KSCHOOL

Del Dato al Conocimiento

Aprendizaje no Supervisado

Antonio Pita Lozano

Máster en Data Science

Índice

Aprendizaje no Supervisado

1. Agrupamiento (Clustering)
   A. Clustering Jerárquico Aglomerativo
   B. Clustering K-means
2. Reducción de Dimensionalidad
   C. Componentes Principales

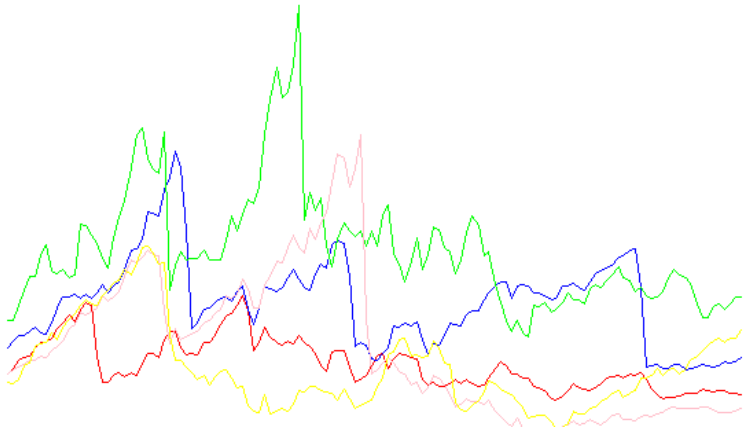
Aprendizaje no Supervisado

KSCHOOL

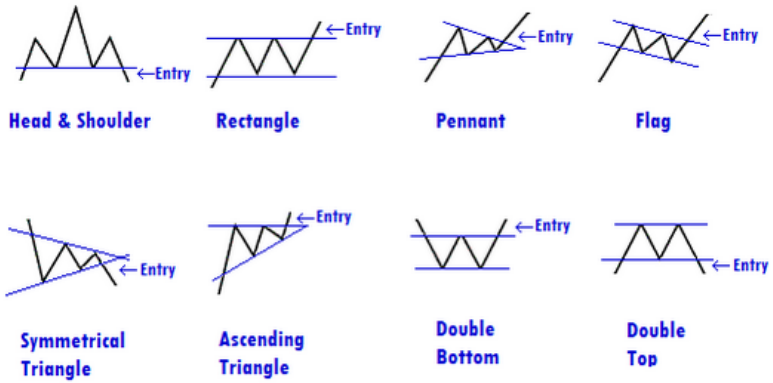
Del Dato al Conocimiento

Las técnicas de aprendizaje no supervisado buscan patrones en los datos sin disponer de información previa del objetivo del análisis.

Datos



Patrones/Similitudes/Cercanía



Finalidad

- Encontrar Patrones
- Agrupar Elementos
- Reducir Dimensionalidad

Agrupamiento (Clustering)

Las técnicas de agrupamiento tienen como objetivo asociar los elementos entre si, mediante distancias o métricas de similitud, conformando grupos de elementos con características comunes.

Jerárquicos

Descomposición jerárquica del conjunto de datos. Dos aproximaciones, bottom-up mediante agrupamiento de elementos y top-down mediante la división del conjunto total.

- **Jerárquico Aglomerativo**
- **Jerárquico Divisivo**

Basados en Centroides

Crean participaciones sucesivas del conjunto de datos hasta estabilizarlo en el óptimo de una métrica fijada. Normalmente, utilizan algoritmos heurísticos que convergen a óptimos locales.

- **K-means**
- **K-medians**
- **Fuzzy C-means**

Basados en Modelos

Se utilizan modelos que se ajusten a los datos. Dos tipologías, modelos basados en distribuciones estadísticas o modelos basados en la densidad de los puntos en el espacio.

- **GMM (distribuciones)**
- **DBSCAN (densidad)**

Para conjuntos de datos de tamaño elevado se aconseja utilizar técnicas de pre-clustering como **Canopy**

Clustering Jerárquico Aglomerativo

Objetivo

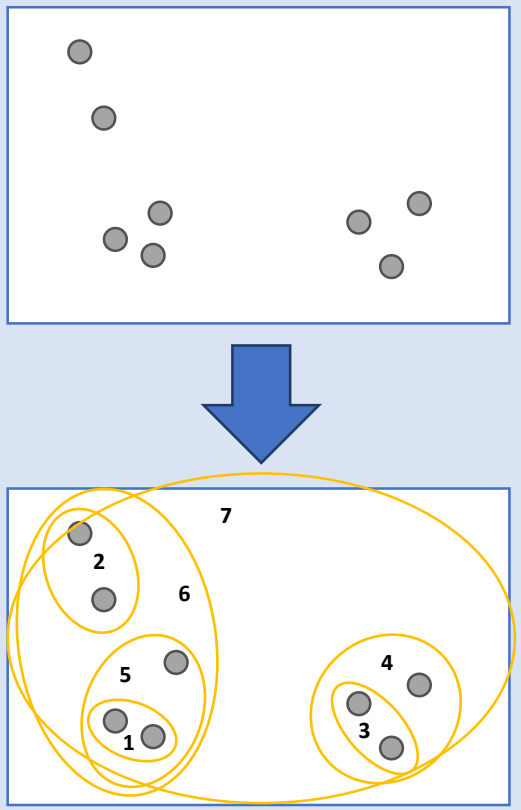
Agrupar los elementos, en una estrategia bottom-up, en función a su parecido (medido mediante una distancia) hasta conseguir tenerlos todos agrupados en un único grupo.

Desarrollo

Se comienza con todos los elementos representados en un espacio vectorial. Se elige una distancia (por defecto distancia euclídea) y se agrupan los dos elementos cuya distancia sea menor (este nuevo grupo pasa a ser un elemento en la siguiente iteración).

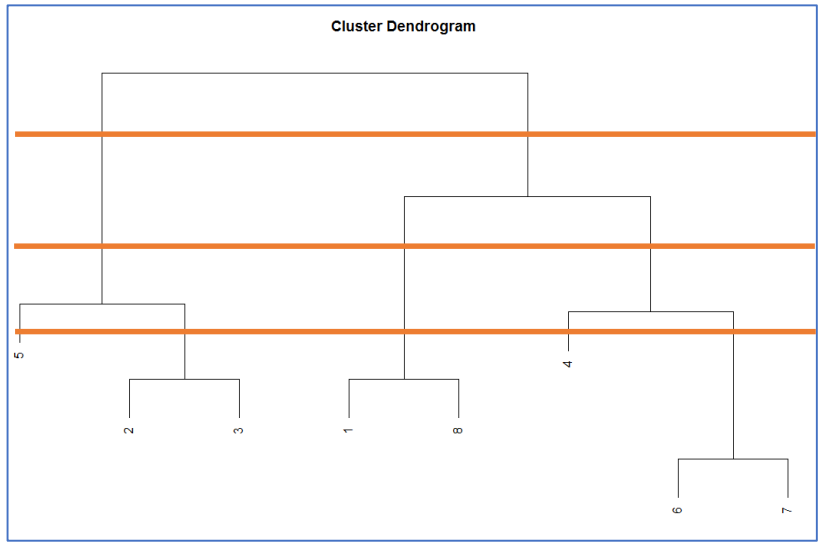
Se forma iterativa se van agrupando dos elementos hasta que todos conforman un único elemento.

Para agrupar dos elementos se calcula la distancia entre ambos, por defecto (método completo) es el máximo de la distancia entre las parejas de puntos, uno de cada elemento.



Clustering Jerárquico Aglomerativo: Dendograma

Representación Gráfica

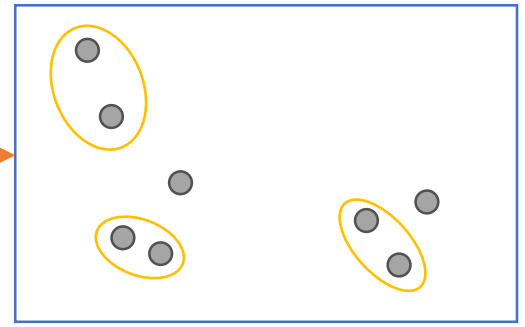


K=2

K=3

K=5

Clusters



Clustering K-means

Objetivo

Agrupar los elementos en un número prefijado k de clusters, mediante una estrategia iterativa.

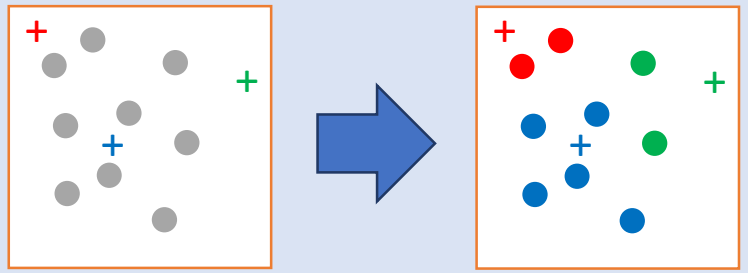
Desarrollo

Se seleccionan al azar (importante fijar semilla para su reproducibilidad) k elementos (llamados centroides) del espacio dimensional y mediante un algoritmo iterativo de dos fases se transforman esos k elementos hasta su convergencia, es decir, en las iteraciones el cambio de los centroides es menor a un umbral. Estos k elementos representan y permiten calcular los clusters.

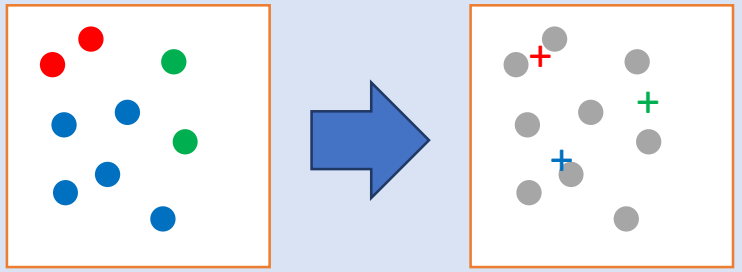
La primera fase del algoritmo es la asignación de elementos a los centroides. Cada elemento se asigna al centroide más próximo mediante una distancia (por defecto distancia euclídea).

La segunda fase del algoritmo es el cálculo de los nuevos centroides. Por defecto, los nuevos centroides son el centro de masas del los elementos asignados y se calcula como la media aritmética de sus componentes.

Asignación de Elementos



Nuevos Centroides



Aprendizaje no Supervisado

Reducción de Dimensionalidad

Las técnicas de reducción de dimensionalidad tienen como objetivo encontrar un nuevo sistema de referencia de los datos donde las nuevas dimensiones van ordenadas por información contenida con el objetivo de seleccionar un número reducido de dimensiones que aglutine la máxima información.

- Componentes Principales
- Singular Value Decomposition
- Low Rank Matrix Aproximation

Componentes Principales

Objetivo

Determinar un nuevo sistema de referencia formado por variables incorreladas linealmente con el objetivo de seleccionar las que presentan la mayor variabilidad

Desarrollo

Se construye la matriz de correlaciones de las variables originales, ésta es simétrica por lo que es diagonalizable y definida positiva por lo que todos los valores propios son positivo.

Se pueden seleccionar las m nuevas variables sintéticas asociadas a los valores propios mayores. Estas variables son incorreladas y acumulan un porcentaje de variabilidad igual a la suma de los valores propios dividido entre el número de variables n.

Los datos originales se representan en el sistema de referencia formado por las m variables sintéticas por lo que la dimensión del dataset pasa de n a m perdiendo la menor variabilidad posible.

Matriz de Correlaciones

$$\sum = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i) * Var(X_j)}}$$

Nuevo Sistema de Referencia

$$\sum = P\Delta P^{-1}, siendo\ \Delta\ diagonal$$

KSCHOOL

Del Dato al Conocimiento

Aprendizaje no Supervisado

Antonio Pita Lozano

Máster en Data Science

https://www.linkedin.com/in/antoniopitalozano/

@anto_pita