

# Clustering

Data Mining - Ocampo Botello Fabiola

Becerril Hernández Aldo  
Lopez Garcia Felipe de Jesus

# Clustering

El clustering es una tarea que tiene como finalidad principal lograr el agrupamiento de conjuntos de objetos no etiquetados, para lograr construir subconjuntos de datos conocidos como Clusters.



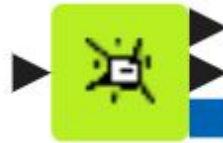
# K-Means

El objetivo de K-medias es agrupar a las observaciones de forma tal que todas las que se encuentren en el mismo grupo sean lo más semejantes entre sí y que las pertenecientes a grupos distintos sean lo más desemejantes entre sí. Las medidas de distancia, como la euclídea, son utilizadas para medir la semejanza y desemejanza. Una medida para indicar cuán bien los centroides representan a los miembros de su grupo es la suma de los errores al cuadrado. K-medias, en cada iteración, intenta reducir el valor de la suma de los errores al cuadrado. La medida consiste en la sumatoria de las distancias al cuadrado de cada observación al centroide de su grupo:



# K-Means - Knime

Este nodo genera los centros de clúster para un número predefinido de clústeres (sin número dinámico de clústeres). K-means realiza un agrupamiento nítido que asigna un vector de datos a exactamente un clúster. El algoritmo termina cuando las asignaciones de clúster ya no cambian. El algoritmo de agrupamiento utiliza la distancia euclidiana en los atributos seleccionados. Los datos no son normalizados por el nodo (si obligatorio, debe considerar utilizar el "Normalizador" como paso de preprocesamiento).



# Hierarchical Clustering

Clustering Jerárquico es un método de data mining para agrupar datos (en minería de datos a estos grupos de datos se les llama clústers). El algoritmo de clúster jerárquico agrupa los datos basándose en la distancia entre cada uno y buscando que los datos que están dentro de un clúster sean los más similares entre sí. En una representación gráfica los elementos quedan anidados en jerarquías con forma de árbol.

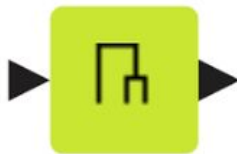
Al igual que el método de K-Means, los algoritmos de agrupamiento jerárquico están dentro de la categoría de algoritmos de aprendizaje no supervisado.



# Hierarchical Clustering - Knime

Agrupar jerárquicamente los datos de entrada. Nota: Este nodo solo funciona en conjuntos de datos pequeños. Mantiene todos los datos en la memoria y tiene una complejidad cúbica.


Este algoritmo funciona de forma aglomerante. Para determinar la distancia entre grupos hay que definir una medida.



# Conjunto de Datos

## About this file

This file contains the basic information (ID, age, gender, income, spending score) about the customers

CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (...)
Unique ID assigned to the customer	Gender of the customer	Age of the customer	Annual Income of the customer	Score assigned by the mall based on customer behavior and spending nature
 200	Female 56% Male 44%	 18 70	 15 137	 1
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77

# Diccionario de Datos

Nombre	Significado	Tipo de dato
CustomerID	Es el número con el que se identifica cada valor de los atributos en cada fila de la tabla.	numérico
Gender	Es la clasificación de género de las personas dentro de este conjunto de datos, pueden ser mujeres u Hombres.	nominal
Age	Es la edad de las personas en años.	numérico
Annual Income	Son los ingresos por año de cada persona.	numérico
Spending Score	Son las puntuaciones de gastos para cada persona propuestas por el centro comercial, pueden tener un valor de entre 1 y 100	numérico



# Tabla de los Promedios Principales de las Estadísticas

	Promedios Principales				
Atributos principales	Cluster_0	Cluster_1	Cluster_2	Cluster_3	Cluster_4
Edad	25.27	45.21	42.93	40.66	32.69
Ingresos anuales	25.72	26.30	55.08	87.65	86.53
Puntuacion de gastos	79.36	20.91	49.71	13.58	82.12