Introducción al cálculo numérico y ciencia de datos

Clase 5

Mg. Víctor Viana

Dr. Diego Passarella

Tacuarembó – abril de 2023

EDUCACION PERMANENTE
Universidad de la República

SEDE TACUAREMBÓ
CENUR NORESTE
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Agenda

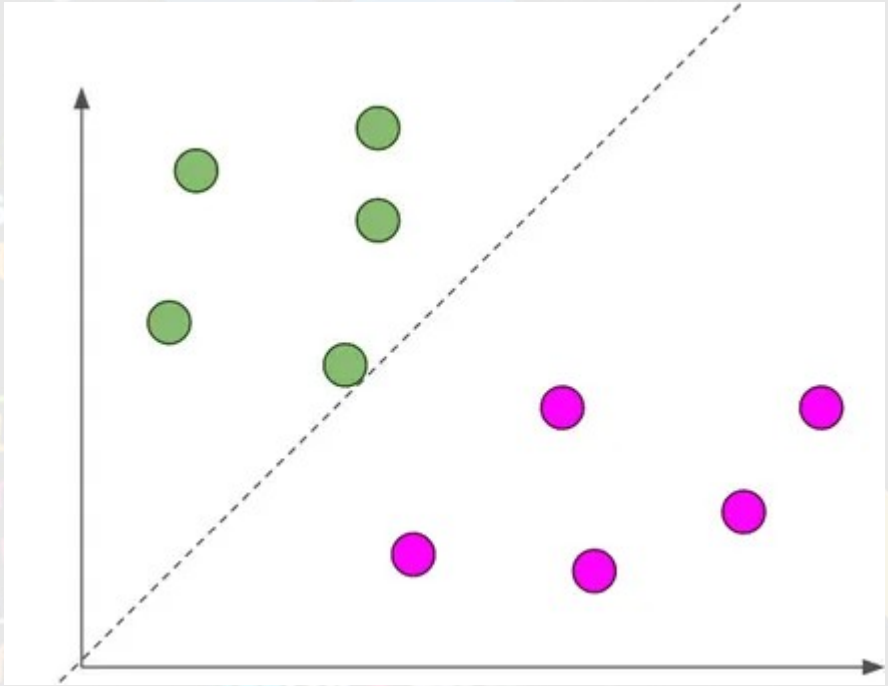
- SVM
- Definición
- Fundamentos matemáticos
- Aplicaciones
- Decision Tree
- Definición
- Aplicaciones

Support Vector Machine

- SVM es uno de los algoritmos de aprendizaje automático supervisado más populares y versátiles.
- Se utiliza tanto para tareas de **clasificación** como de regresión.
- Normalmente se prefiere para conjuntos de datos de tamaño medio y pequeño.
- El objetivo principal de SVM es encontrar el **hiperplano** óptimo que separa linealmente los puntos de datos en dos componentes maximizando el margen.

Support Vector Machine

La línea de puntos es el **hiperplano** que separa los puntos de las clases verde y fucsia.



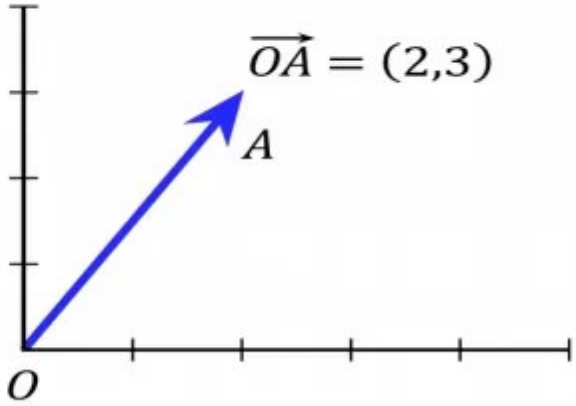
https://ankitnitjsr13.medium.com/math-behind-support-vector-machine-svm-5e7376d0ee4d

Álgebra lineal básica - Vectores

Los vectores son entidades matemáticas que tienen magnitud y dirección. Un punto en el plano 2D puede representarse como un vector entre el origen y ese punto.

Fig.-1

$\vec{OA}$ is a vector and length between O and A is its magnitude.



https://espanol.libretexts.org/Fisica/Libro%3A_F%C3%ADsica_(sin_l%C3%ADmites)/3%3A_Cinem%C3%A1tica_bidimensional/3.2%3A_Vectores

Longitud de los vectores

La longitud de los vectores también se denomina norma. Indica a qué distancia se encuentran los vectores del origen.

Length of vector $x(x_1, x_2, x_3)$ *is calculated as :*

$$\|x\| = \sqrt{x_1^2 + x_2^2 + x_3^2}$$

https://ankitnitjsr13.medium.com/math-behind-support-vector-machine-svm-5e7376d0ee4d

Dirección de los vectores

Direction of vector $x(x_1, x_2, x_3)$ *is calculated as:*

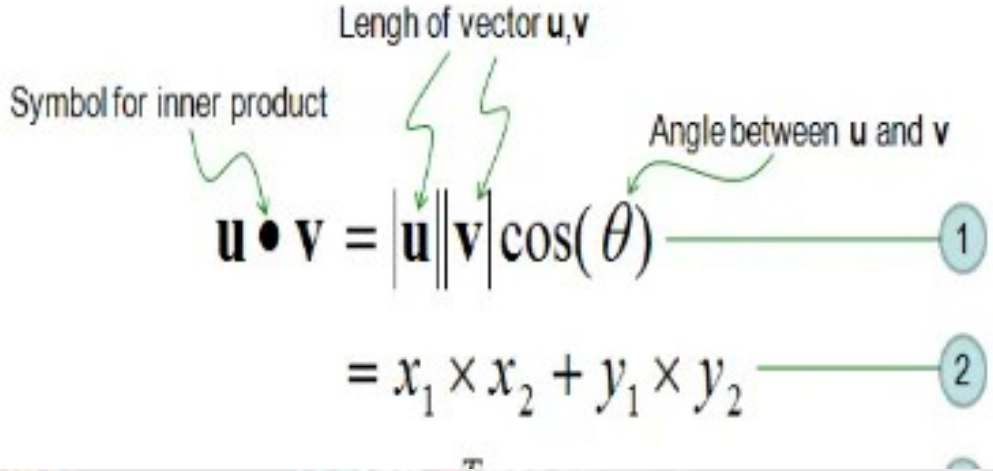
$$\left\{ \frac{x_1}{\|x\|}, \frac{x_2}{\|x\|}, \frac{x_3}{\|x\|} \right\}$$

https://ankitnitjsr13.medium.com/math-behind-support-vector-machine-svm-5e7376d0ee4d

Producto Vectorial

El producto punto entre dos vectores es una cantidad escalar. Indica la relación entre dos vectores.

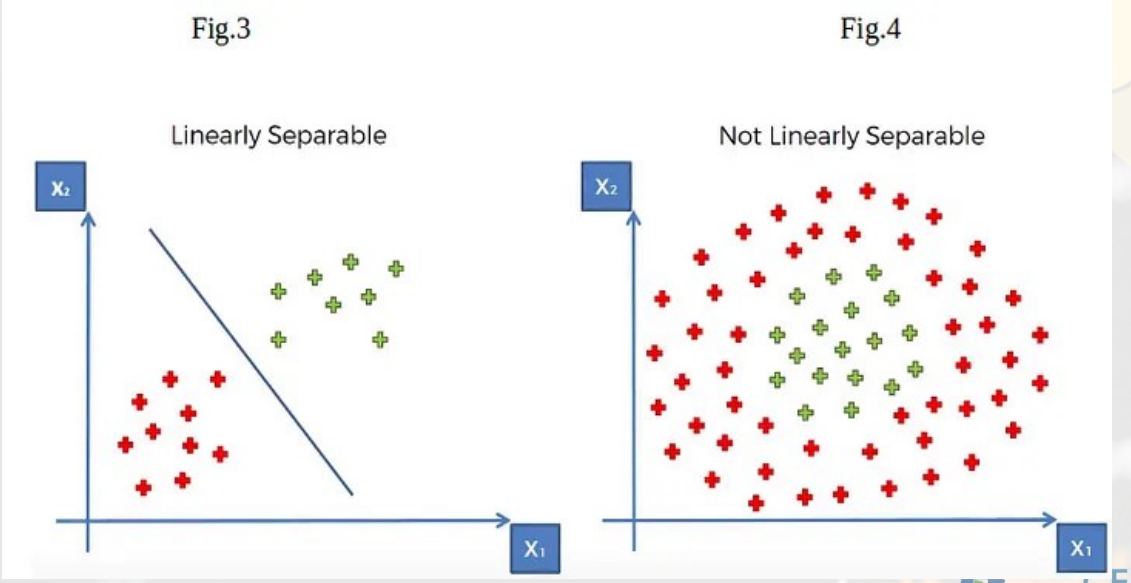
Two vectors u and v and their dot product is calculated as:



https://ankitnitjsr13.medium.com/math-behind-support-vector-machine-svm-5e7376d0ee4d

Hiperplano

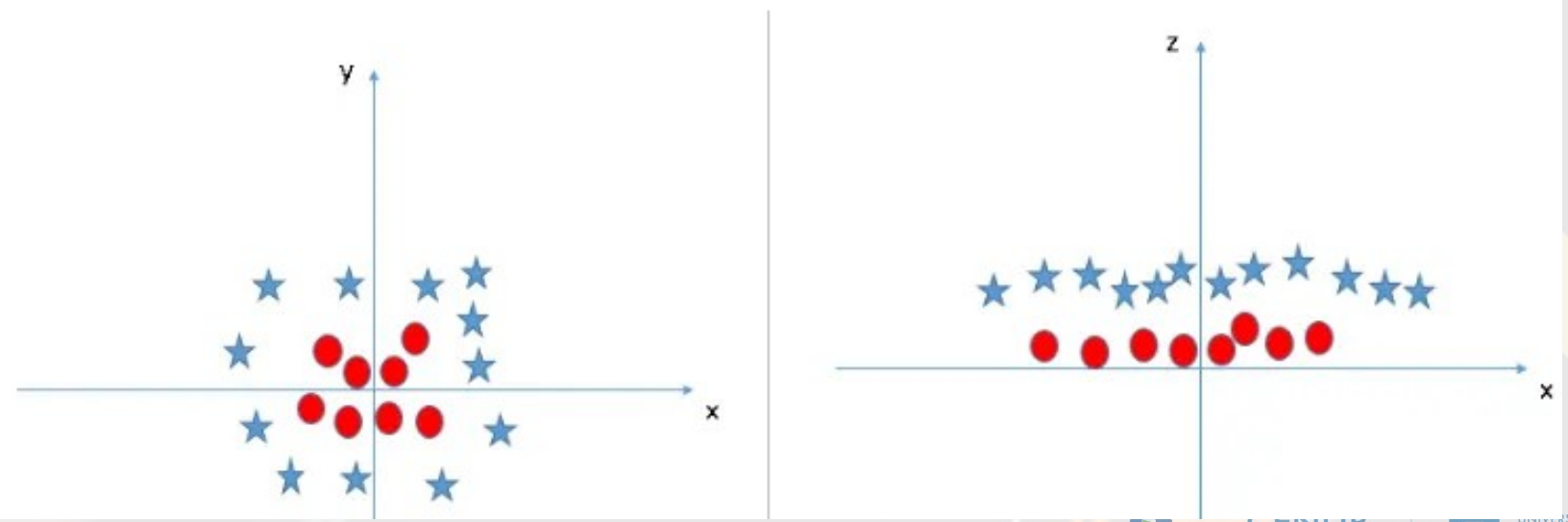
Es un plano que divide linealmente los puntos de datos n-dimensionales en dos componentes. En el caso de 2D, el hiperplano es una línea, en el caso de 3D es un plano, también llamado línea n-dimensional.



https://medium.com/@suvigya2001/the-gaussian-rbf-kernel-in-non-linear-svm-2fb1c822aae0

¿Qué pasa si los puntos no son linealmente separables?

- Este tipo de situación se da muy a menudo en el mundo del aprendizaje automático, ya que los datos en bruto son siempre no lineales.
- Añadimos una dimensión extra a los puntos de datos para hacerlos separables.

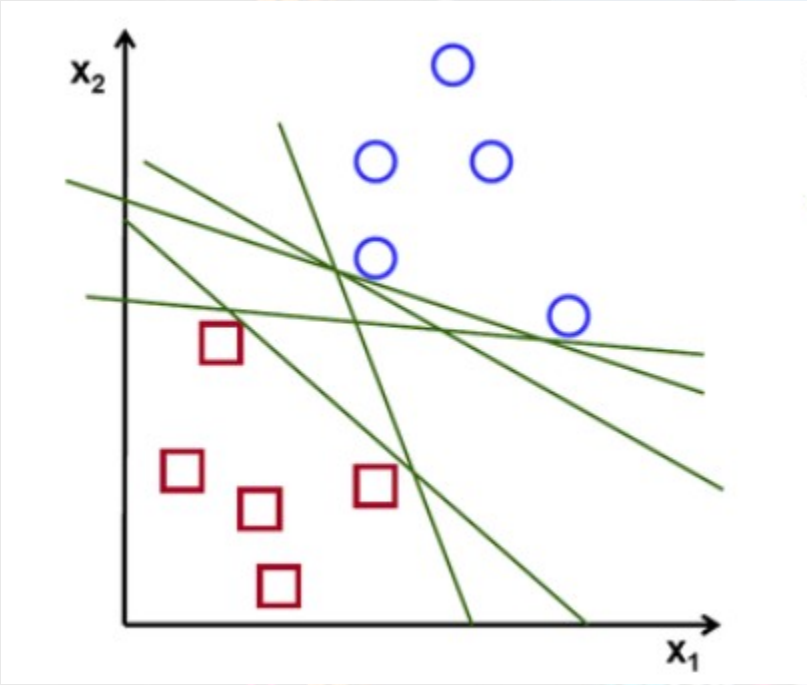


https://ankitnitjsr13.medium.com/math-behind-support-vector-machine-svm-5e7376d0ee4d

Hiperplano optimo

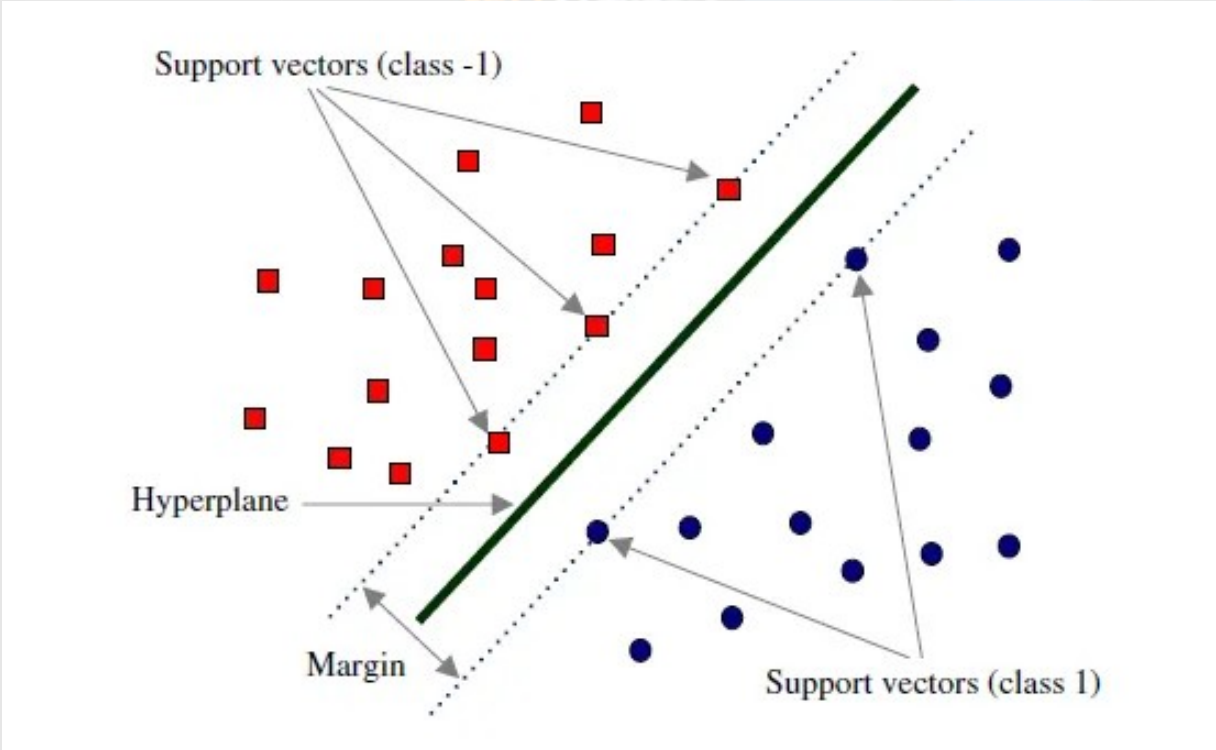
- El **hiperplano óptimo** es uno que divide los puntos de mejor manera.
- Si se elige un hiperplano sub-óptimo, no hay duda de que después de un número de iteraciones de entrenamiento, el error de entrenamiento disminuirá, pero durante la prueba, cuando una instancia no vista llegue, resultará en un alto error de prueba.
- En ese caso, es necesario elegir un plano óptimo para obtener una buena precisión.

Hiperplano optimo(II)



https://idus.us.es/bitstream/handle/11441/43808/Mart%C3%ADn%20Guare%C3%B1o%2C%20Juan%20Jos%C3%A9%20TFG.pdf?sequence=1&isAllowed=y

¿Cómo elegir el hiperplano óptimo?



https://www.researchgate.net/publication/264420380_INCORPORATION_OF_A_LANGUAGE_MODEL_INTO_A_BRAIN_COMPUTER_INTERFACE_BASED_SPELLER/figures?lo=1

¿Cómo elegir el hiperplano óptimo?(II)

- La distancia entre el hiperplano y el hiperplano óptimo se conoce como margen, y los puntos de datos más cercanos se conocen como vectores de soporte.
- El margen es un área que no contiene ningún punto de datos.
- Elegir el hiperplano óptimo implica elegir uno entre el conjunto de hiperplanos que se encuentre a la mayor distancia de los puntos de datos más cercanos.
- Si el hiperplano óptimo está muy cerca de los puntos de datos, entonces el margen será muy pequeño y generalizará bien para los datos de entrenamiento, pero cuando lleguen datos no vistos, fallará en generalizar bien como se explicó anteriormente.
- Así que nuestro objetivo es maximizar el margen para que nuestro clasificador sea capaz de generalizar bien para los casos no vistos.

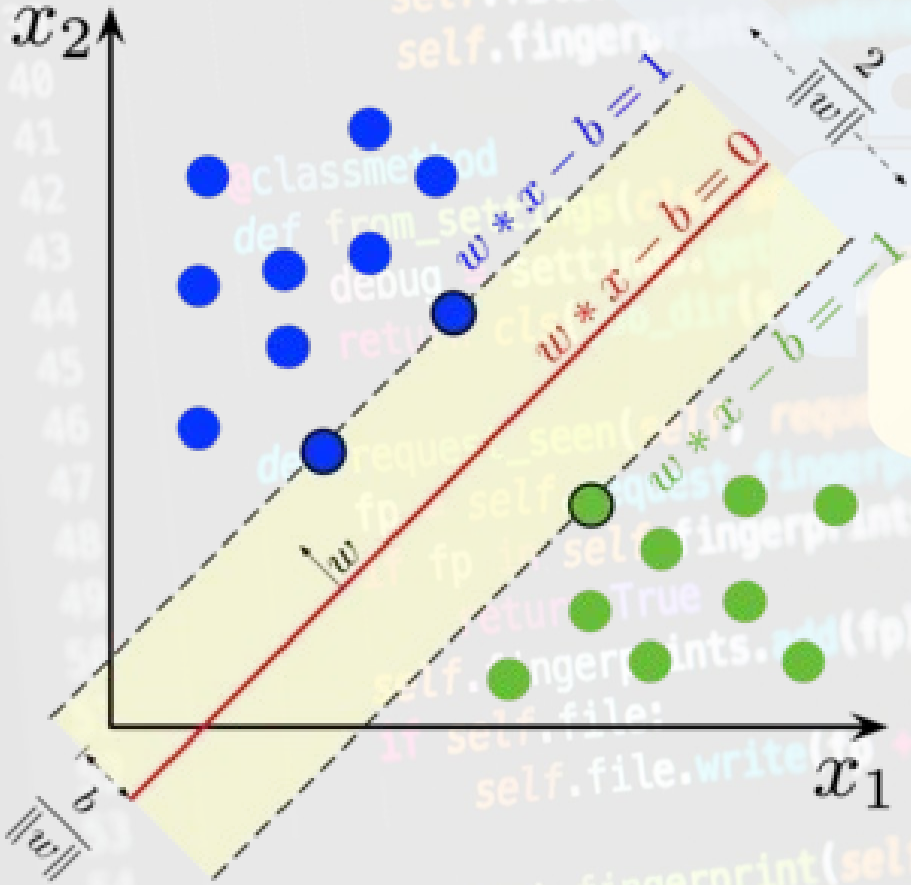
¿Cómo elegir el hiperplano óptimo?(III)

- Supongamos que tenemos n puntos de entrenamiento, cada observación i tiene p características (es decir, x_i tiene p dimensiones), y está en dos clases y_i=-1 o y_i = 1.

- Supongamos que esas dos clases de observaciones son linealmente separables.

- Podemos dibujar un hiperplano a través de nuestro espacio de características de tal manera que todos los casos de una clase están en un lado del hiperplano, y todos los casos de la otra clase están en el lado opuesto.

¿Cómo elegir el hiperplano óptimo?(III)



Wikipedia

SEDE TACUAREMBÓ
CENUR NORESTE
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

¿Cómo elegir el hiperplano óptimo?(IV)

Dicho hiperplano deberá cumplir las siguientes desigualdades:

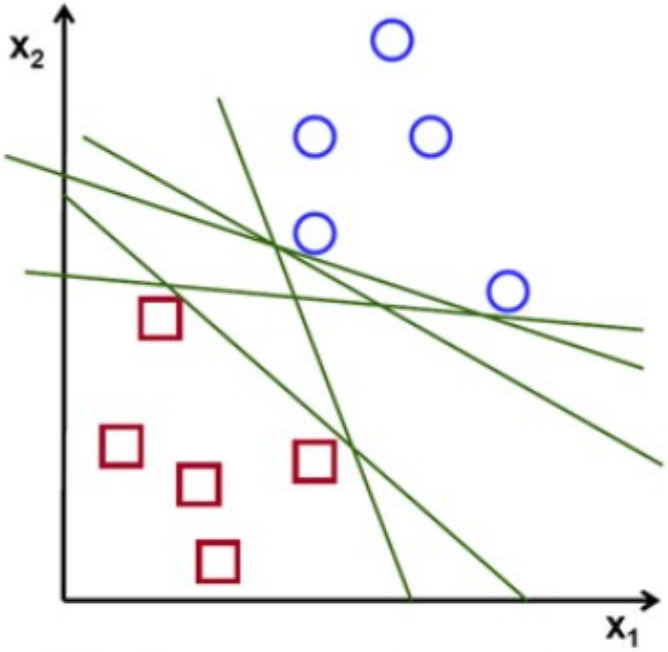
$$w \cdot x + b \geq 0 \, , si \, y = 1 \, \forall \, i = 1 \ldots n$$

$$w \cdot x + b \leq 0 \, , si \, y = -1 \, \forall \, i = 1 \ldots n$$

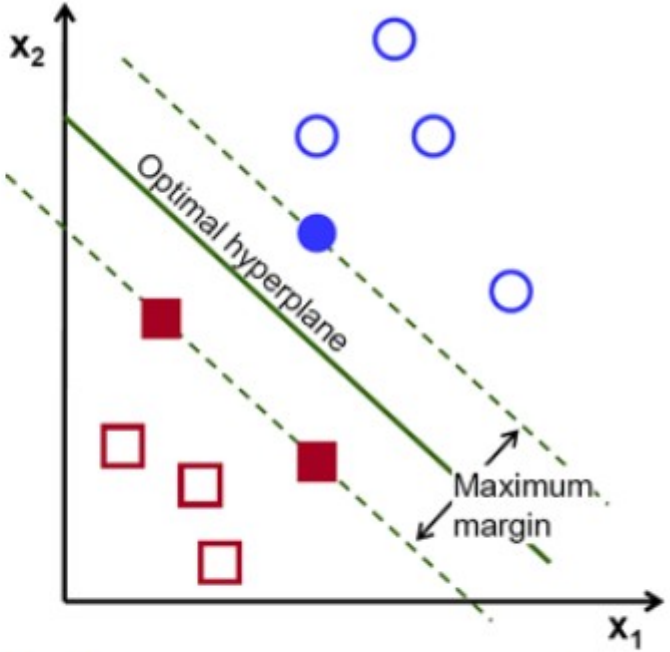
Escrito de forma mas compacta:

$$y(w \cdot x) + b \geq 0 \, , \forall \, i = 1 \ldots n$$

Hiperplano optimo(V)



(a) En esta representación apreciamos la NO unicidad de solución.



(b) Mientras que en esta, una vez impuesto el margen máximo, el hiperplano es único.

https://idus.us.es/bitstream/handle/11441/43808/Mart%C3%ADn%20Guare%C3%B1o%2C%20Juan%20Jos%C3%A9%20TFG.pdf?sequence=1&isAllowed=y

Hiperplano optimo(VI)

- El hiperplano que permite separar las dos clases no suele ser único.
- La selección de un hiperplano de entre todos los posibles hiperplanos de separación se realizará a partir del concepto de **margen**, que se define como la distancia mínima entre dicho hiperplano y el ejemplo más cercano a cada clase.

Hiperplano optimo(VI)

- Maximizar el margen es equivalente a minimizar la norma de *w*.

$$min \frac{1}{2}\|w\|^2$$

$$s.a.$$

$$y_i\left(w \cdot x_i\right)+b \geq 1, \forall i=1...n$$

Caso Lineal No Separable

- En los problemas reales encontrar un conjunto con dos clases totalmente separables es escasamente probable, entre otras cosas por la existencia de ruido en los datos.
- La idea para tratar con este tipo de casos con ruido es introducir un conjunto de variables reales y positivas, variables artificiales, ξi, i = 1,… n, de forma que permitan algunos ejemplos no separables, es decir:

$$min \frac{1}{2}\|w\|^2 + C\sum_{\square}^{\square} \xi_i$$
$$s.a.$$
$$y_i(w \cdot x_i) + b \geq 1 - \xi_i, \forall i = 1...n$$
$$\xi_i \geq 0, \forall i = 1...n$$

Ejemplo – Básico

Ver ejemplo en el Notebook: icncd_clase5_SVM_basico.ipynb

Ejemplo – Básico con scikit-learn

Ver ejemplo en el Notebook: icncd_clase5_SVM_skl_1.ipynb

Ejemplo – Iris flowers

Ver ejemplo en el Notebook: [icncd_clase5_SVM_skl_2.ipynb]

Ejemplo – Pronóstico del clima

Ver ejemplo en el Notebook: [icncd_clase5_SVM_clima.ipynb]

Arboles de Decisión

- El algoritmo de **árbol de decisión** (Decision Tree, DT) pertenece a la familia de algoritmos de aprendizaje supervisado.
- El objetivo de utilizar un DT es crear un modelo de entrenamiento que pueda utilizarse para predecir la clase o el valor de la variable objetivo mediante el aprendizaje de reglas de decisión sencillas inferidas a partir de datos anteriores.
- En los DT, para predecir la etiqueta de clase de un registro, partimos de la raíz del árbol. Comparamos los valores del atributo raíz con el atributo del registro. Basándonos en la comparación, seguimos la rama correspondiente a ese valor y saltamos al siguiente nodo.

Tipos de árboles de decisión

- **DT de variable categórica**: tiene una variable objetivo categórica
- **DT de Variable Continua**: tiene una variable objetivo continua.

Tipos de árboles de decisión(II)

- Ejemplo: Supongamos que tenemos un problema para predecir si un cliente pagará su prima de renovación con una compañía de seguros (sí/no).
- Sabemos que los ingresos de los clientes son una variable importante, pero la compañía de seguros no dispone de datos sobre los ingresos de todos los clientes.
- Ahora bien, como sabemos que se trata de una variable importante, podemos construir un árbol de decisión para predecir los ingresos de los clientes en función de la ocupación, el producto y otras variables.
- En este caso, estamos prediciendo valores para las variables continuas.

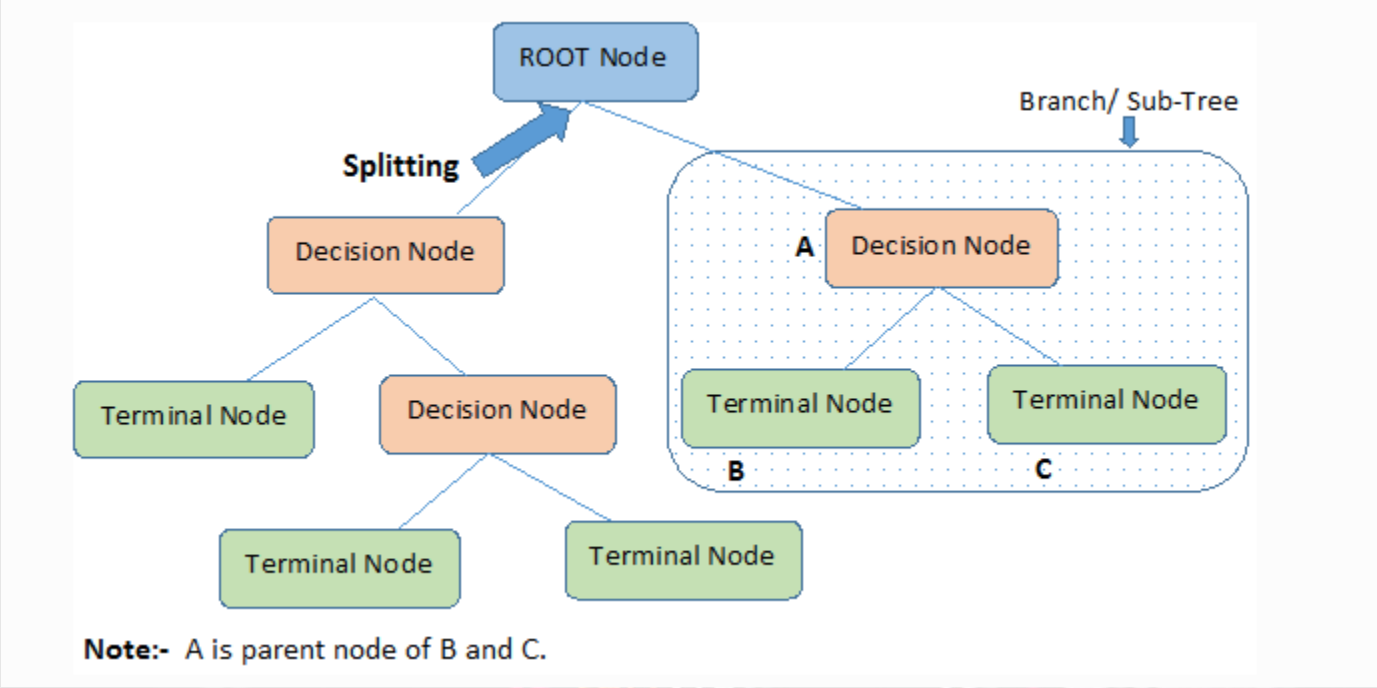
Terminología relacionada con los DT

- **Nodo Raíz**: Representa a toda la población o muestra y ésta a su vez se divide en dos o más conjuntos homogéneos.
- **División**: Es un proceso de división de un nodo en dos o más subnodos.
- **Nodo de decisión**: Cuando un subnodo se divide en otros subnodos, se denomina nodo de decisión.
- **Nodo terminal**: Los nodos que no se dividen se denominan nodos hoja o nodos terminales.

Terminología relacionada con los DT(II)

- **Poda**: Cuando eliminamos subnodos de un nodo de decisión, este proceso se denomina poda. Se puede decir que es el proceso contrario a la división.
- **Rama / Subárbol**: Una subsección de todo el árbol se denomina rama o subárbol.
- **Nodo padre e hijo**: Un nodo dividido en subnodos se denomina nodo padre de subnodos, mientras que los subnodos son hijos de un nodo padre.

Terminología relacionada con los DT(II)



https://www.kdnuggets.com

Suposiciones

- Al principio, todo el conjunto de entrenamiento se considera la raíz.
- Se prefiere que los valores de las características sean categóricos. Si los valores son continuos, se discretizan antes de construir el modelo.
- Los registros se distribuyen recursivamente en función de los valores de los atributos.
- El orden de colocación de los atributos como raíz o nodo interno del árbol se realiza utilizando algún enfoque estadístico.

¿Cómo trabajan los DT?

- Los árboles de decisión utilizan múltiples algoritmos para decidir dividir un nodo en dos o más subnodos. La creación de subnodos aumenta la homogeneidad de los subnodos resultantes.
- La selección del algoritmo también se basa en el tipo de variables objetivo. Veamos algunos algoritmos utilizados en DT:
  - ID3 → (extensión de D3)
  - C4.5 → (sucesor de ID3)
  - CART → (árbol de clasificación y regresión)
  - CHAID → (detección automática de interacciones Chi-cuadrado Realiza divisiones multinivel al calcular árboles de clasificación)
  - MARS → (splines de regresión adaptativa multivariante)

ID3

- El algoritmo ID3 construye árboles de decisión utilizando un enfoque de búsqueda greedy descendente a través del espacio de posibles ramas sin retroceso.
- Un algoritmo greedy, como su nombre indica, siempre elige la opción que le parece mejor en ese momento.

Medidas de selección de atributos

- Si el conjunto de datos consta de N atributos, decidir qué atributo colocar en la raíz o en distintos niveles del árbol como nodos internos es un paso complicado.
- Seleccionar aleatoriamente cualquier nodo para que sea la raíz no puede resolver el problema.
- Si seguimos un enfoque aleatorio, puede darnos malos resultados con baja precisión.

Medidas de selección de atributos(II)

- Para resolver este problema de selección de atributos se sugiere utilizar algunos criterios como :
  - Entropía,
  - Ganancia de información,
  - Índice de Gini,
  - Ratio de ganancia,
  - Reducción de la varianza
  - Chi-cuadrado

Ejemplo – Decisión sobre la capacidad de pago

Ver ejemplo en el Notebook: icncd_clase5_dt_ejemplo1.ipynb