Clustering Jerárquico

Clustering Jerárquico

Clustering Jerárquico

Permite obtener una estructura de árbol o jerárquica de grupos. Existen dos estrategias de identificación de grupos: Divisiva y Aglomerativa.

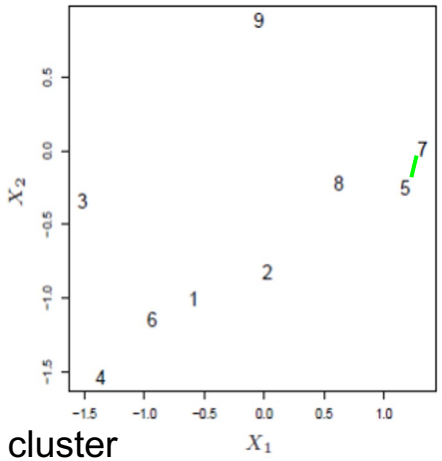
1. Divisiva: Se comienza con un único gran grupo, el cual se va dividiendo hasta llegar a tantos grupos como observaciones tiene el conjunto de datos.

1. Aglomerativa: Se comienza con tantos grupos como puntos existen en el conjunto de datos. Luego, estos se van mezclando hasta llegar a un solo gran grupo.

El resultado final del proceso se realiza usando un diagrama denominado Dendrograma.

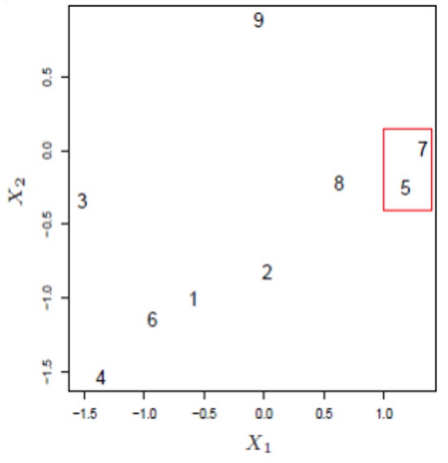
Clustering Jerárquico

n samples
n clusters



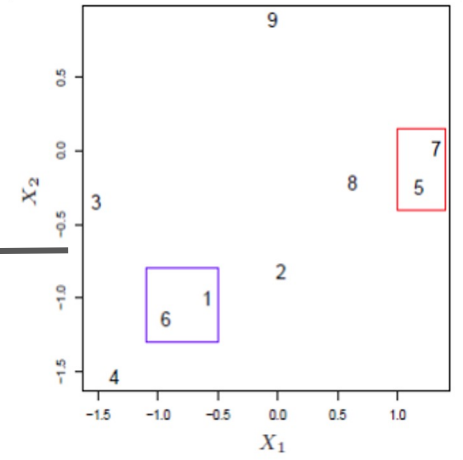
medida de distancia entre samples (‘affinity’), usualmente la euclídea

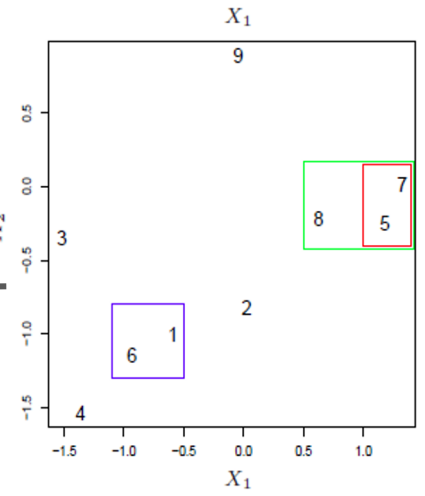
junto los dos cluster que están a menor distancia



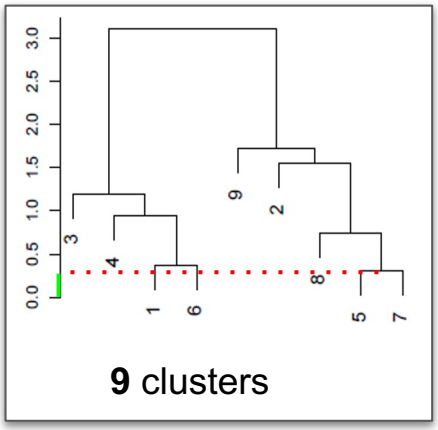
n-1 clusters

distancia entre clusters de >=1 elementos (‘linkage’)





(...)



9 clusters

1 cluster

linkage

Dendograma

Clustering Jerárquico

En el corazón de estos métodos se encuentra la necesidad de medir distancia (o similitud) entre clusters (representado por un conjunto de puntos).

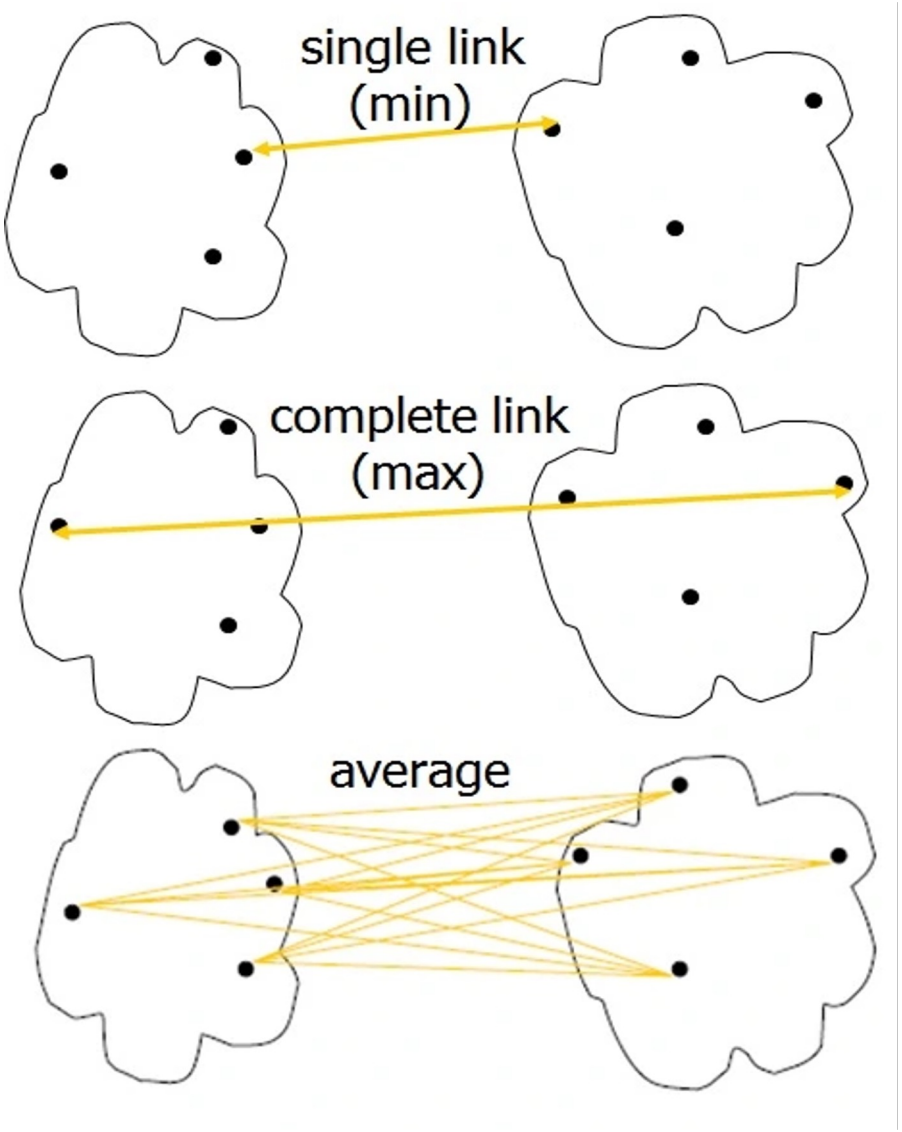
Existen varios métodos. Ejemplos de estos son:

- Distancia Mínima (Single-linkage)
- Distancia Máxima (Complete-linkage)
- Distancia Promedio

Nombres	Fórmula
Agrupamiento de máximo o completo enlace	$\max\{\, d(a,b) : a \in A,\, b \in B \,\}.$
Agrupamiento de mínimo o simple enlace	$\min\{\, d(a,b) : a \in A,\, b \in B \,\}.$
Agrupamiento de enlace media o promedio, o UPGMA	$\frac{1}{\lvert A\rvert\lvert B\rvert}\sum_{a\in A}\sum_{b\in B} d(a,b).$

En la siguiente slide se explica lo mismo, pero con un diagrama.

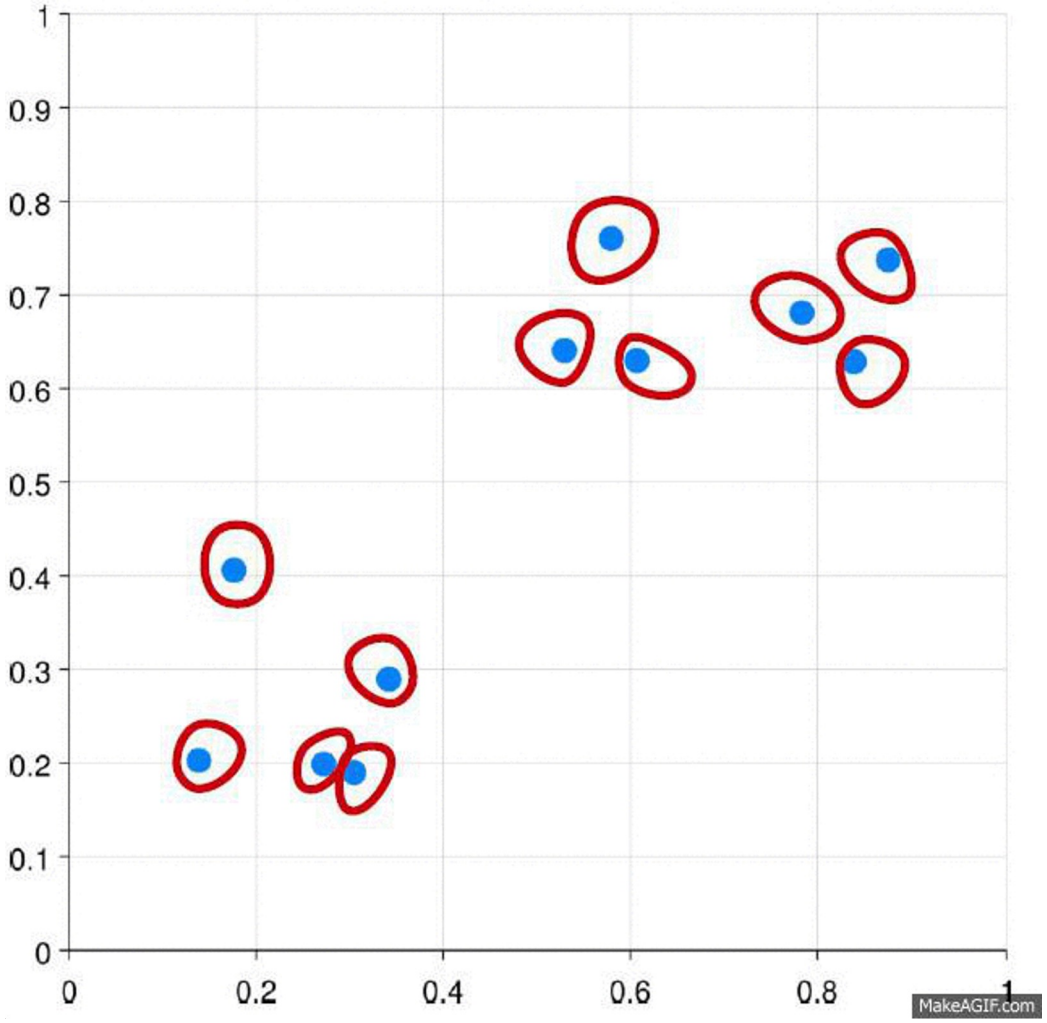
Clustering Jerárquico



Fuente: https://mlalgorithm.wordpress.com/2016/06/08/hierarchical-clustering/

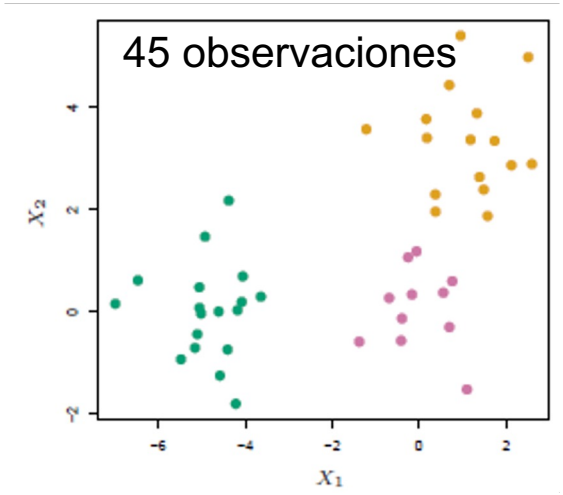
Clustering Jerárquico

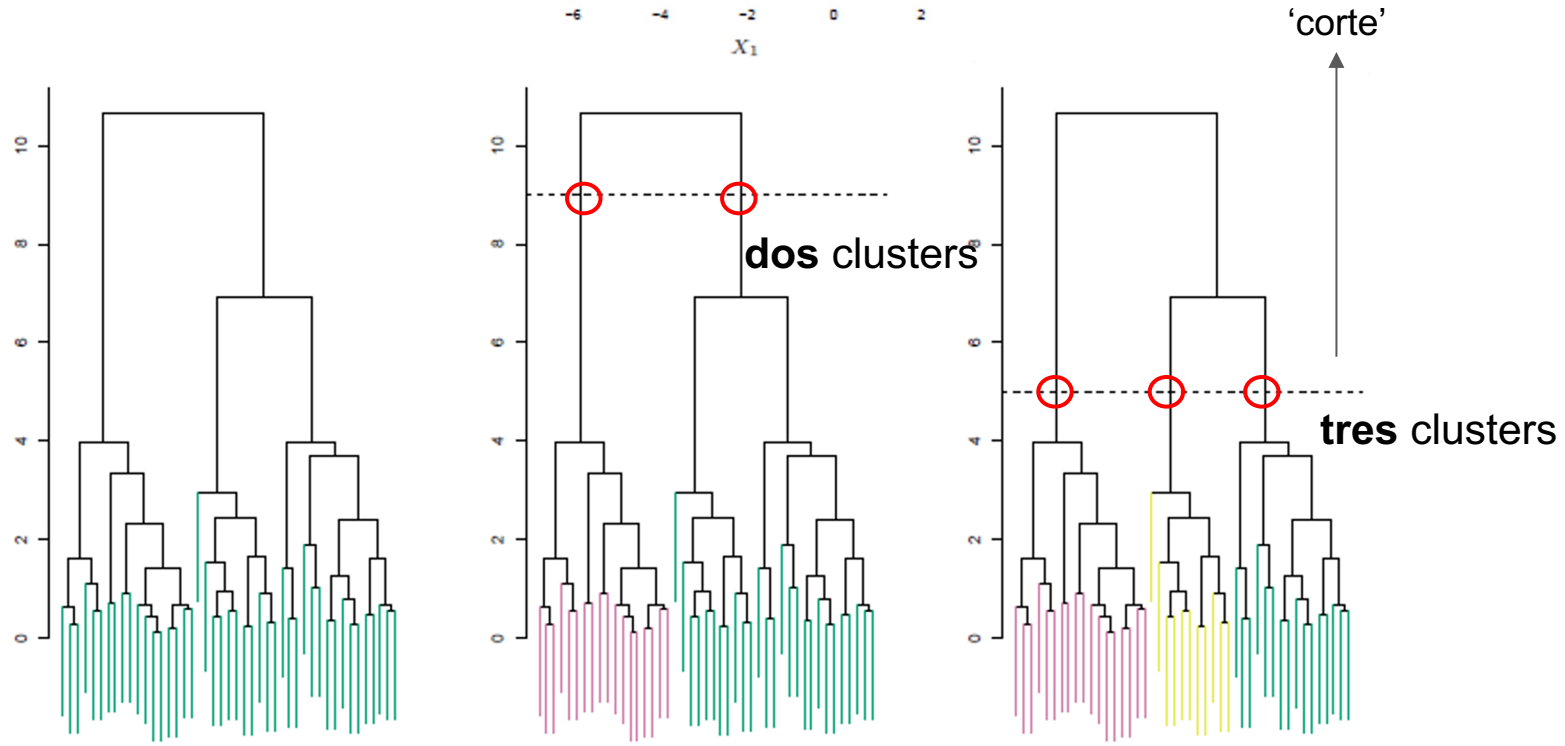
Gif animado



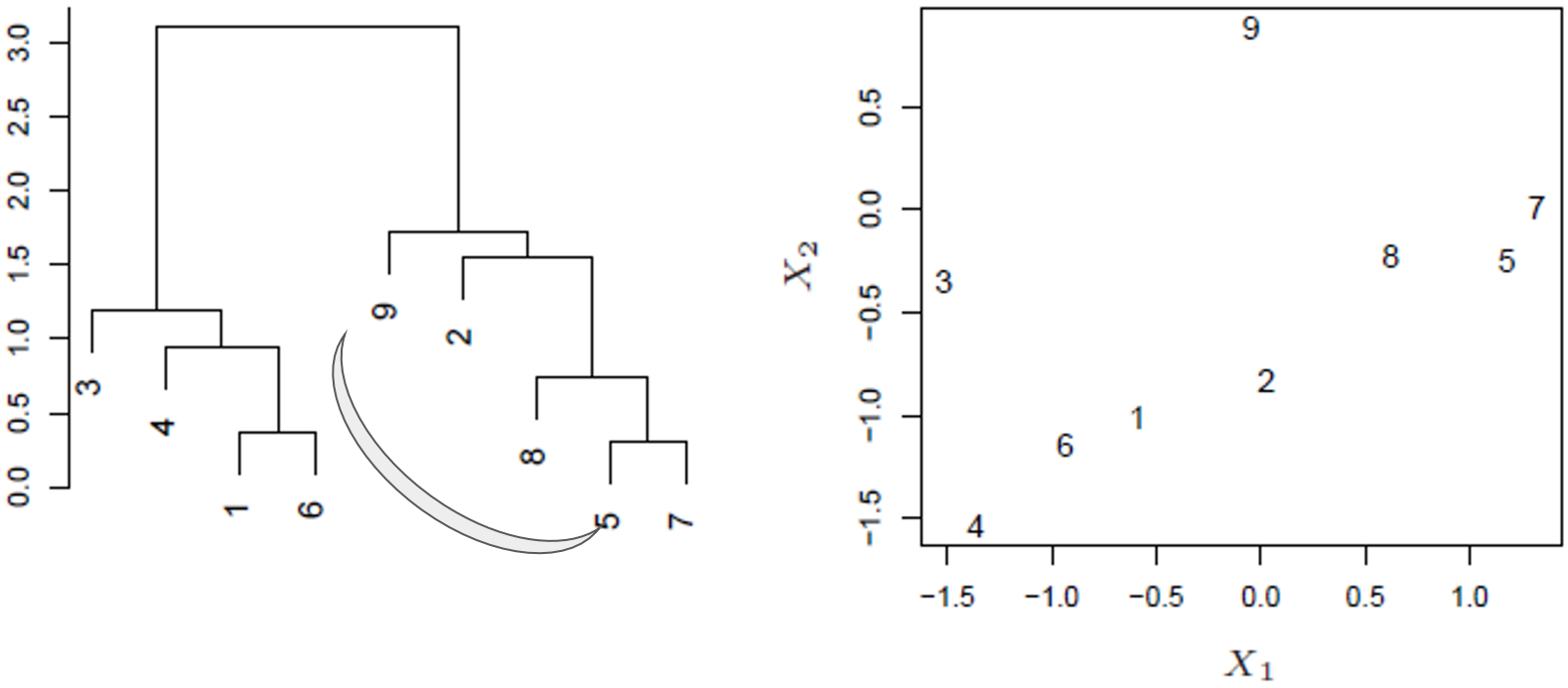
Fuente: https://mlalgorithm.wordpress.com/2016/06/08/hierarchical-clustering/

Clustering Jerárquico





Clustering Jerárquico



Clusters que se obtienen cortando el dendograma a una dada altura están anidados con los clústeres que se obtienen al cortar el dendograma en una altura superior

Clustering Jerárquico

Ventajas y desventajas

+ Pueden revelar detalles finos en la relación de los datos

+ Proveen un dendograma interpretable

+ Son determinísticos - producen el mismo resultado si se corre el mismo modelo con el mismo input

- Son computacionalmente costosos

En Scikit Learn:

sklearn.cluster.AgglomerativeClustering