)





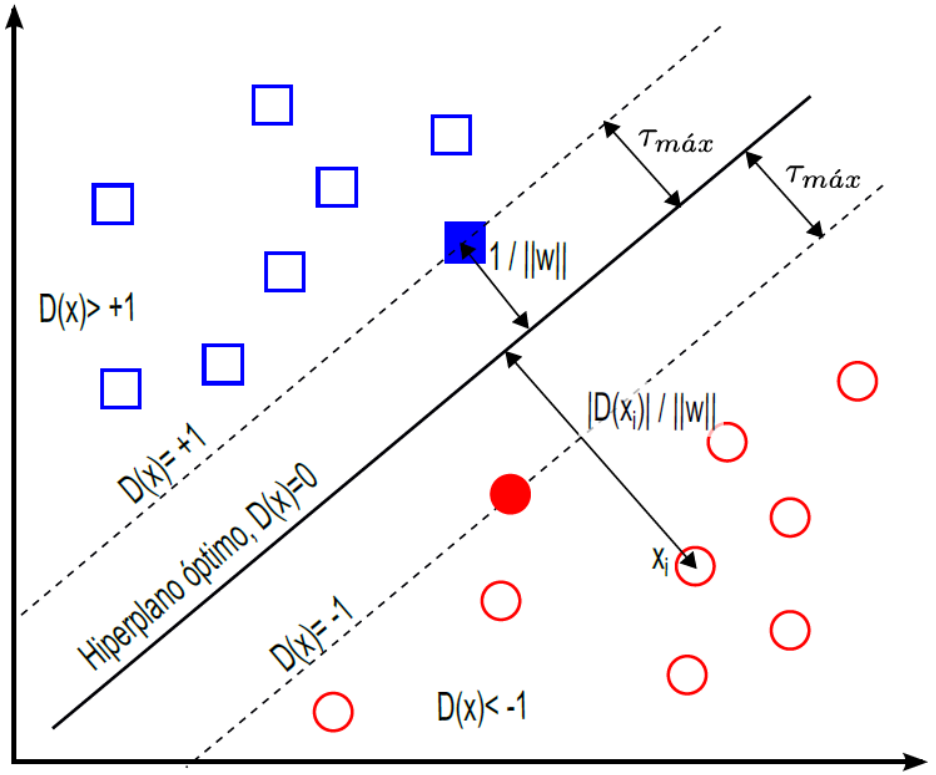
SVM (DISTANCIA DE UN HIPERPLANO)

Por geometría, se sabe que la distancia entre un hiperplano de desaparición D (x) y un ejemplo X´ viene dada por: :

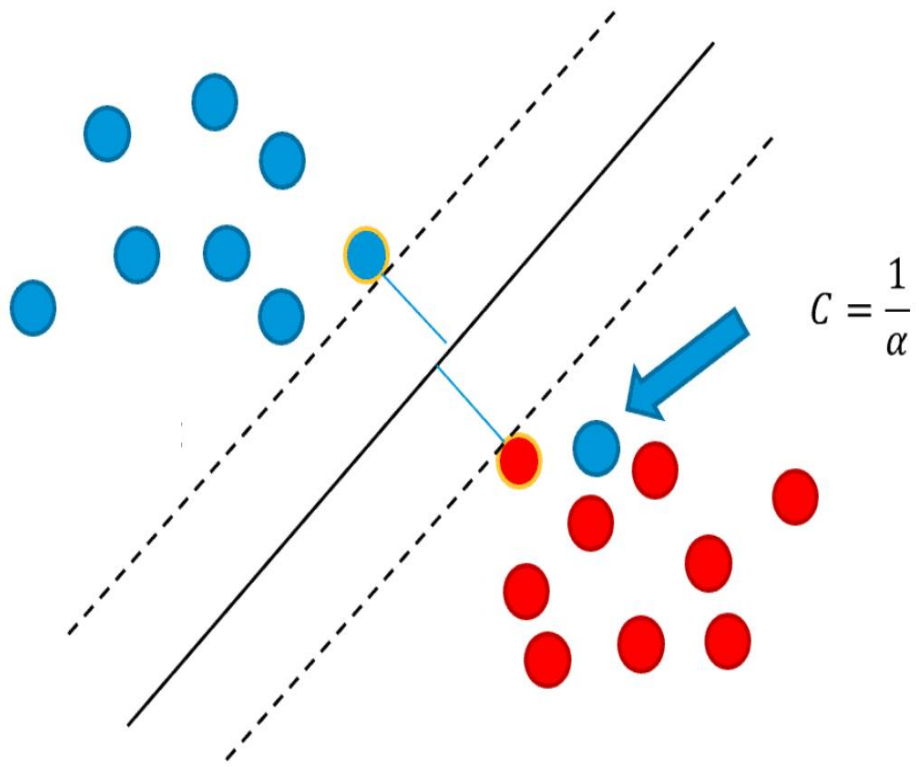
$$Distancia\left(D(\boldsymbol{x}), \boldsymbol{x}'\right) = \frac{|D(\boldsymbol{x}')|}{\|\boldsymbol{w}\|}$$

Por tanto, todos los ejemplos de entrenamiento cumplirán que la distancia de cada uno de ellos en el hiperplano de separación óptimo es mayor o igual que dicho margen:

$$\frac{y_i D(\boldsymbol{x}_i)}{\|\boldsymbol{w}\|} \geq \tau, \qquad i = 1, \ldots, n$$

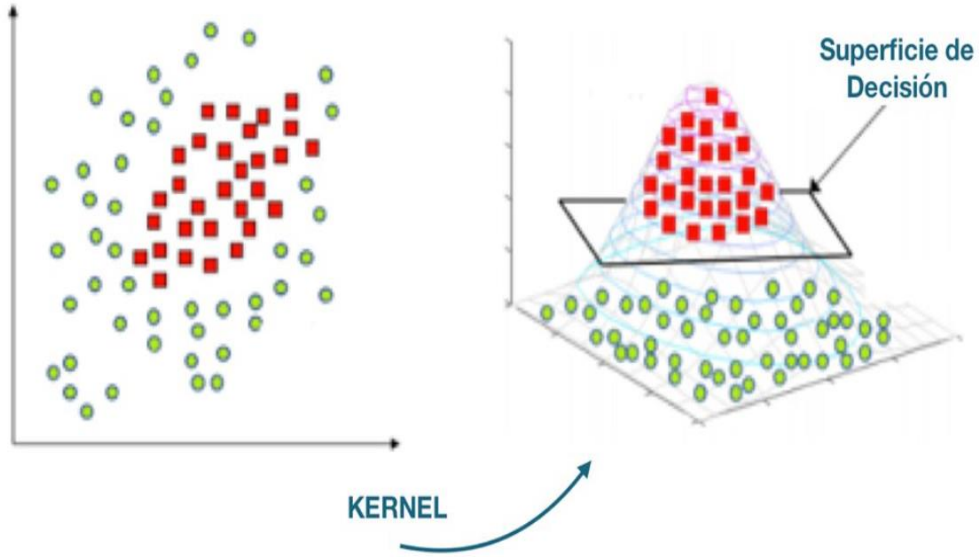


SVM (parámetro de Regularización)



- Es bastante frecuente que los datos tenga ruido, que no estén etiquetados perfectamente, o que el problema sea complejo para ser clasificado correctamente.
- Para estos casos, podemos decirle al algoritmo SVM, que preferimos que generalice bien para la mayoría de los casos, aunque algunos pocos casos del conjunto de entrenamiento no estén perfectamente clasificados.
- Tenga en cuenta que lo que normalmente vamos buscando es la construcción de modelos de aprendizaje automático es que generalicen bien.
- Para controlar la cantidad de regularización, podemos usar el hiper-parámetro *C*.
- A tener en cuenta: Un valor de C muy grande conlleva a modelos simples pero podría estar sobreajustando los datos de entrenamiento. Caso contrario produce modelos más complejos con gran propensión a clasificar mal.

SVM (Kernel)



- La mayoría de los eventos reales no son separables linealmente por lo que se dificulta la definición del Hiperplano de Separación Optimo.
- Para solucionar ese problema se utiliza un aprroach denominado Kernel.
- El approach kernel consiste en inventar una dimensión nueva en la que podamos encontrar un hiperplano para separar las clases.

Ejemplos de funciones kernel

Se presentan aquí algunos ejemplos de funciones kernel:

- Kernel lineal:

$$K_L(\boldsymbol{x}, \boldsymbol{x}') =< \boldsymbol{x}, \boldsymbol{x}' >$$

- kernel polinómico de grado-p:

$$K_P(\boldsymbol{x}, \boldsymbol{x}') = \left[\gamma < \boldsymbol{x}, \boldsymbol{x}' > +\tau\right]^p, \ \gamma > 0$$

- kernel gaussiano:

$$K_G(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\gamma \left\|\boldsymbol{x} - \boldsymbol{x}'\right\|^2\right) \equiv \exp\left(-\gamma < \left(\boldsymbol{x} - \boldsymbol{x}'\right), \left(\boldsymbol{x} - \boldsymbol{x}'\right) >\right), \ \gamma > 0$$

- kernel sigmoidal:

$$K_S(\boldsymbol{x}, \boldsymbol{x}') = \tanh(\gamma < \boldsymbol{x}, \boldsymbol{x}' > +\tau)$$

SVM

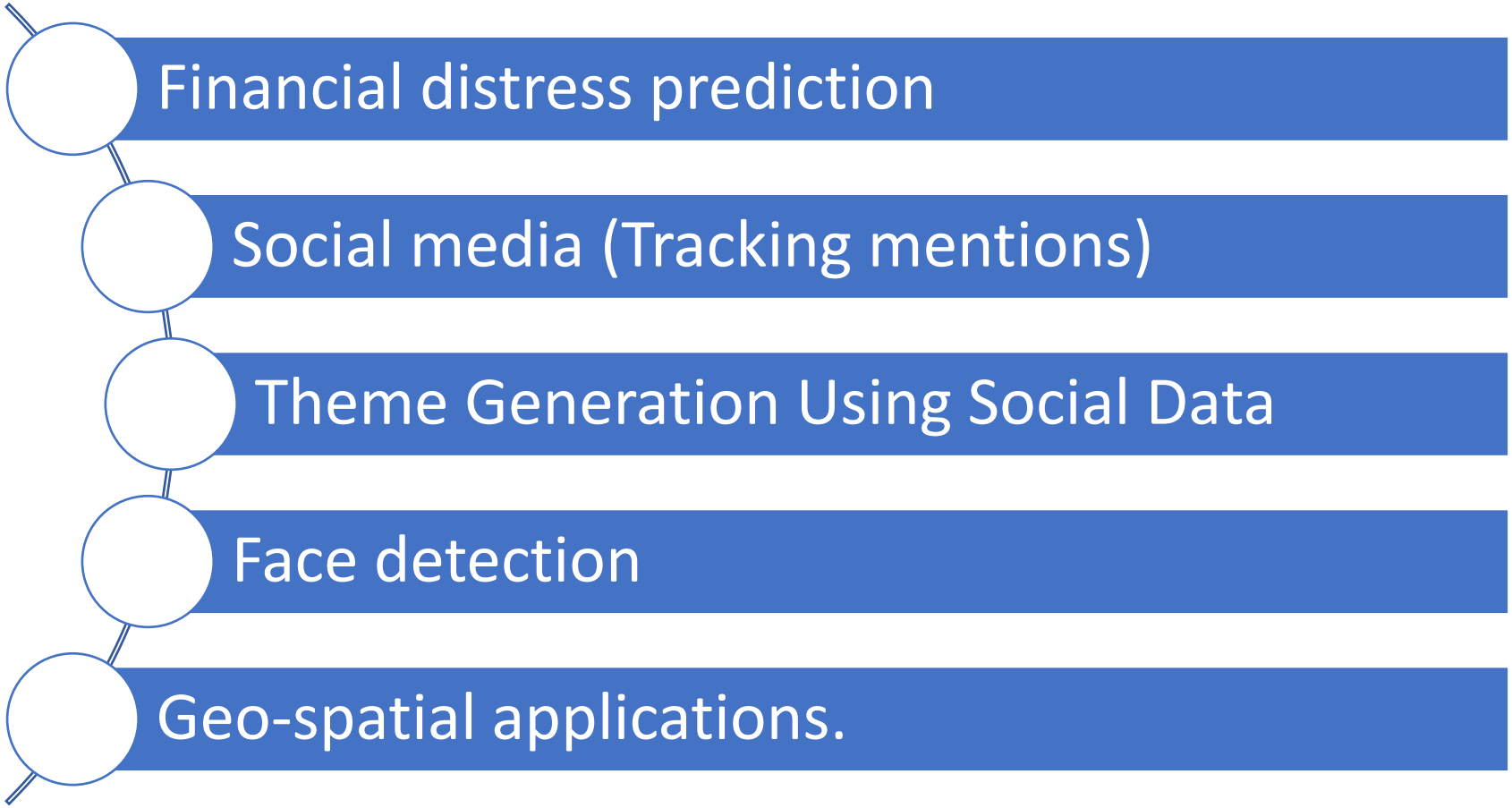
VENTAJAS

- ✓ Los clasificadores de Máquinas de Vectores de Soporte ofrecen una buena precisión y realizan predicciones más rápidas en comparación con el algoritmo de Naive Bayes.
- ✓ También utilizan menos memoria porque utilizan un subconjunto de puntos de entrenamiento en la fase de decisión.
- ✓ Este algoritmo funciona bien con un claro margen de separación y con un espacio dimensional elevado.

DESVENTAJAS

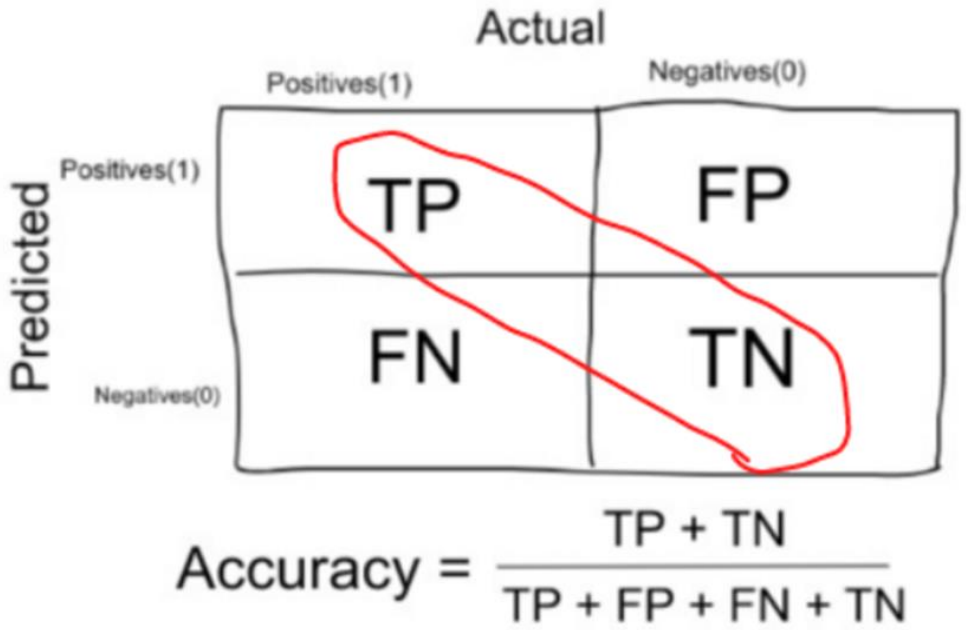
- ✓ Las Máquinas de Vectores de Soporte no son adecuadas para grandes conjuntos de datos debido a su alto tiempo de formación y también requiere más tiempo de formación en comparación con Naive Bayes.
- ✓ Funciona mal con clases superpuestas y también es sensible al tipo de núcleo utilizado.

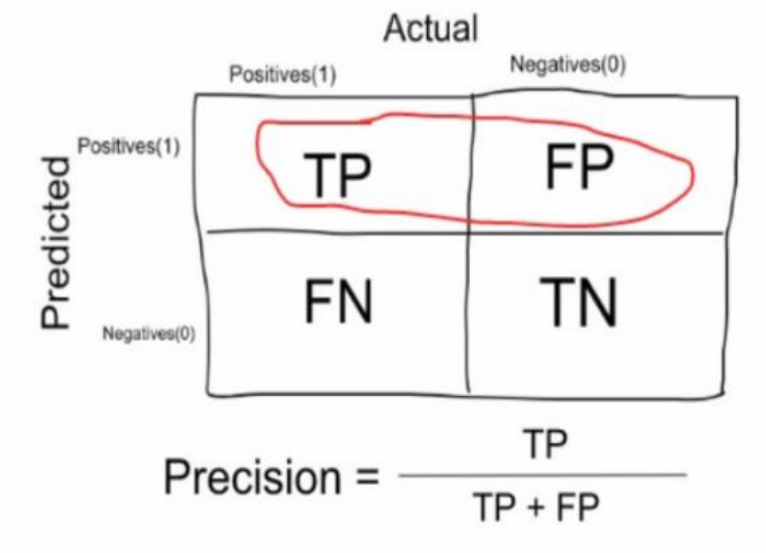
Aplicaciones del algoritmo SVM



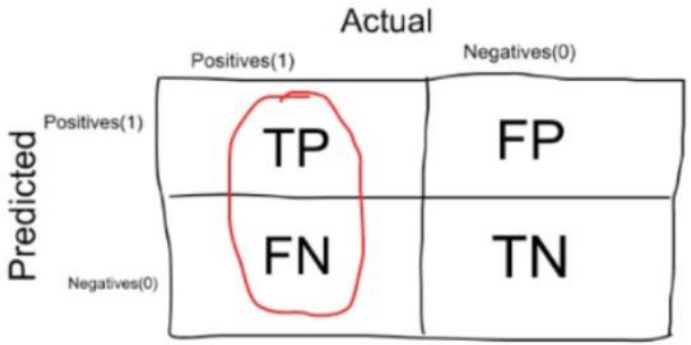
Métricas

1. **True Positives (TP):** cuando la clase real del punto de datos era 1 (Verdadero) y la predicha es también 1 (Verdadero)
2. **Verdaderos Negativos (TN):** cuando la clase real del punto de datos fue 0 (Falso) y el pronosticado también es 0 (Falso).
3. **False Positives (FP):** cuando la clase real del punto de datos era 0 (False) y el pronosticado es 1 (True).
4. **False Negatives (FN):** Cuando la clase real del punto de datos era 1 (Verdadero) y el valor predicho es 0 (Falso).

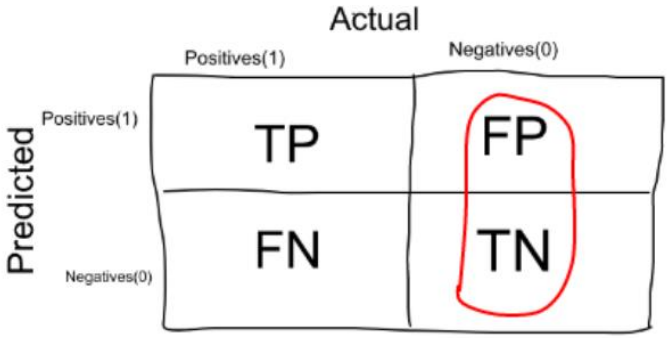




$$\text{Precision} = \frac{TP}{TP + FP}$$



$$\text{Recall} = \frac{TP}{TP + FN}$$



$$\text{Specificity} = \frac{TN}{TN + FP}$$

Métricas

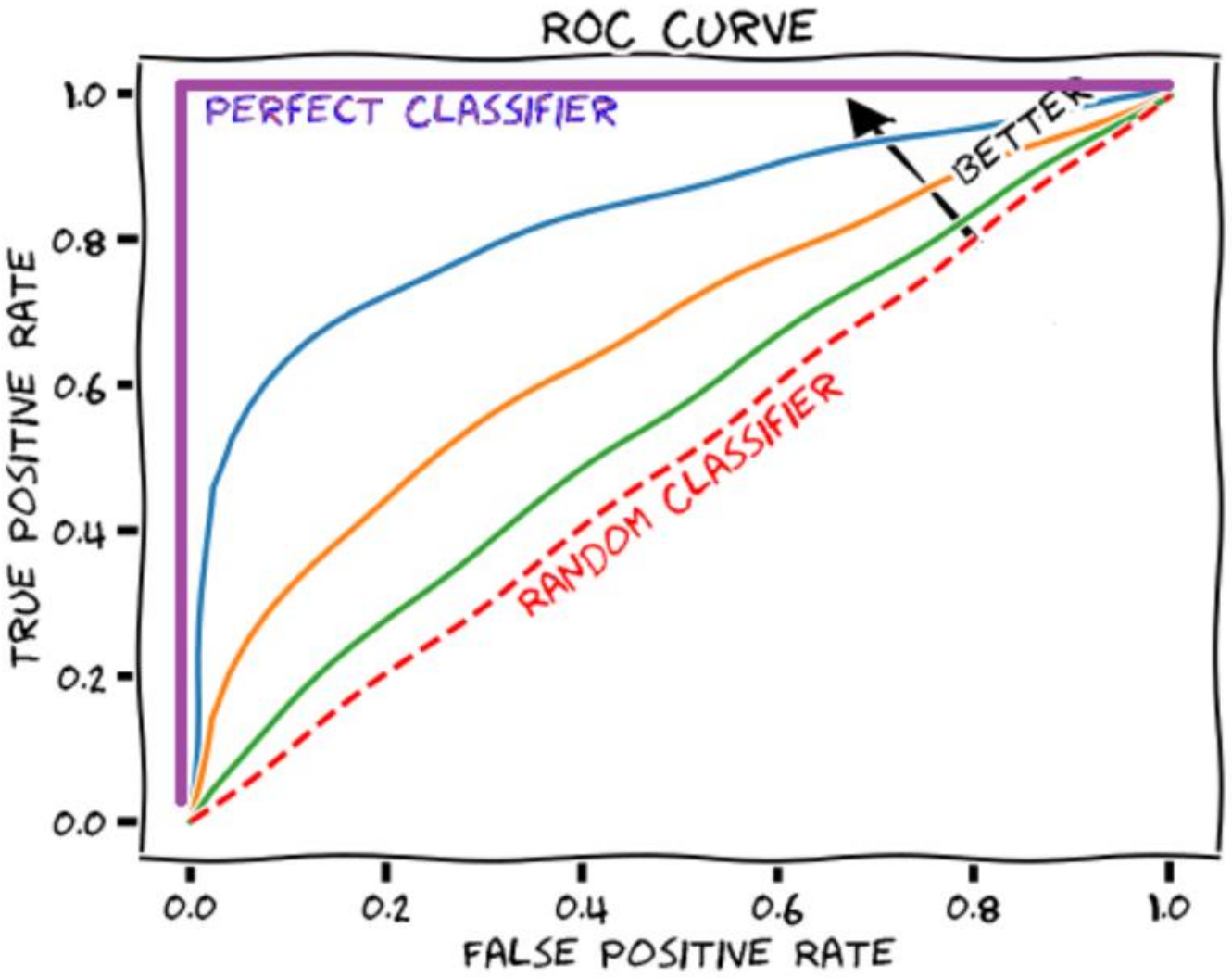
Métricas

Curva ROC

ROC es un acrónimo para Receiver Operating Characteristic (Característica Operativa del Receptor). Es una gráfica que enfrenta el ratio de falsos positivos (eje x) con el ratio de falsos negativos (eje y).

La curva ROC es útil por dos principales motivos:

- Permite comparar diferentes modelos para identificar cual otorga mejor rendimiento como clasificador
- El área debajo de la curva (AUC) puede ser utilizado como resumen de la calidad del modelo



Bibliotecas

Manipulación y análisis de datos

Creación de vectores y matrices
Colección de funciones matemáticas

Generación de gráficos

Biblioteca de visualización basada en Matplotlib

Biblioteca de aprendizaje automático