Machine Learning

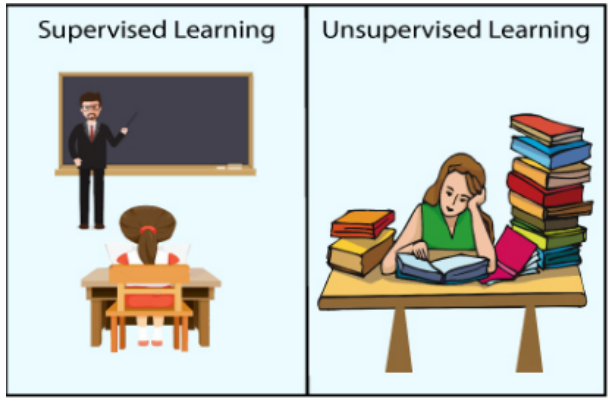
Rama de la Inteligencia Artificial (AI), basada en la idea de que los sistemas pueden aprender de datos, identificar patrones y tomar decisiones

Tipo de algoritmos

Supervisados

Son aquellos donde se tiene etiqueta o variable respuesta. Se fundamenta en crear modelos con información histórica para luego poder predecir en el futuro. Uso de conjuntos de datos etiquetados para entrenar algoritmos que clasifiquen datos o predigan resultados.

- Regresión
- Clasificación



No supervisados

Es un tipo de aprendizaje automático en el que se utilizan conjuntos de datos no etiquetados para encontrar todo tipo de patrones desconocidos en los datos. No hay una variable objetivo (variable de salida)

- Clustering
- Reglas de asociación
- Reducción de Dimensionalidad

Existen múltiples algoritmos, en este documento comentaremos aquellos más populares

Supervisados

Regresión

- Requiere la predicción de una variable continua.
- Puede tener como entrada valores continuos o discretos.
- Un problema con múltiples variables de entrada a menudo se denomina regresión lineal múltiple.
- Predicen valores numéricos. Es decir, la variable target en un problema de regresión es de tipo cuantitativa.

Ejemplos

- Predicción comportamiento de clientes
- Pronósticos de demanda
- Pronóstico de Revenue y Profit

La regresión es una técnica de aprendizaje supervisado que ayuda a encontrar la correlación entre variables, y nos permite predecir la variable de salida continua basada en uno o más variables predictoras.

Clasificación

- Requiere variable objetivo con dos o más clases.
- Puede tener variables de entrada discretas o continuas.
- Un problema con dos clases se denomina problema de clasificación binaria y con más de dos, clasificación multiclase.
- El objetivo final es predecir la clase más probable de un elemento, en función de un conjunto de variables de entrada. Para este tipo de algoritmos, la variable target o respuesta, es una variable de tipo categórica.

Ejemplos

- Clasificación de imágenes
- Categorización de productos
- Detección de fraude
- Detección de spam

Regresión logística

Árbol de decisión

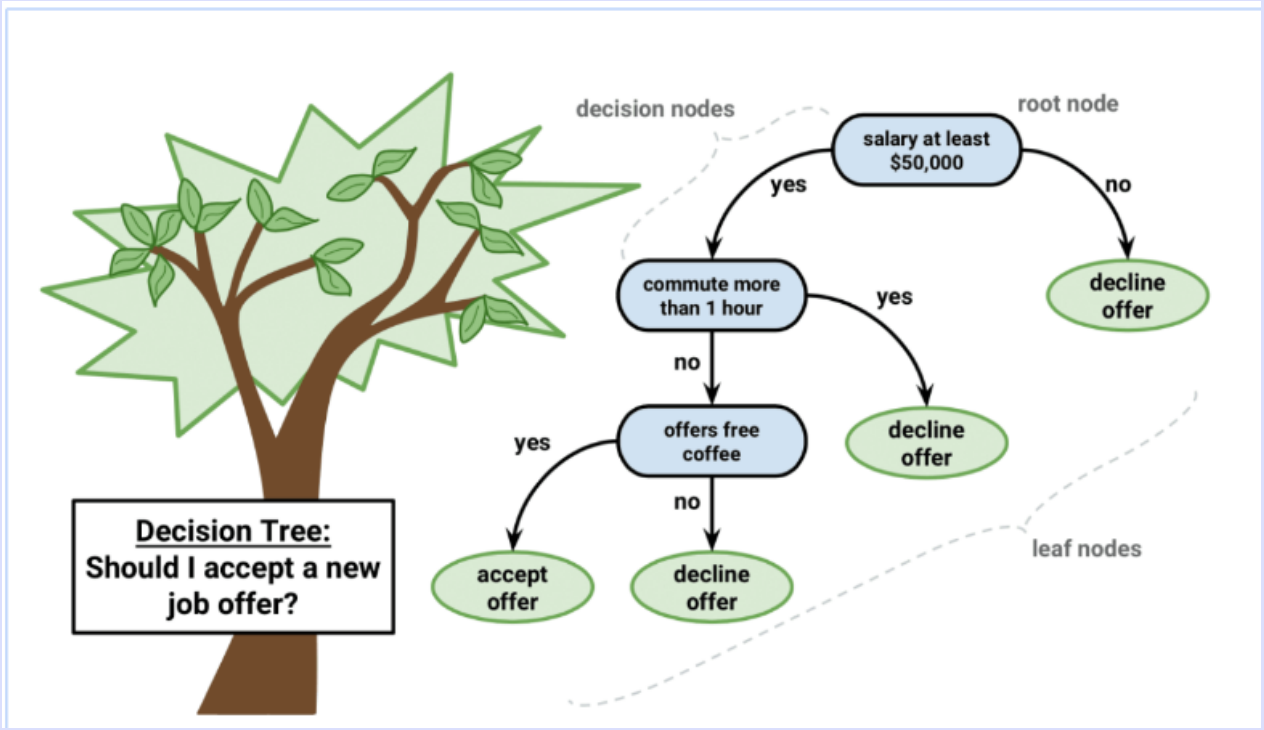
K-Nearest-Neighbor

Aclaración: Es importante recordar, que muchos de los algoritmos que se utilizan para clasificación también pueden utilizarse para problemas de regresión, como ser por ejemplo: Árboles o KNN. Cabe aclarar que por supuesto, utilizan diferentes librerías. ☺

Supervisados - Clasificación

Árbol de decisión

- Aprenden de los datos generando reglas de tipo if-else.
- Separan los datos en grupos cada vez más pequeños de subsets de un dataset original.
- A cada división se la conoce con el nombre de nodo. Cuando un nodo no conduce a nuevas divisiones se le denomina hoja, para luego ser considerada como ramas del árbol.



Ventajas

- Relativamente robusto cuando la complejidad no es tan alta.
- Conjunto de reglas con booleanos, sus resultados son fáciles de entender e interpretar.

Desventajas

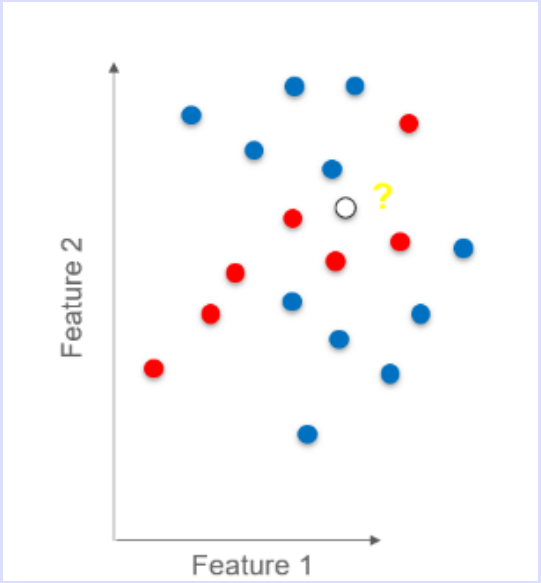
- Se ven influenciadas por los outliers, creando árboles con ramas muy profundas que no predicen bien para nuevos casos.
- Crear árboles demasiado complejos puede conllevar que no se adapten bien a los nuevos datos.
- Se pueden crear árboles sesgados si una de las clases es más numerosa que otra es decir, si hay desbalance de clases.

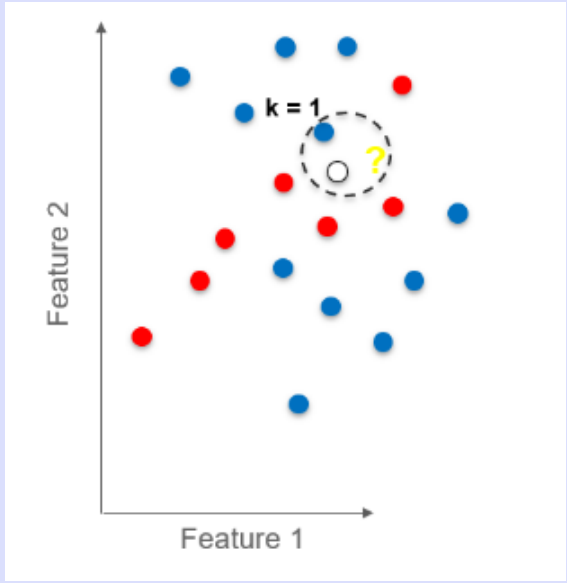
Supervisados - Clasificación

K-Nearest-Neighbor(Vecinos cercanos)

- Este algoritmo, puede usarse para clasificar nuevas muestras (valores discretos) o para predecir (regresión, valores continuos). Sirve esencialmente para clasificar valores, buscando los puntos de datos "más similares" (por cercanía) aprendidos en la etapa de entrenamiento del modelo y haciendo conjeturas de nuevos puntos basado en esa clasificación o regresión.
- El algoritmo clasifica cada dato nuevo en el grupo que corresponda, según tenga k vecinos más cerca de un grupo o de otro. Es decir, calcula la distancia del elemento nuevo a cada uno de los existentes, y ordena dichas distancias de menor a mayor para ir seleccionando el grupo al que pertenecer.

- Dada una nueva instancia, de la cual no sabemos cuál es su clase, vamos a recurrir a sus vecinos cercanos para clasificarla. La pregunta seria entonces, ¿La clasificamos como rojo o como azul?





Si tomamos K=1, solo miraremos al vecino más cercano.
Aclaración: K es el nro de vecinos. La predicción será: Azul.

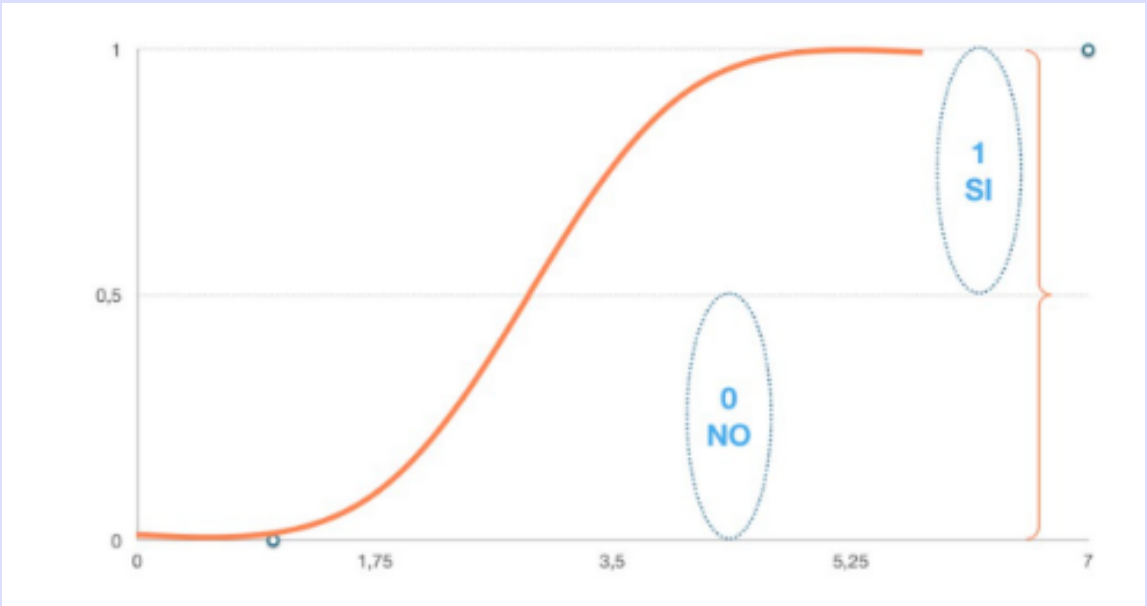
Supervisados - Clasificación

Regresión logística

- La Regresión Logística es una técnica de aprendizaje automático que proviene del campo de la estadística. Es un método para problemas de clasificación, en los que se obtienen un valor binario entre 0 y 1.
- Por ejemplo, un problema de clasificación es identificar si una operación dada es fraudulenta o no, asociándole una etiqueta "fraude" a unos registros y "no fraude" a otros. Entonces, la Regresión Logística describe y estima la relación entre una variable binaria dependiente y las variables independientes.

La Regresión Logística lleva el nombre de la función utilizada en el núcleo del método, la Función Logística es también llamada función Sigmoide. Esta función es una curva en forma de S que puede tomar cualquier número de valor real y asignar a un valor entre 0 y 1.

Si la curva va a infinito positivo la predicción se convertirá en 1, y si la curva pasa el infinito negativo, la predicción se convertirá en 0. Si la salida de la función Sigmoide es mayor que 0.5, podemos clasificar el resultado como 1 o SI, y si es menor que 0.5 podemos clasificarlo como 0 o NO.



No Supervisados - Clustering

Clustering

También conocidas como agrupamiento o segmentación, tienen como principal función encontrar una estructura o un patrón en una colección de datos no clasificados. Es decir, intentan encontrar grupos en los datos que compartan atributos en común. Ejemplos de algoritmos de cluster: k-means. clustering jerárquico, modelos de mixturas gaussianas, o algoritmos basados en densidad como DBSCAN.

Técnicas para codificación de categorías

Label Encoder (LE): utilizamos esta técnica cuando La variable categórica es ordinal y el número de categorías es bastante grande. En este útimo caso si usamos la codificación one-hot puede llevar a un alto consumo de memoria.

SAFETY-LEVEL (TEXT)	SAFETY-LEVEL (NUMERICAL)
None	0
Low	1
Medium	2
High	3
Very-High	4

One Hot Encoding: existe la manera de trabajar con variables categóricas haciéndolas variables dummy, a través del uso de la técnica de transformación de datos One Hot Encoding (OHE).

Color
Red
Red
Yellow
Green
Yellow

→

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

Clustering Jerárquico

En estos algoritmos se generan sucesiones ordenadas (jerarquías) de conglomerados. Puede ser agrupando clústers pequeños en uno más grande o dividiendo grandes clusters en otros más pequeños.

Jerárquicos aglomerativos: Inicialmente cada instancia es un clúster. Las estrategias aglomerativas parten de un conjunto de elementos individuales y van "juntando" los elementos que más se parezcan hasta quedarse con un número de clusters que se considere óptimo.

Jerárquicos divisivos: Inicialmente todas las instancias están en un solo clúster y luego se van dividiendo, tal cual su nombre lo indica. Las estrategias divisivas, parten del conjunto de elementos completos y se van separando en grupos diferentes entre sí, hasta quedarse con un número de clusters que se considere óptimo.

Cluster no Jerárquico

La cantidad de clústeres óptima se define de antemano, y los registros se asignan a los clústeres según su cercanía. Existen múltiples algoritmos de Tipo No Jerárquico, como ser por ejemplo: K - Means o DBSCAN.

No Supervisados - Reglas de asociación

Reglas de asociación

- Los algoritmos de reglas de asociación tienen como objetivo encontrar relaciones dentro un conjunto de transacciones, en concreto, items o atributos que tienden a ocurrir de forma conjunta. En este contexto, el término transacción hace referencia a cada grupo de eventos que están asociados de alguna forma, por ejemplo:
  - La cesta de la compra en un supermercado.
  - Los libros que compra un cliente en una librería.

- Se refiere a la identificación de los objetos que se encuentran juntos en un evento o registro dado.
- Una regla de asociación tiene un antecedente (lado izquierdo) y un consecuente (lado derecho). Ambos lados de la regla son un conjunto de elementos. Si el conjunto de elementos X es el antecedente y conjunto de elementos Y es el consecuente, entonces la regla de asociación se escribe como:

$$X \longrightarrow Y$$

- Las reglas no deben ser interpretadas como una relación causal, sino como una asociación entre dos o más elementos.

- Ejemplo: sistemas de recomendación utilizados por varias tiendas virtuales como por ejemplo Mercado libre, Amazon.

No Supervisados - Reducción de la dimensionalidad

Reducción de la dimensionalidad

Buscamos reducir la cantidad de features de un dataset, pero reteniendo la mayor cantidad de "información" posible.

Tenemos dos aplicaciones principales con esta técnica:

1- Eliminar variables

2- Encontrar grupos

¿Para qué lo aplicaríamos?

- Para enfrentar "La Maldición de la Dimensionalidad" es decir, tenemos tantos features que termina siendo algo negativo para nuestro modelo de ML.
- Reducir el input en un modelo de regresión o clasificación.
- Visualizar mucho mejor nuestros datos.
- Compresión de archivos.
- Detectar features relevantes en datasets o variables altamente correlacionadas.

Algoritmos de aplicación:

- PCA: Principal Component Analysis.
- Auto-Encoders con Redes Neuronales.
- MDS: Multidimensional scaling.

Principal Component Analysis.

- El método gira los datos de forma que, desde un punto de vista estadístico, no exista una correlación entre las características rotadas pero que conserven la mayor cantidad posible de la varianza de los datos originales.
- Es decir, el PCA reduce la dimensionalidad de un conjunto de datos proyectándose sobre un subespacio de menor dimensionalidad.

by María Belén Camandone