SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

INSTITUTO POLITÉCNICO NACIONAL
ESCOM

Inteligencia Artificial

UNIDAD IV

Aprendizaje de máquina

ESCOM-IPN Daniel Aguilar Velázquez - Rodolfo Romero Herrera – Gabriela de Jesús López Ruiz

Unidad de competencia

- Construye algoritmos de aprendizaje válidos a partir de los diferentes tipos de aprendizaje de máquina.

Contenido

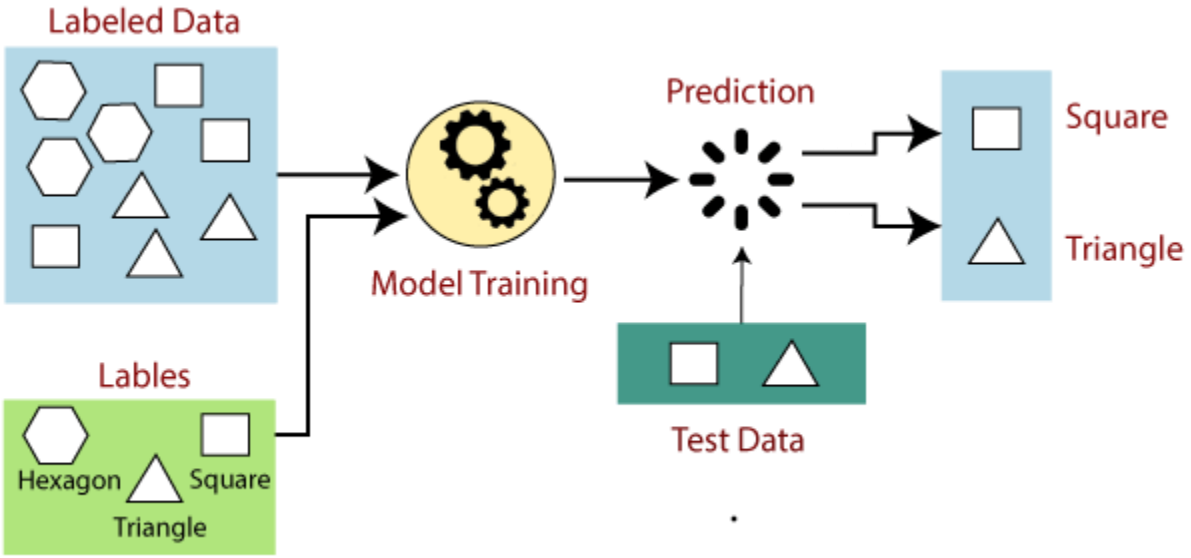
Contenido

Aprendizaje

Se denomina **aprendizaje** al proceso de adquisición de conocimientos y habilidades, posibilitado mediante el estudio, la enseñanza o la experiencia. Para nuestro caso de estudio, el **aprendizaje automático** (Machine Learning) es el subapartado de la inteligencia artificial (IA) que se centra en desarrollar sistemas que aprenden, o mejoran el rendimiento, en función de los datos que consumen. Es decir es la capacidad que tiene un modelo de generar su propia base de conocimiento, mediante el uso de información.

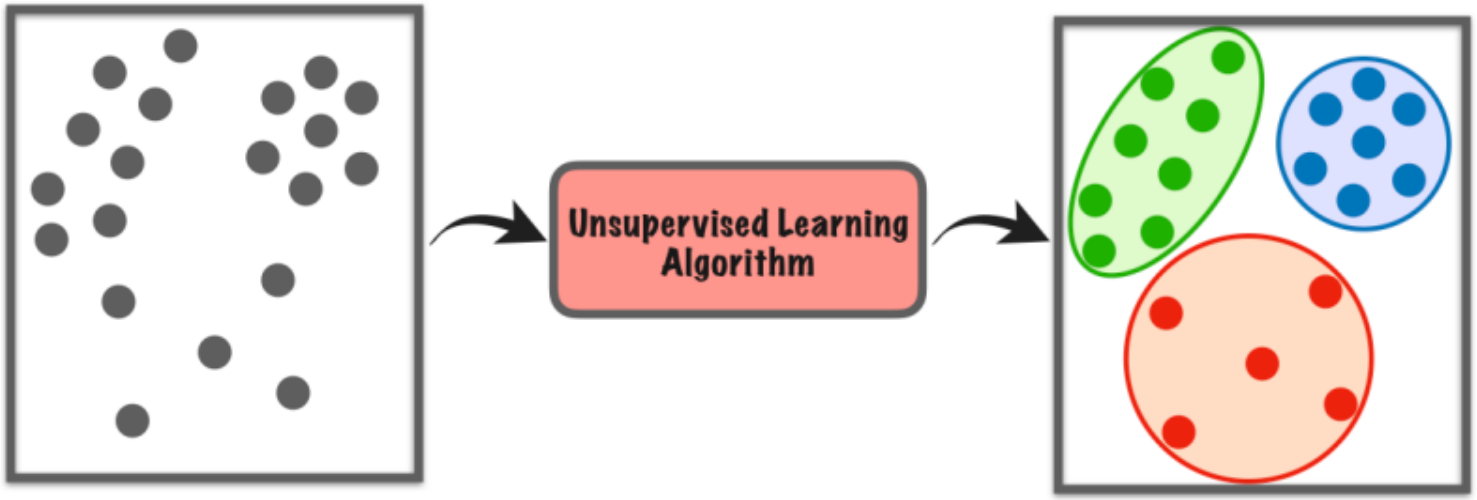
Aprendizaje supervisado

Los métodos de **aprendizaje supervisado** requieren la creación de un conjunto de datos de aprendizaje, con los cuales se le pueda enseñar a un modelo de inteligencia artificial mediante ejemplo, es decir, es necesario tener un conjunto de datos de entrada y sus respectivos resultados esperados.



Aprendizaje no supervisado

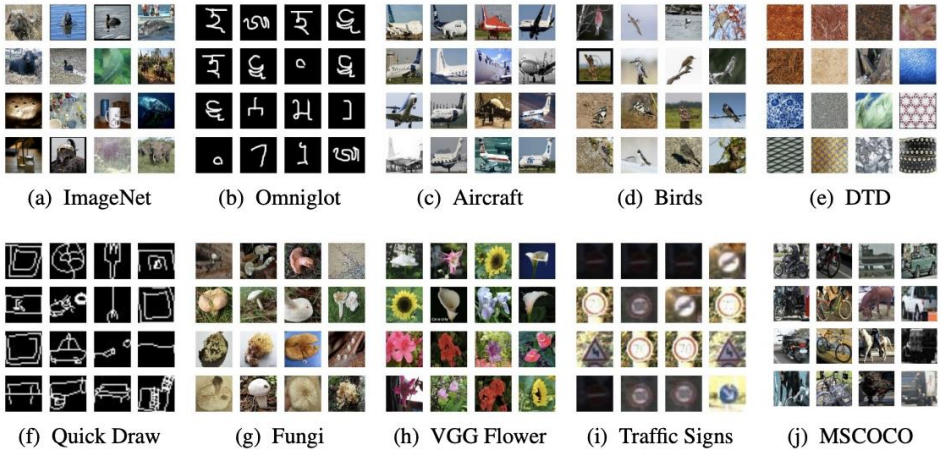
En los métodos de **aprendizaje no supervisado** se elimina la necesidad de un conjunto de datos previamente etiquetados, es decir que se conozca su resultado esperado, ya que será responsabilidad del modelo decidir las clases o categorías a las que pertenecen los datos de entrada.



Características de un conjunto de datos

Un **conjunto de datos** (conocido también como **dataset**) es una colección de datos habitualmente tabulados. En el caso de datos tabulados, un conjunto de datos contiene los valores para cada una de las variables organizadas como columnas, como por ejemplo la altura y el peso de un objeto, que corresponden a cada miembro del conjunto de datos, que están organizados en filas. Cada uno de estos valores se conoce con el nombre de dato. El conjunto de datos también puede consistir en una colección de documentos o de archivos.

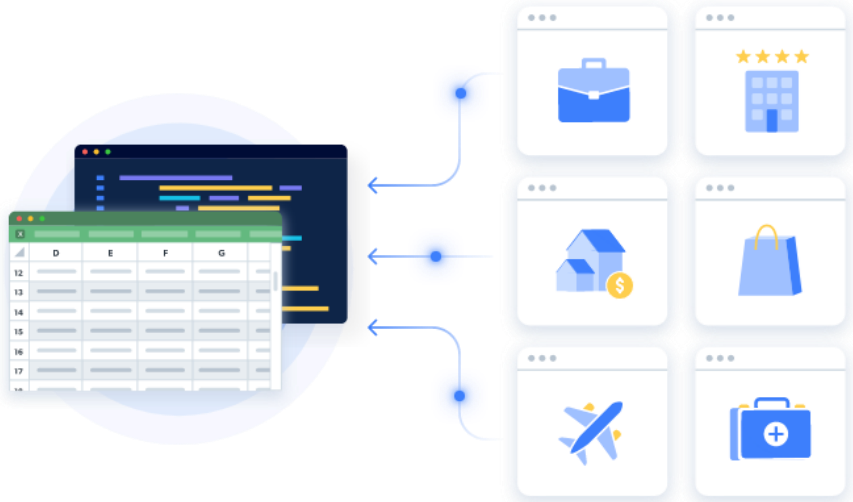
Conjuntos de datos tan grandes que aplicaciones tradicionales de procesamiento de datos no los pueden tratar se llaman big data.



Tipos de características

Según el estudio, es un factor clave de las propiedades del conjunto: **dispersión**, **curtosis**, etc. Los valores pueden ser **números**, como **números reales** o **enteros**, por ejemplo, que representan la altura de una persona en centímetros, pero también pueden ser **datos nominales** (es decir, que no consisten en valores numéricos), por ejemplo, que representan la etnia de una persona. De manera más general, los valores pueden ser de cualquiera de los tipos descritos como nivel de medición.

En estadística, los conjuntos de datos generalmente provienen de observaciones reales obtenidas al muestrear una población estadística, y cada fila corresponde a las observaciones de un elemento de esa población.



Problemas en los conjuntos de datos

Los **conjuntos de datos** pueden presentar diferentes desafíos a tener en cuenta, como los que se muestran a continuación:

- **Tamaño de la muestra**: El tamaño de la muestra se le conoce como aquel número determinado de sujetos o cosas que componen la muestra extraída de una población, necesarios para que los datos obtenidos sean representativos de la población. Un tamaño de muestra reducido puede resultar en un agente inteligente poco flexible al enfrentarse a situaciones nuevas, existen diferentes formas de combatir este problema, como lo es la recolección de nuevos datos o la fabricación de ellos mediante procesos de **data augmentation**.

- **Desequilibrio de clases**: Este problema se presenta cuando el número de registros para una de las clases a clasificar es inferior al resto. Cuando el desequilibrio es pequeño, uno a dos, esto no supone un problema, pero cuando es grande es un problema para la mayoría de los modelos de clasificación. Una de las formas de tratar este problema es mediante el **remuestreo**.

Problemas en los conjuntos de datos

- **Complejidad**: El problema de complejidad se presenta cuando existe un número muy grande de variables para cada registro, ésto implica que los modelos deberán aprender y manejar cada una de las variables lo que aumentará, significativamente la complejidad del aprendizaje. Para el manejo de este problema se debe realizar una selección de variables mediante el **análisis causal**.

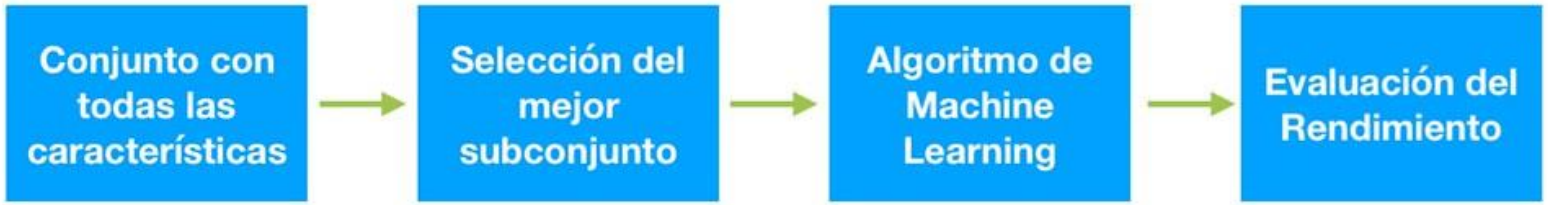
- **Datos ruidosos**: Los datos ruidosos son datos que no tienen sentido por la existencia de demasiadas variaciones. Para el manejo de este problema se puede realizar un proceso de **normalización** o **discretización** sobre las variables que presenten ruido.

- **Valores atípicos**: Los valores atípicos son puntos de datos observados que se alejan de la línea de mínimos cuadrados. Tienen grandes "errores", donde el "error" o residual es la distancia vertical de la línea al punto. Para el manejo de este problema se puede **reducir el peso** que tienen los valores atípicos.

Selección de características

La **Selección de Características** es el proceso de seleccionar las más importante y/o relevante características de un conjunto de datos, con el objetivo de mejorar el rendimiento de predicción de los predictores, proporcionar predictores más rápidos y más rentables y proporcionar una mejor comprensión del proceso subyacente que generó los datos. Los **métodos de filtro** se utilizan generalmente como un paso de preprocesamiento de datos, la selección de características es independiente de cualquier algoritmo de Machine Learning.

Las características se clasifican según los puntajes estadísticos que tienden a determinar la correlación de las características con la variable de resultado.



Selección de características

Existen diversas motivaciones para ejecutar un proceso de selección de características en machine learning. Entre ellas están:

Interpretabilidad: A menos variables de entrada, más fácil es explicar cómo afecta cada una de ellas en el resultado final. Esto se puede hacer de dos formas:

- **Eliminando variables irrelevantes**.
- **Entendiendo mejor los datos.**

Reducir costos computacionales del entrenamiento.

Evitar el sobreajuste u overfitting: El **sobreajuste** es un concepto en la ciencia de datos, qué ocurre cuando un modelo estadístico se ajusta exactamente a sus datos de entrenamiento. Cuando esto sucede, el algoritmo desafortunadamente no puede funcionar con precisión contra datos invisibles, frustrando su propósito.

Algoritmos de clasificación

Los algoritmos de clasificación se utilizan cuando el resultado deseado es una etiqueta discreta. Es decir, son útiles cuando la respuesta a la pregunta sobre la empresa se aloja dentro de un conjunto finito de resultados posibles. En el caso de que el modelo entrenado es para predecir cualquiera de las dos clases objetivos, verdadero o falso, por ejemplo, se le conoce como clasificación binaria. Algunos ejemplos de esto son: predecir si un alumno aprobará o no, predecir si un cliente comprará un producto nuevo o no.

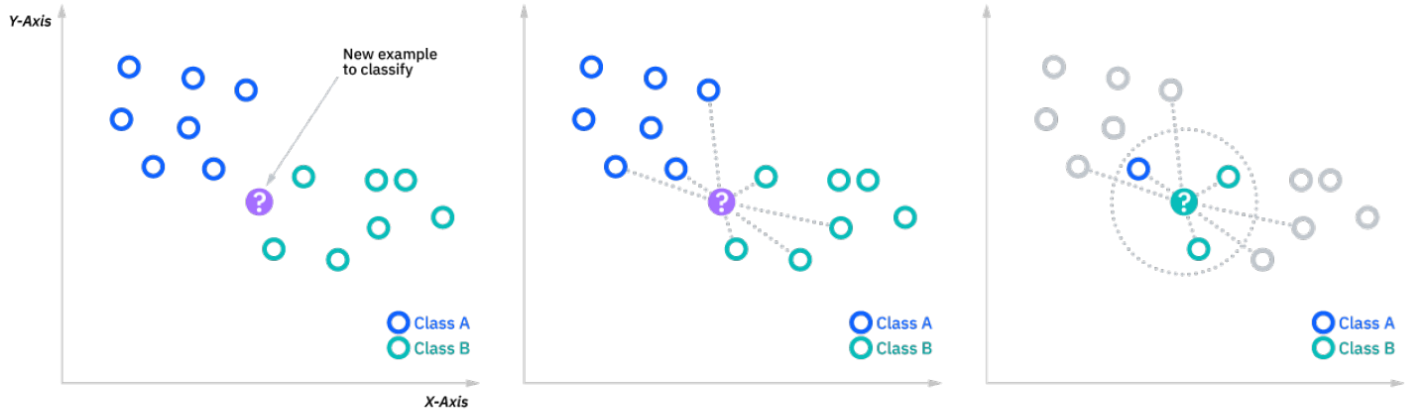
Por su parte, si se quiere predecir más de dos clases objetivos, se le conoce como clasificación multicategoría. Un ejemplo de esto es predecir qué asignaturas un alumno tendrá las más clasificaciones. Este tipo de clasificación es útil para la segmentación del cliente, la categorización de imágenes y audio y análisis de texto.

Algoritmos basados en distancia

Los **algoritmos de clasificación basados en distancia** son algoritmos que permiten evaluar la similitud entre los pares de elementos que conforman un conjunto de datos.

Algoritmo de distancia mínima o KNN

Es un método usado para agrupar objetos basados en la similaridad de sus atributos, dado un conjunto de datos, cada uno con un número n de atributos, el algoritmo del vecino más cercano decide si un nuevo elemento x pertenece a la misma categoría que su vecino más cercano. En otras palabras, el algoritmo de **KNN** asigna una categoría a un nuevo elemento del conjunto de datos basado en la categoría de la mayoría de sus K vecinos, donde K es un número natural. El algoritmo de mínima distancia es un caso especial del algoritmo KNN donde K = 1.

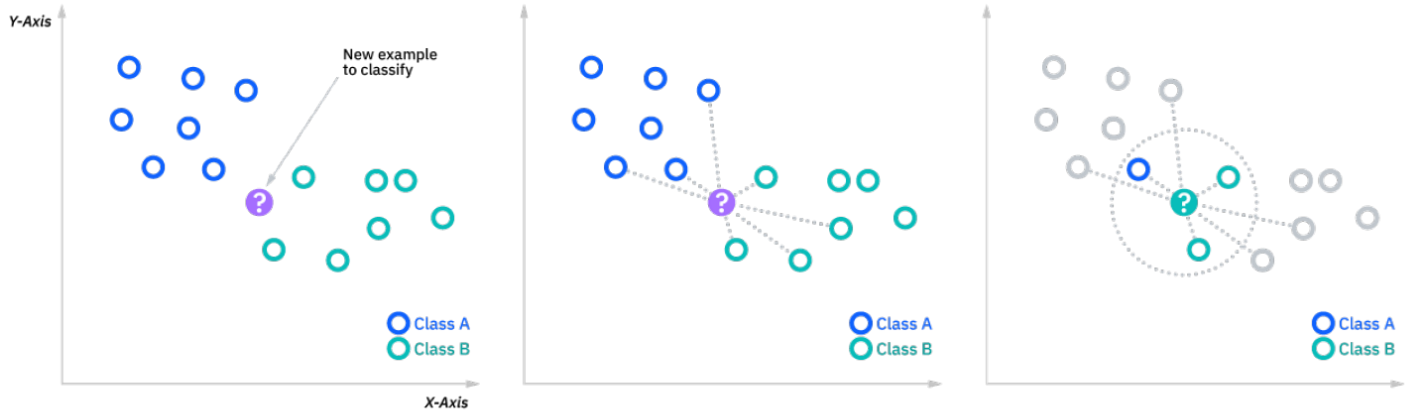


KNN

Para el algoritmo de KNN lo primero es determinar un número K correspondiente al número de vecinos con los que queremos verificar. Posteriormente se debe medir la distancia del nuevo elemento, a los elementos actuales, para el caso de elementos numéricos se puede hacer uso de la distancia euclidiana:

$$d(\boldsymbol{x},\boldsymbol{y}) = |\boldsymbol{x}-\boldsymbol{y}| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}.$$

Es necesario almacenar las clases de las K distancias más cercanas a nuestro nuevo elemento, y posteriormente se decide a qué clase pertenece el nuevo elemento. Adicionalmente es posible modificar el algoritmo, agregando una condicional de distancia máxima permitida m y si ningún elemento se encuentra a una distancia menor a la distancia máxima se crea una nueva clase para el nuevo elemento.



Árboles de decisión

Un **árbol de decisión** es un algoritmo de aprendizaje supervisado no paramétrico, que se utiliza tanto para tareas de clasificación como de regresión. Tiene una estructura de árbol jerárquico, que consta de un nodo raíz, ramas, nodos internos y nodos hoja. Este tipo de estructura de diagrama de flujo también crea una representación fácil de digerir de la toma de decisiones, lo que permite que diferentes grupos en una organización comprendan mejor por qué se tomó una decisión.

El aprendizaje del árbol de decisiones emplea una estrategia de divide y vencerás mediante la realización de una búsqueda codiciosa para identificar los puntos de división óptimos dentro de un árbol. Este proceso de división se repite de forma recursiva de arriba hacia abajo hasta que todos o la mayoría de los registros se hayan clasificado bajo etiquetas de clase específicas.

Árboles de decisión

El **algoritmo ID3** es utilizado dentro del ámbito de la inteligencia artificial. Su uso se engloba en la búsqueda de hipótesis o reglas en él, dado un conjunto de ejemplos. El conjunto de ejemplos deberá estar conformado por una serie de tuplas de valores, cada uno de ellos denominados atributos, en el que uno de ellos, (el atributo a clasificar) es el objetivo, el cual es de tipo binario.

De esta forma el algoritmo trata de obtener las hipótesis que clasifiquen ante nuevas instancias, si dicho ejemplo va a ser positivo o negativo.

ID3 realiza esta labor mediante la construcción de un **árbol de decisión**.
Los elementos son:

- **Nodos**: Los cuales contendrán atributos.
- **Arcos**: Los cuales contienen valores posibles del nodo padre.
- **Hojas**: Nodos que clasifican el ejemplo como positivo o negativo.

Algoritmo ID3

Id3(Ejemplos, Atributo-objetivo, Atributos)

Si todos los ejemplos son positivos devolver un nodo positivo
Si todos los ejemplos son negativos devolver un nodo negativo
Si Atributos está vacío devolver el voto mayoritario del valor del atributo objetivo

Ejemplos
En otro caso
Sea A Atributo el MEJOR de atributos
Para cada v valor del atributo hacer
Sea Ejemplos(v) el subconjunto de ejemplos cuyo valor de atributo A es v
Si Ejemplos(v) está vacío devolver un nodo con el voto mayoritario del Atributo objetivo de Ejemplos
Sino Devolver Id3(Ejemplos(v), Atributo-objetivo, Atributos/{A})

La elección del mejor atributo se establece mediante la entropía. Eligiendo aquel que proporcione una mejor ganancia de información. La función elegida puede variar, pero en su forma más sencilla es como esta:

$$-\left(\frac{|p|}{|d|}\right)\log_2\left(\frac{|p|}{|d|}\right)-\left(\frac{|n|}{|d|}\right)\log_2\left(\frac{|n|}{|d|}\right)$$

Algoritmo C4.5

C4.5 es un algoritmo usado para generar un árbol de decisión desarrollado por Ross Quinlan. C4.5 es una extensión del algoritmo **ID3** desarrollado anteriormente por Quinlan. Los árboles de decisión generados por C4.5 pueden ser usados para clasificación.

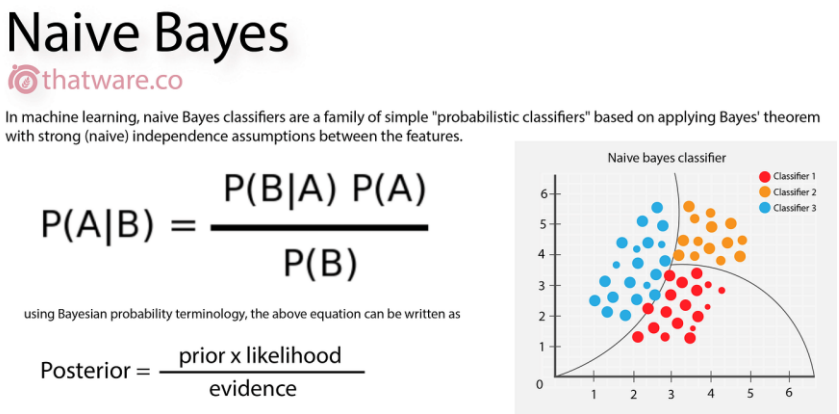
Los datos de entrenamiento son un grupo S = s1, s2, ... de ejemplos ya clasificados. Cada ejemplo si = x1, x2, ... es un vector donde x1, x2, ... representan los atributos o características del ejemplo. Los datos de entrenamiento son aumentados con un vector C = c1, c2, ... donde c1, c2, ... representan la clase a la que pertenece cada muestra.

En cada nodo del árbol, C4.5 elige un atributo de los datos que más eficazmente dividen el conjunto de muestras en subconjuntos enriquecidos en una clase u otra. Su criterio es el normalizado para ganancia de información (diferencia de entropía) que resulta en la elección de un atributo para dividir los datos. El atributo con la mayor ganancia de información normalizada se elige como parámetro de decisión. El algoritmo C4.5 divide recursivamente en sublistas más pequeñas.

Naive Bayes

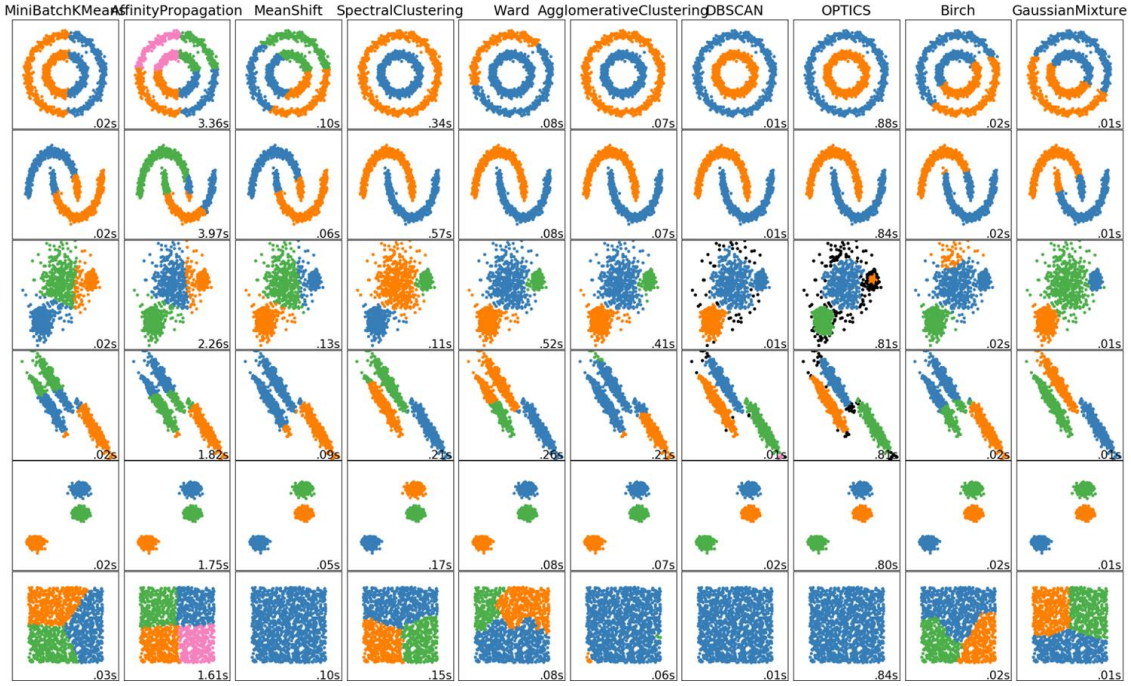
En teoría de la probabilidad y minería de datos, un clasificador **Naive Bayes** es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales. Es a causa de estas simplificaciones, que se suelen resumir en la hipótesis de independencia entre las variables predictoras, que recibe el apelativo de naive, es decir, ingenuo.

En términos simples, un clasificador de **Naive Bayes** asume que la presencia o ausencia de una característica particular no está relacionada con la presencia o ausencia de cualquier otra característica, dada la clase variable. Por ejemplo, una fruta puede ser considerada como una manzana si es roja, redonda y de alrededor de 7 cm de diámetro. Un clasificador de Naive Bayes considera que cada una de estas características contribuye de manera independiente a la probabilidad de que esta fruta sea una manzana, independientemente de la presencia o ausencia de las otras características.



Algoritmos de agrupamiento

Un **algoritmo de agrupamiento** (en inglés, **clustering**) es un procedimiento de agrupación de una serie de vectores de acuerdo con un criterio. Esos criterios son por lo general distancia o similitud. La cercanía se define en términos de una determinada función de distancia, como la euclídea, aunque existen otras más robustas o que permiten extenderla a variables discretas.



Algoritmos basados en distancia

K-means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Se suele usar la distancia cuadrática.

El algoritmo consta de tres pasos:

1. **Inicialización**: una vez escogido el número de grupos, k, se establecen k centroides en el espacio de los datos, por ejemplo, escogiendo aleatoriamente.
2. **Asignación objetos a los centroides**: cada objeto de los datos es asignado a su centroide más cercano.
3. **Actualización centroides**: se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Se repiten los pasos 2 y 3 hasta que los centroides no se mueven, o se mueven por debajo de una distancia umbral en cada paso.

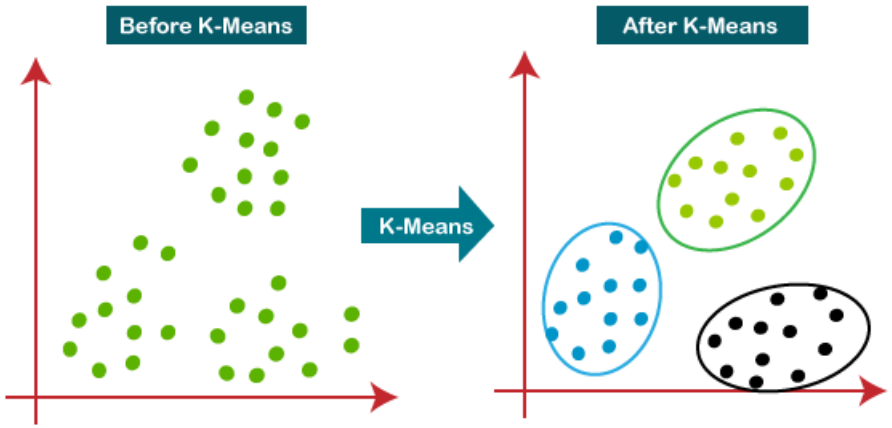
K-means

El algoritmo k-means resuelve un problema de optimización, siendo la función a optimizar (minimizar) la suma de las distancias cuadráticas de cada objeto al centroide de su cluster.

Los objetos se representan con vectores reales de **d** dimensiones (x1,x2,...,xn) y el algoritmo k-means construye k grupos donde se minimiza la suma de distancias de los objetos, dentro de cada grupo S={S1,S2,...,Sk}, a su centroide. El problema se puede formular de la siguiente forma:

$$\min_{\mathbf{S}} E(\boldsymbol{\mu}_i) = \min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

donde **S** es el conjunto de datos cuyos elementos son los objetos xj representados por vectores, donde cada uno de sus elementos representa una característica o atributo. Tendremos k grupos o clusters con su correspondiente centroide **μi.**

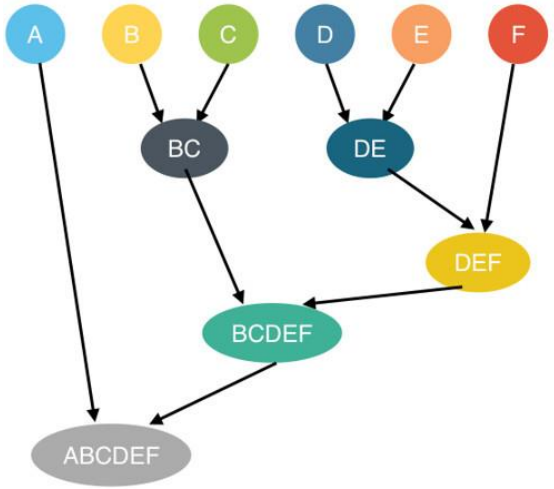


Algoritmos basados en jerarquías

En minería de datos, el **agrupamiento jerárquico** es un método de análisis de grupos puntuales, el cual busca construir una jerarquía de grupos. Estrategias para agrupamiento jerárquico generalmente caen en dos tipos:

- **Aglomerativas**: Este es un acercamiento ascendente: cada observación comienza en su propio grupo, y los pares de grupos son mezclados mientras uno sube en la jerarquía.
- **Divisivas**: Este es un acercamiento descendente: todas las observaciones comienzan en un grupo, y se realizan divisiones mientras uno baja en la jerarquía.

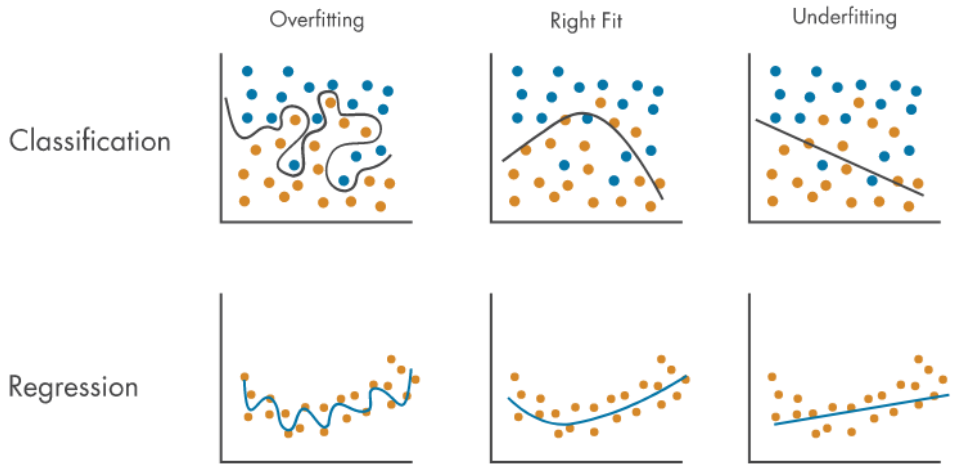
En general, las mezclas y divisiones son determinadas con un algoritmo voraz. Los resultados del agrupamiento jerárquico son usualmente presentados en un dendrograma.



Métodos de validación

Los **métodos de validación** se utilizan para determinar de manera sistemática, el mérito, el valor y el rendimiento de un agente en función de ciertos criterios respecto a un conjunto de normas. Los resultados de la validación permiten realizar ajustes a lo modelos para intentar mejorar el rendimiento de los mismos, se pueden dividir en:

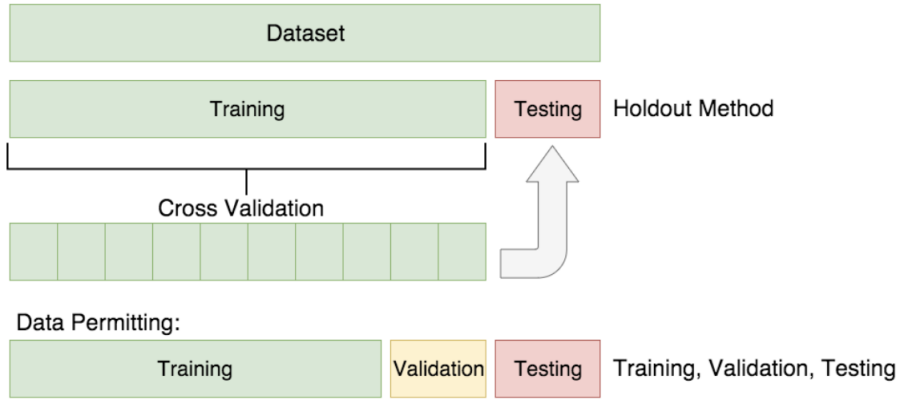
- **Métodos de validación de algoritmos de clasificación**: Entrenamiento y prueba, validación cruzada y matriz de confusión.
- **Métodos de validación de algoritmos de agrupamiento**: Medidas de validación internas y externas.



Métodos de validación de algoritmos de clasificación

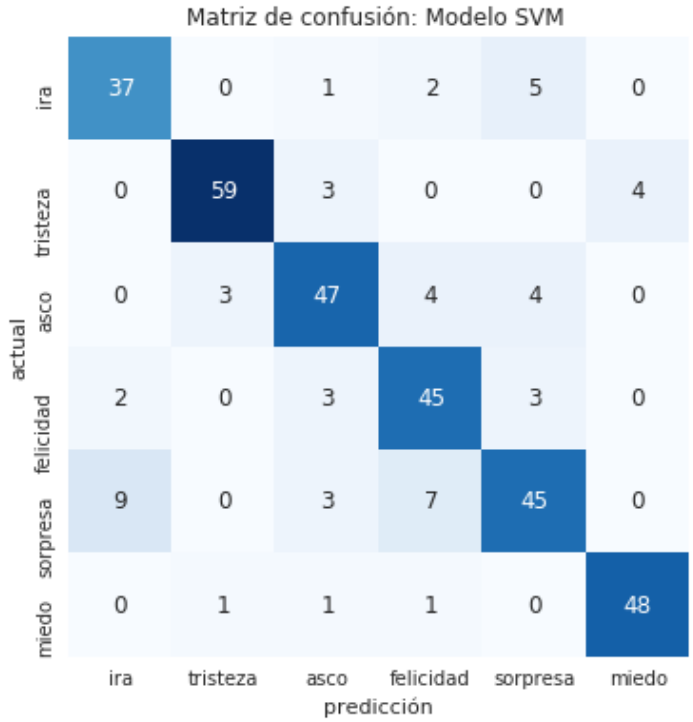
Para la validación dentro de los algoritmos de clasificación es necesario primeramente realizar una segmentación de los datos datos de entrenamiento, donde se deberán dejar grupos bien definidos: **entrenamiento**, **validación** y **pruebas**, siendo el conjunto de entrenamiento el más extenso, una disposición común para el tamaño de los conjuntos es 70%, 15% y 15% respectivamente.

El **cross-validation** o la **validación cruzada** es un método estadístico que permite estimar la forma en la que van a funcionar y si es la manera más adecuada del desarrollo de los modelos predictivos del campo de machine learning. El cross-validation también es conocido como out-of-sample testing, donde el objeto modelos son los de tipo de predicción.



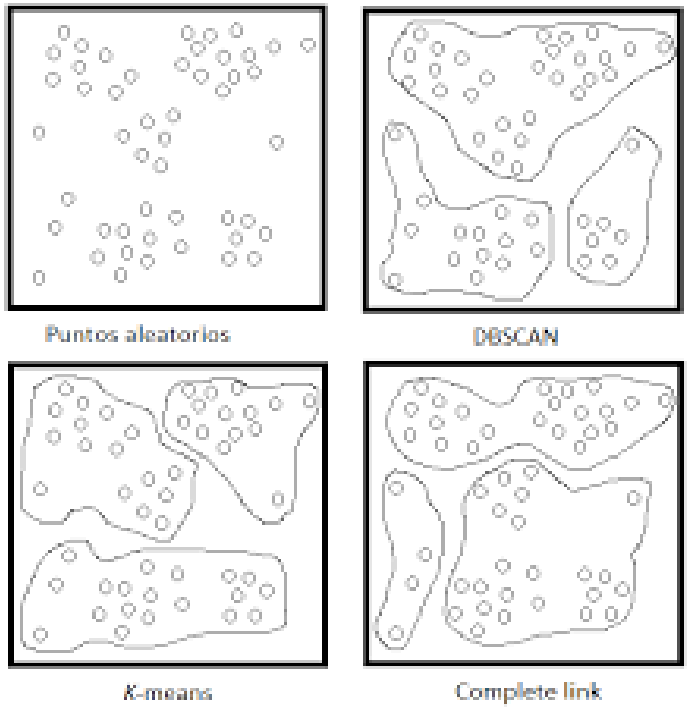
Métodos de validación de algoritmos de clasificación

En el campo de la inteligencia artificial una **matriz de confusión** es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases.



Métodos de validación de algoritmos de agrupamiento

Es de importancia evaluar el resultado de los algoritmos de agrupamiento, sin embargo, es difícil definir cuando el resultado de un agrupamiento es aceptable. Por esta razón existen técnicas e índices para la validación de un agrupamiento realizado. Existen dos tipos de validación: La **validación externa** y la **validación interna** son las dos categorías más importantes para la validación de clustering. La principal diferencia es si se usa o no información externa para la validación, es decir, información que no es producto de la técnica de agrupación utilizada



Métricas de Validación Interna

Como el objetivo del clustering es agrupar objetos similares en el mismo clúster y objetos diferentes ubicarlos en diferentes clúster, las métricas de validación interna están basadas usualmente en los dos siguientes criterios:

- **Cohesión**: El miembro de cada clúster debe ser lo más cercano posible a los otros miembros del mismo clúster.

- **Separación**: Los clúster deben estar ampliamente separados entre ellos. Existen varios enfoques para medir esta distancia entre clúster: distancia entre el miembro más cercano, distancia entre los miembros más distantes o la distancia entre los centroides.

- **Sum of Squared Within (SSW)**: Medida interna especialmente usada para evaluar la Cohesión de los clústeres que el algoritmo de agrupamiento generó. Siendo k el número de clústeres, x un punto del clúster Ci y mi el centroide del clúster Ci.

$$SSE = \sum_{i=1}^{k} \sum_{x \epsilon C_i} dist^2(m_i, x)$$

Métricas de Validación Interna

- **Sum of Squared Between (SSB)**: Es una medida de separación utilizada para evaluar la distancia inter-clúster (Separación). Siendo k el número de clústeres, nj el número de elementos en el clúster j, cj el centroide del clúster j y x es la media del data set.

$$SSB = \sum_{j=1}^{k} n_j \, dist^2 \, (c_j - \bar{x})$$

- **Sum of Squares based Indexes**: Los índices o medidas basadas en las "sumas de cuadrados" presentadas anteriormente se caracterizan por medir o cuantificar la dispersión de los puntos a nivel inter-cluster e intra-cluster. Los índices son:

Ball y Hall (1965)	$\frac{SSW}{k}$
Calinski y Harabasz (1974)	$\frac{SSB/(k-1)}{SSW/(n-k)}$
Hartigan (1975)	$log(\frac{SSB}{SSW})$
Xu (1997)	$d * \log\left(\sqrt{\frac{SSW}{dN^2}}\right) + \log(k)$

Siendo k el número de clústeres, N el número de datos y d la dimensión de los datos.

Métricas de Validación Externa

Cuando se tiene información externa tal como la clase de cada dato, es común y ampliamente utilizado el siguiente análisis: Se tiene la clase de cada dato en el dataset, es decir, se tiene de antemano el número de clúster y a cual clúster pertenece cada dato.

- **Verdadero positivo**: Este término hace referencia a aquellos puntos que fueron ubicados por el algoritmo en el mismo clúster que indicaba la clase con la que se contaba de antemano.

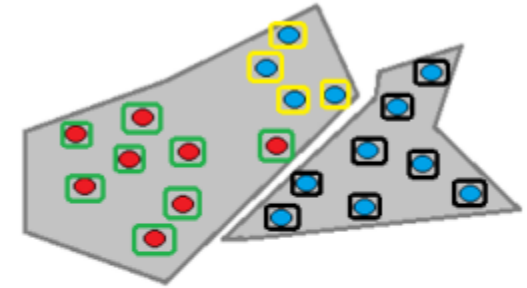
- **Falso positivo**: Hace referencia a aquellos puntos que fueron ubicados por el algoritmo en el clúster j y que en realidad pertenecían a otro clúster.

- **Falsos negativos**: Hacen referencia a aquellos elementos del clúster j que fueron ubicados en un clúster diferente al que indicaba su etiqueta. En el ejemplo presentado, el clúster j tiene todos sus elementos asignados correctamente, luego no hay falsos negativos.

- **Verdadero negativo**: Hace referencia a aquellos elementos que fueron ubicados correctamente fuera del clúster j, es decir, aquellos elementos ajenos al clúster en cuestión y que efectivamente no correspondían a este.

Métricas de Validación Externa

Con la terminología anterior aclarada es posible introducir las siguientes métricas ampliamente utilizadas y provenientes del campo de Information Retrieval: la Precisión y el Recall.



Hipótesis	Verdad: P	Verdad: N
P	VP	FP
N	FN	VN

$$Precisión = \frac{a}{a + b}$$

$$Recall = \frac{a}{a + c}$$