

Agrupacion Aglomerativa Jerarquica

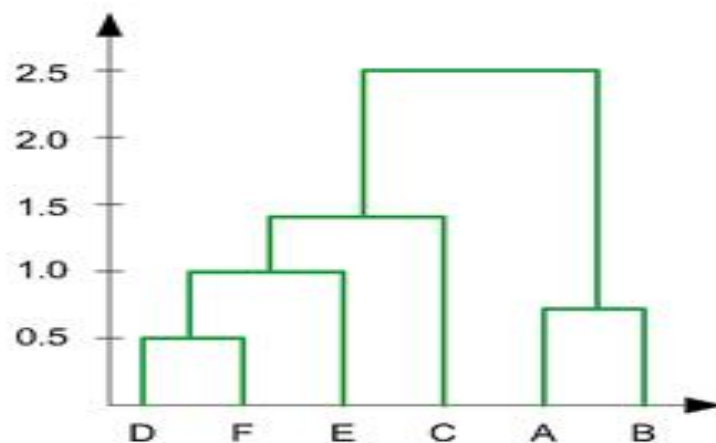
Jose Antonio Mejia

18 de Septiembre del 2018

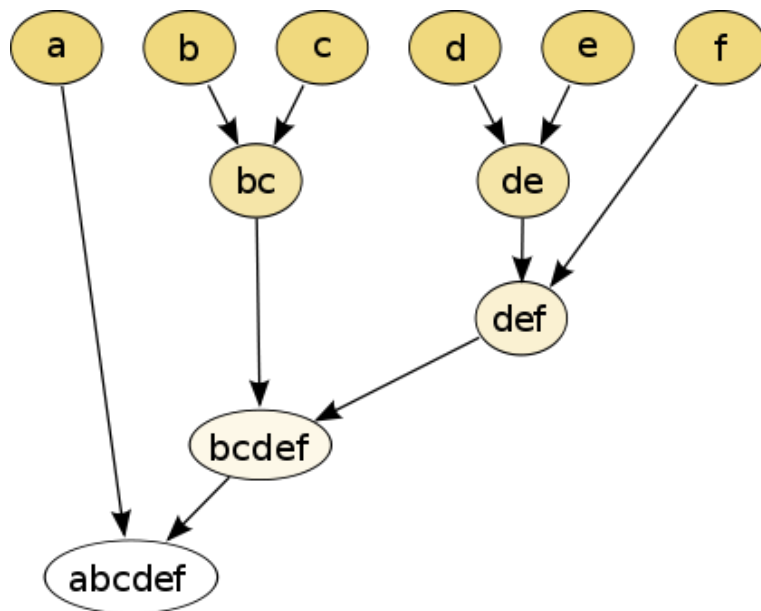
1 Que son los algoritmos de agrupacion jerarquica.

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 Cual es su utilidad?

Los algoritmos de agrupación jerárquica son utilizados en gran medida para ver que tan relacionados están los datos que analizarán como se veo esto? Gracias a que estos algoritmos fueron diseñados para crear un dendrograma con el resultado final, en este se puede ver que tan relacionados están los datos en base a sus distancias por poner un ejemplo. Un ejemplo más claro es que tengamos a personajes de los Simpsons por el ejemplo Bart, Lisa, Skinner, Homero y Moe. Cualquiera que vio o ve los Simpson sabe que Bart y Lisa están relacionados por ser hermanos y ellos a Homero ya que es su padre, luego sabemos que Homero y Moe están relacionados ya que Moe es un amigo cercano a la familia de Homero y por último el que menor relación tiene con todos ellos es Skinner ya que solo es el director de la escuela donde Bart y Lisa van. Para demostrar mejor esto utilizaremos más abajo los datos que di anteriormente y usando este ejemplo para que se entienda mejor el cómo funciona el algoritmo. Pero en pocas palabras sirve para determinar que datos están más relacionados en un conjunto de datos.

3 Criterios de vinculación y distancia euclidiana

El primer paso para realizar este algoritmo es generar una matriz de distancias. Para generar esta matriz tendremos que usar la distancia euclidiana que es una fórmula matemática para determinar la distancia entre dos puntos en un espacio euclideo.

$$d_E(P, Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} \quad (1)$$

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:

3.0.1 Agrupación de enlaces máximos o completos

$$\max \{ d(a, b) : a \in A, b \in B \} \quad (2)$$

3.0.2 Agrupación mínima o de enlace único

$$\min \{ d(a, b) : a \in A, b \in B \} \quad (3)$$

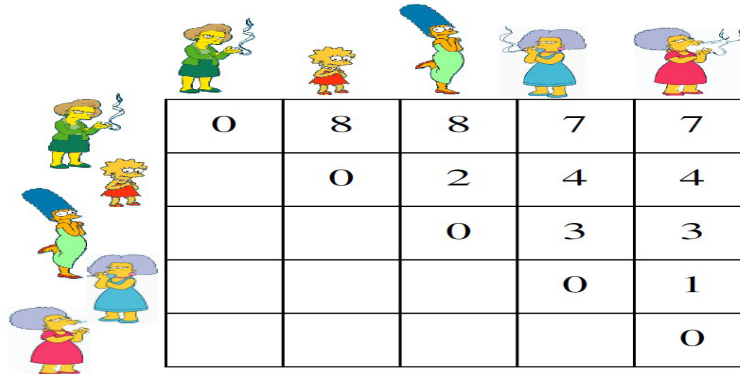
3.0.3 Agrupación media o media de enlaces, o UPGMA

$$\frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b) \quad (4)$$

4 Pasos

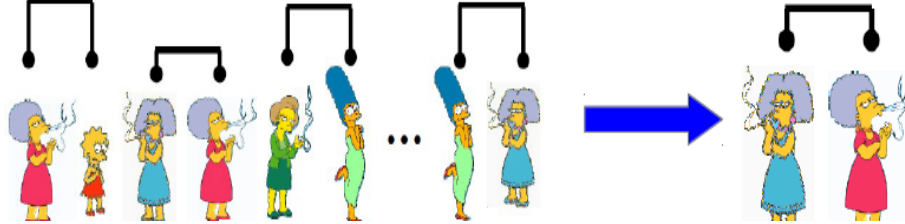
1. El primer paso es utilizar la formula matematica de distancia euclidea para sacar la matriz de distancia.

El segundo paso es usar criterio de enlace unico. Pondre el ejemplo de los Simpson de nuevo para entender mejor el como funciona. Teniendo los siguientes resultados de la matriz de distancia ya dada:



0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0

2. Teniendo esta tabla primero que todo comparamos que valores estan mas cercanos comparando cada personaje con cada uno:



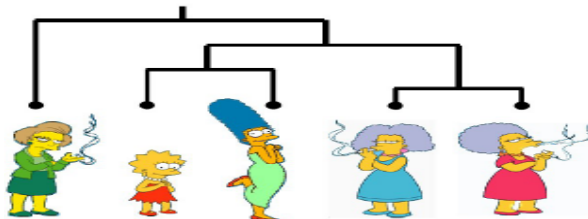
3. Una vez terminado de comparar cada uno de los datos vemos que las 2 hermanas de marge estan muy relacionadas(cercanas) que el resto a compartir datos muy similares por lo que una vez hecho esto se combinan los datos y se crea 1 cluster que las contiene a las 2. En esta parte es donde principalmente se utiliza la Agrupacion minima para determinar que valores seran los que quedaran al combinar a las hermanas de marge creando un solo cluster que sera de la siguiente forma.

$$\min d(7, 7) : 7 \min d(4, 4) : 4 \min d(3, 3) : 3 \min d(0, 1) : 0 \quad (5)$$

4. Al aplicar esto la tabla resultaria de este modo:

	0	8	8	7
		0	2	4
			0	3
				0

5. Este paso se repite sucesivamente hasta que ya no hayan datos por analizar dando como resultado el siguiente dendograma:



Como podemos observar al final vemos que los que estan mas relacionados

son Marge, Lisa y las hermanas de Marge y que al final la maestra de Lisa es la que menos relacionada esta con ese grupo y cualquiera que vea la serie de los Simpson estara de acuerdo a que las relaciones en el dendograma estan correctas.

5 Conclusiones

- El Algoritmo de Agrupamiento Aglomerativo Jerárquico es excelente al momento de mostrar resultados gracias a la creacion del dendograma pero tiene dos principales desventajas:
- la primera es que tiene un gran acumulacion de errores ya que si la matriz es demasiado grande y al momento de agrupar hay un minimo error este se propaga durante el resto de la construccion del dendograma sin ser posible repararlo.
- La segunda desventaja es que requiere demasiada memoria ya que al tratar con un conjunto de datos mayor este muestra su mal rendimiento.

6 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/>