Agrupacion por Jerarquias

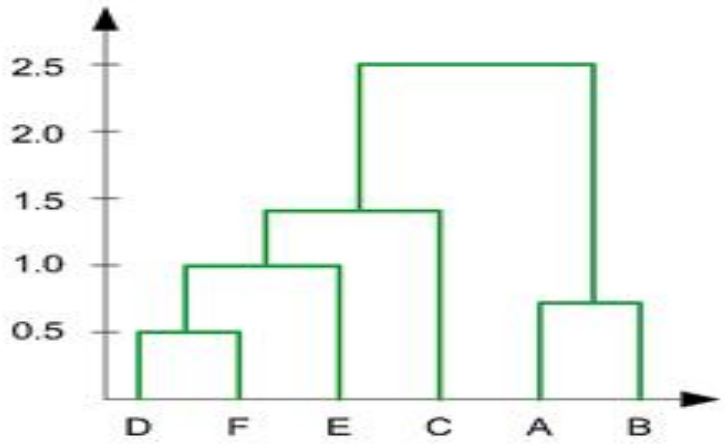
Jose Antonio Mejia

August 2018

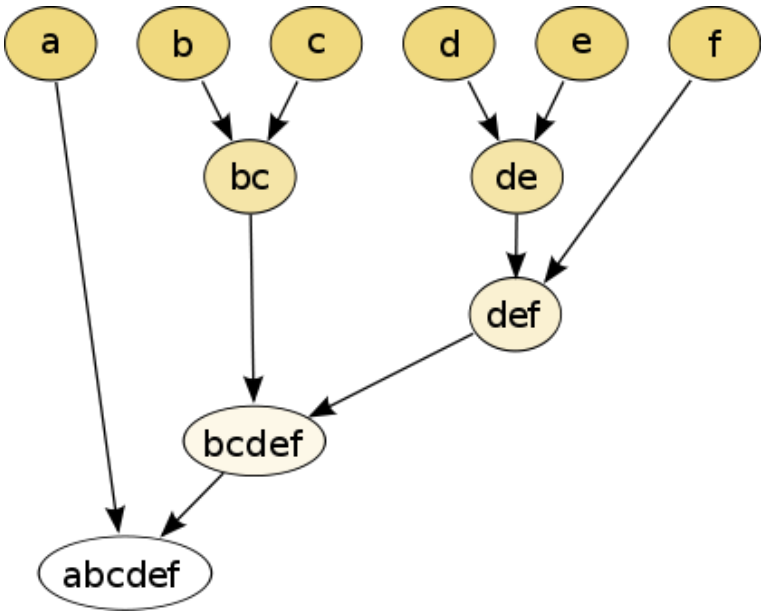
1 Que son?

Los algoritmos de agrupación jerárquica en realidad se dividen en 2 categorías: de arriba hacia abajo o de abajo hacia arriba. Los algoritmos ascendentes tratan cada punto de datos como un único grupo al principio y luego fusionan (o aglomeran) pares de clústeres hasta que todos los clústeres se hayan fusionado en un solo clúster que contenga todos los puntos de datos. La agrupación jerárquica ascendente se denomina, por lo tanto, agrupamiento aglomerativo jerárquico o HAC. Esta jerarquía de conglomerados se representa como un árbol (o dendrograma). La raíz del árbol es el clúster único que reúne todas las muestras, siendo las hojas los grupos con solo una muestra.

1.1 Dendrograma



1.2 Arbol



2 Criterios de vinculación

El criterio de vinculación determina la distancia entre conjuntos de observaciones en función de las distancias por pares entre las observaciones. Algunos criterios de vinculación de uso común entre dos conjuntos de observaciones A y B son:

2.0.1 Agrupación de enlaces máximos o completos

$$\max \{\, d(a,b) : a \in A,\, b \in B \,\} \tag{1}$$

2.0.2 Agrupación mínima o de enlace único

$$\min \{\, d(a,b) : a \in A,\, b \in B \,\} \tag{2}$$

Esta sera la que yo usare para realizar el proyecto.

2.0.3 Agrupación media o media de enlaces, o UPGMA

$$\frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a,b) \tag{3}$$

3 Pasos

1. Comenzamos tratando cada punto de datos como un único clúster, es decir, si hay X puntos de datos en nuestro conjunto de datos, entonces tenemos X grupos. Luego seleccionamos una métrica de distancia que mide la distancia entre dos grupos. Como ejemplo, utilizaremos una vinculación promedio que define la distancia entre dos grupos como la distancia promedio entre los puntos de datos en el primer grupo y los puntos de datos en el segundo grupo.

2. En cada iteración, combinamos dos clusters en uno. Los dos clusters que se combinarán se seleccionan como aquellos con el enlace promedio más pequeño. Es decir, de acuerdo con nuestra métrica de distancia seleccionada, estos dos grupos tienen la menor distancia entre ellos y, por lo tanto, son los más similares y deben combinarse.

3. El paso 2 se repite hasta que lleguemos a la raíz del árbol, es decir, solo tenemos un clúster que contiene todos los puntos de datos. De esta forma, podemos seleccionar cuántos clústeres queremos al final, simplemente eligiendo cuándo dejar de combinar los clusters, es decir, cuando dejamos de construir el árbol.

4 Links

https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68 https://github.com/shubhamjha97/hierarchical-clusteringintroduction https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/