UANL

FCFM

CLUSTERING

DATA MINING

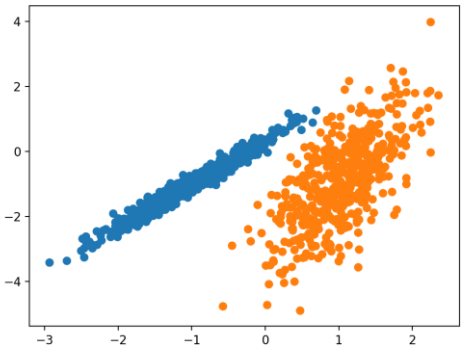
Anahí Alemán Alvarado 1821952
Ricardo Zarek Sánchez Olivares 1795134
Juan Pablo Nasser Benavides 1753367
Eduardo Almaguer Alanís 1741322
Oscar Saúl Vega Macias 1626997

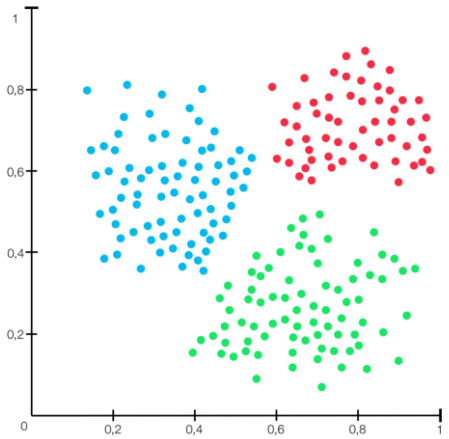
¿Qué es?

También conocido como agrupamiento, es una de las técnicas de minería de datos, el proceso consiste en la división de los datos en grupos de objetos similares.

Las técnicas de Clustering son las que utilizando algoritmos matemáticos se encargan de agrupar objetos. Usando la información que brindan las variables que pertenecen a cada objeto se mide la similitud entre los mismos, y una vez hecho esto se colocan en clases que son muy similares internamente y a la vez diferente entre los miembros de las diferentes clases.

Fuente: https://quantdare.com/quantitative-clustering-with-machine-learning/



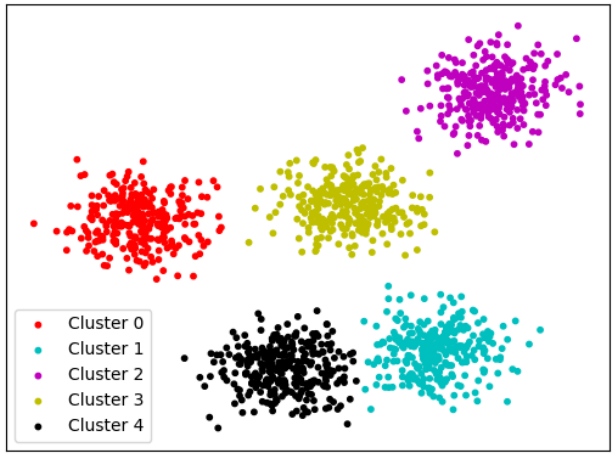


Un **cluster** es una colección de objetos de datos. Similares entre sí dentro del mismo grupo. Disimilar a los objetos en otros grupos.

Análisis de cluster: dado un conjunto de puntos de datos tratar de entender su estructura. Encuentra similitudes entre los datos de acuerdo con las características encontradas en los datos. Es un aprendizaje no supervisado ya que no hay clases predefinidas.

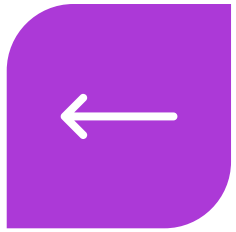
Fuente:https://medium.com/@jaywrkr/miner%C3%ADa-de-datos-4-3bc03e7fc234#:~:text=La%20agrupaci%C3%B3n%20(Clustering)%20busca%20agrupaciones,una%20peque%C3%B1a%20distancia%20entre%20s%C3%AD.

Aplicaciones



Fuente: https://medium.com/@jaywrkr/miner%C3%ADa-de-datos-4-3bc03e7fc234#:~:text=La%20agrupaci%C3%B3n%20(Clustering)%20busca%20agrupaciones,una%20peque%C3%B1a%20distancia%20entre%20s%C3%AD.

Métodos de agrupación

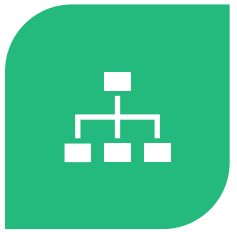


ASIGNACIÓN JERÁRQUICA FRENTE A PUNTO

DATOS NUMÉRICOS Y/O SIMBÓLICOS

DETERMINÍSTICA VS. PROBABILÍSTICA

EXCLUSIVO VS. SUPERPUESTO



JERÁRQUICO VS. PLANO.

DE ARRIBA A ABAJO Y DE ABAJO A ARRIBA

Fuente: https://medium.com/@jaywrkr/miner%C3%ADa-de-datos-4-3bc03e7fc234#:~:text=La%20agrupaci%C3%B3n%20(Clustering)%20busca%20agrupaciones,una%20peque%C3%B1a%20distancia%20entre%20s%C3%AD.

ALGORITMOS DE CLUSTERING

Simple K-Means

Este algoritmo debe definir el número de clusters que se desean obtener, así se convierte en un algoritmo voraz para particionar. Pasos:

1. Se determina la cantidad de clusters en los que se quiere agrupar la información, en este caso las simulaciones.

2. Se asume de forma aleatoria los centros por cada clusters. Una vez encontrados los primeros centroides el algoritmo hará los tres pasos siguientes:
   - Determina las coordenadas del centroide.
   - Determina la distancia de cada objeto a los centroides.
   - Agrupa los objetos basados en la menor distancia.

3. Finalmente quedarán agrupados por clusters, los grupos de simulaciones según la cantidad de clusters que el investigador definió en el momento de ejecutar el algoritmo.

Fuente: https://www.ecured.cu/Clustering

X-Means

Este algoritmo es una variante mejorada del **K-Means**.

Su ventaja fundamental está en haber solucionado una de las mayores deficiencias presentadas en K-Means, el hecho de tener que seleccionar a priori el número de clusters que se deseen obtener, a **X-Means** se le define un límite inferior **K-min** (número mínimo de clusters) y un límite superior **K-Max** (número máximo de clusters) y este algoritmo es capaz de obtener en ese rango el número óptimo de clusters, dando de esta manera más flexibilidad al usuario.

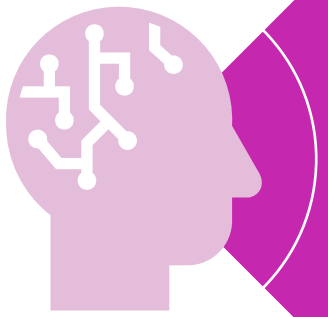
Fuente: https://www.ecured.cu/Clustering

Cobweb

Pertenece a la familia de algoritmos **jerárquicos**. Se caracteriza por la utilización de aprendizaje incremental, esto quiere decir, que realiza las agrupaciones instancia a instancia. Durante la ejecución del algoritmo se forma un árbol (**árbol de clasificación**) donde las hojas representan los segmentos y el nodo raíz engloba por completo el conjunto de datos.

Además, en el algoritmo también hay que tener en cuenta dos parámetros muy importantes:

Acuity: es un parámetro muy necesario, pues la utilidad de categoría está basada en la estimación de la media y la desviación estándar del valor de un atributo para un nodo en particular.



Cut-off: este parámetro es usado para evitar el crecimiento descontrolado de la cantidad de segmentos. Indica el grado de mejoría que se debe producir en la utilidad de categoría

Fuente: https://www.ecured.cu/Clustering

EM

Este algoritmo pertenece a una familia de modelos que se conocen como Finite Mixture Models, los cuales se pueden utilizar para segmentar conjuntos de datos. Está clasificado como un método de particionado y recolocación, o sea, **Clustering Probabilístico**. Se trata de obtener la FDP (Función de Densidad de Probabilidad) desconocida a la que pertenecen el conjunto completo de datos.

El algoritmo EM, procede en dos pasos que se repiten de forma iterativa:

- Expectation: Utiliza los valores de los parámetros, iniciales o proporcionados por el paso Maximization, obteniendo diferentes formas de la FDP buscada.
- Maximization: Obtiene nuevos valores de los parámetros a partir de los datos proporcionados por el paso anterior.

Finalmente se obtendrá un conjunto de clusters que agrupan el conjunto de proyectos original. Cada uno de estos cluster estará definido por los parámetros de una distribución

Fuente: https://www.ecured.cu/Clustering