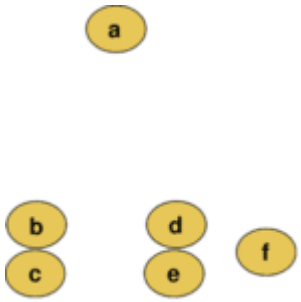
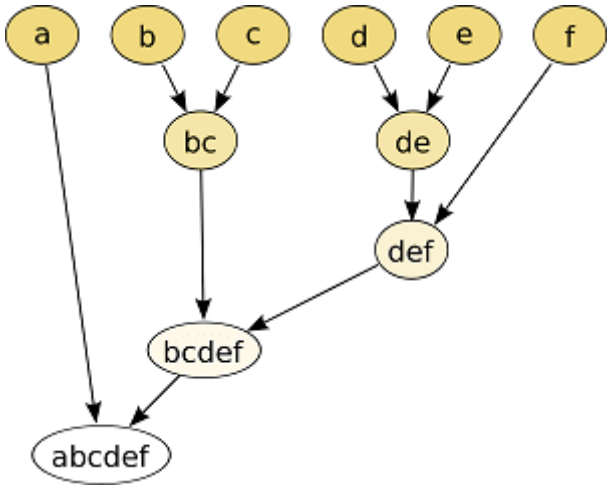
Clustering Jerárquico

Que es

- 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. Lo mejor para explicarlo es una imagen. Así que para ilustrar mejor este tema de agrupación en categorías voy a retomar un ejemplo gráfico muy difundido – y a la vez es el más descriptivo que he encontrado – que es el que exponen en la Wikipedia.
- En la primera imagen vemos cómo están distribuidos los datos y a qué distancia se encuentran unos de otros. En la segunda, vemos un ejemplo de clustering jerárquico dónde los datos se agrupan en función de la distancia (en este caso distancia euclidiana) entre ellos.





- los algoritmos de agrupamiento jerárquico están dentro de la categoría de **algoritmos de aprendizaje no supervisado**.
- Recordamos que en los algoritmos de aprendizaje no supervisado la data no está etiquetada previamente y solo se cuenta con variables independientes (las características de los datos). A partir de esto, los algoritmos de aprendizaje no supervisado buscan patrones en los datos sin hacer una predicción específica como objetivo (no hay variable dependiente).

Como funciona

- Se pueden definir dos tipos de clustering jerárquico dependiendo de la dirección en la que el algoritmo ejecute el agrupamiento:
  - **Tipo Aglomerativo**: Empezamos a agrupar desde cada elemento individual. Al inicio cada punto o dato está en un clúster separado. A cada paso, los dos clústeres más cercanos se fusionan. Estas fusiones de clústeres se siguen produciendo de forma sucesiva produciendo una jerarquía de resultados de clustering. Al final del proceso solo queda un único clúster que aglutina todos los elementos.
  - **Divisible**: Comenzamos a la inversa, partimos de un único clúster que aglomera todos los datos y vamos dividiendo en clústers más pequeños

Medir la distancia

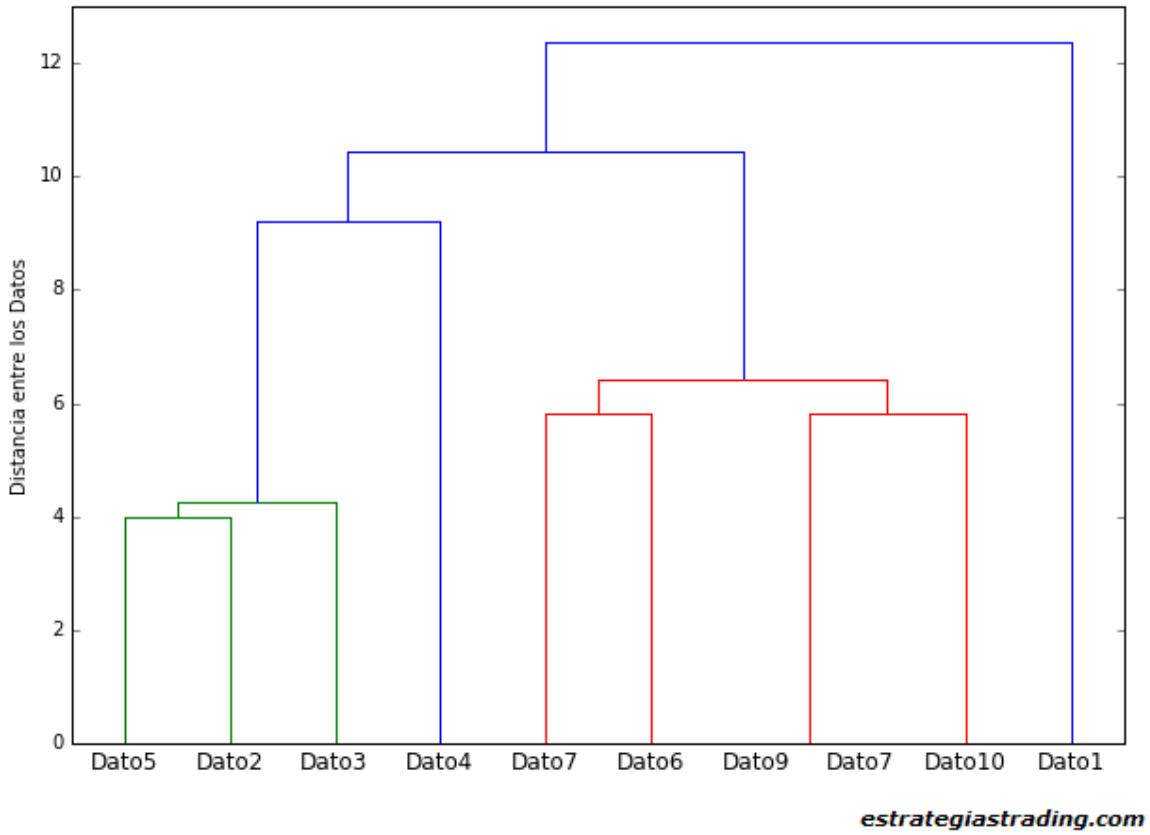
Además de la dirección en la que se haga la agrupación, también tenemos que definir cómo vamos a medir la distancia entre clústeres. Podemos tener 4 tipos de medidas de distancia entre clústeres:

Conexión completa: La distancia se mide entre los dos puntos más lejanos de cada clúster

Conexión simple: Es la opuesta a la conexión completa. Toma la distancia mínima entre dos puntos de cada clúster.

Distancia entre medias: La distancia entre dos clústeres se calcula como la distancia entre las medias de cada uno.

La distancia promedio entre pares: Es el promedio entre todas las distancias que podemos obtener entre todos los pares de puntos.



Las líneas verticales del dendrograma ilustran las fusiones (o divisiones) realizadas en cada etapa del clustering. Podemos ver la distancia, los distintos niveles de asociaciones entre los datos individuales y también las asociaciones entre clústers.

Comentarios sobre el clustering Jerárquico

- A diferencia del método de K-Means, el cuando efectuamos un clustering jerárquico no es necesario determinar previamente el número de clústeres que vamos a formar. Podemos obtener el número de clústeres directamente del modelo.
- No es apropiado para datasets muy grandes.
- Al igual que con otros métodos de minería de datos es importante recordar que debemos estandarizar las variables antes de aplicar el algortimo.
- La elección de la medida de distancia es muy importante para que el clustering tenga sentido. Con distintas medidas tendremos obtendremos distintos agrupamientos.

Video

https://youtu.be/T76paW6fJBI