Hierarchical Clustering

Agenda

- Hierarchical Clustering
- Linkage methods
- Hierarchical Agglomerative Clustering
- Dendrogram
- Interpretation of Dendrogram
- Hierarchical Divisive Clustering
- Key points

Hierarchical Clustering

- Type of clustering to group objects in cluster based on their similarity

- Hierarchical Agglomerative Clustering - it’s a bottom-up approach to build clusters. Algorithm starts with one cluster and merges nearest objects or clusters until one big cluster is formed

- Hierarchical Divisive Clustering - it’s a top-down approach to build clusters. It is inverse of agglomerative clustering

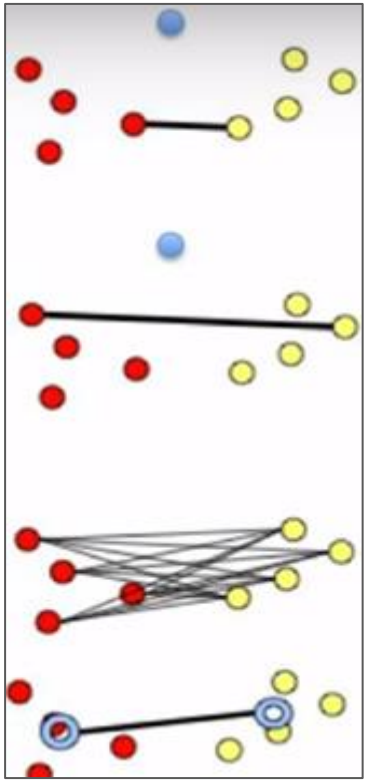
Linkage Methods

- In order to decide which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a measure of dissimilarity between sets of observations is required
- Linkage criterion specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets.
- Choice of distance metric and linkage criteria may influence the final results

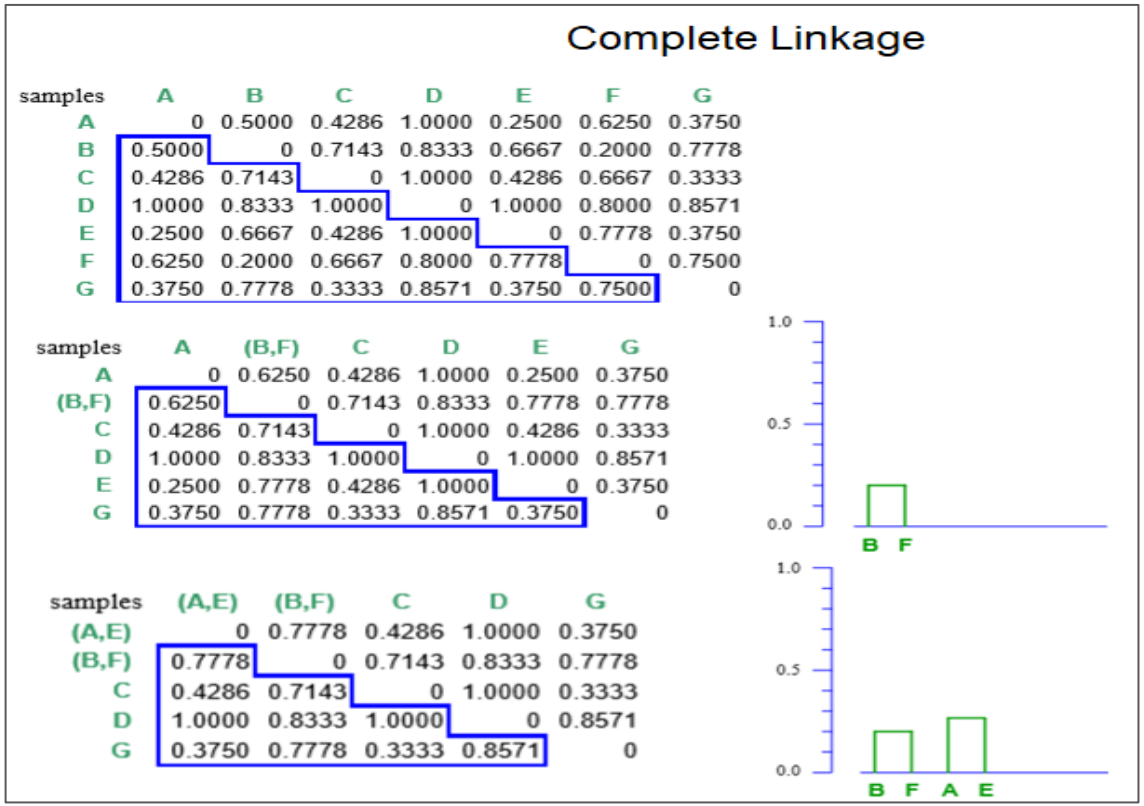
Hierarchical Agglomerative Clustering

- Common inter cluster distance measurement techniques are:

1. Single linkage: minimum distance between closest data points from the two clusters is considered
2. Complete linkage: distance between two farthest data points from the two clusters is considered
3. Average linkage: average distance is considered
4. Centroid distance: distance between centroid of different clusters is considered



Hierarchical Agglomerative Clustering

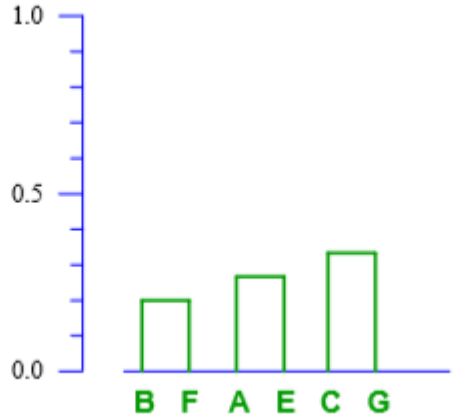


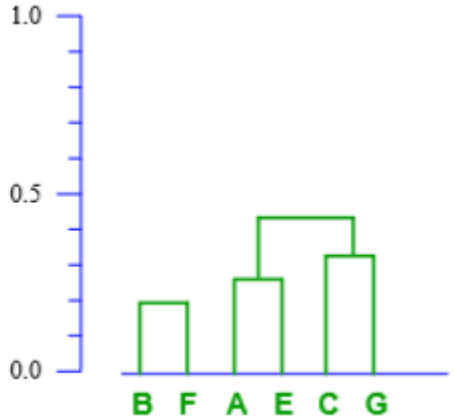
Complete linkage is used here

- We have 7 samples (initial 7 clusters)
- Distance (using complete linkage) is calculated between all the data points (Clusters) and displayed in a matrix
- Closest points are merged into one cluster (B,F)
- Plot the graph
- Repeat the above steps until only one cluster is left

Hierarchical Agglomerative Clustering

samples	(A,E)	(B,F)	(C,G)	D
(A,E)	0	0.7778	0.4286	1.0000
(B,F)	0.7778	0	0.7778	0.8333
(C,G)	0.4286	0.7778	0	1.0000
D	1.0000	0.8333	1.0000	0

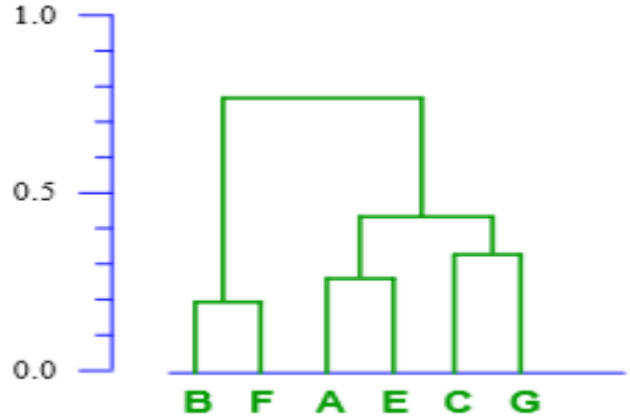


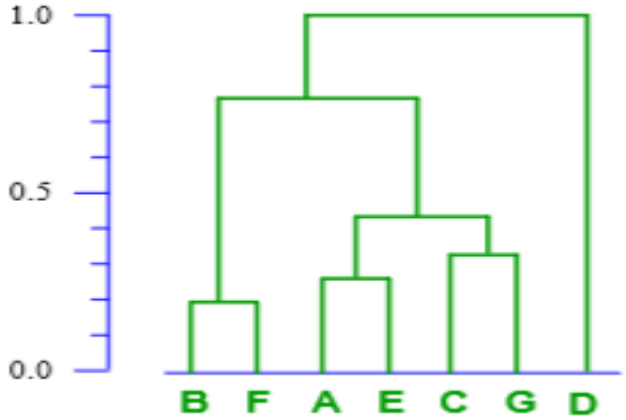


Hierarchical Agglomerative Clustering

samples	(A,E,C,G)	(B,F)	D
(A,E,C,G)	0	0.7778	1.0000
(B,F)	0.7778	0	0.8333
D	1.0000	0.8333	0

samples	(A,E,C,G,B,F)	D
(A,E,C,G,B,F)	0	1.0000
D	1.0000	0

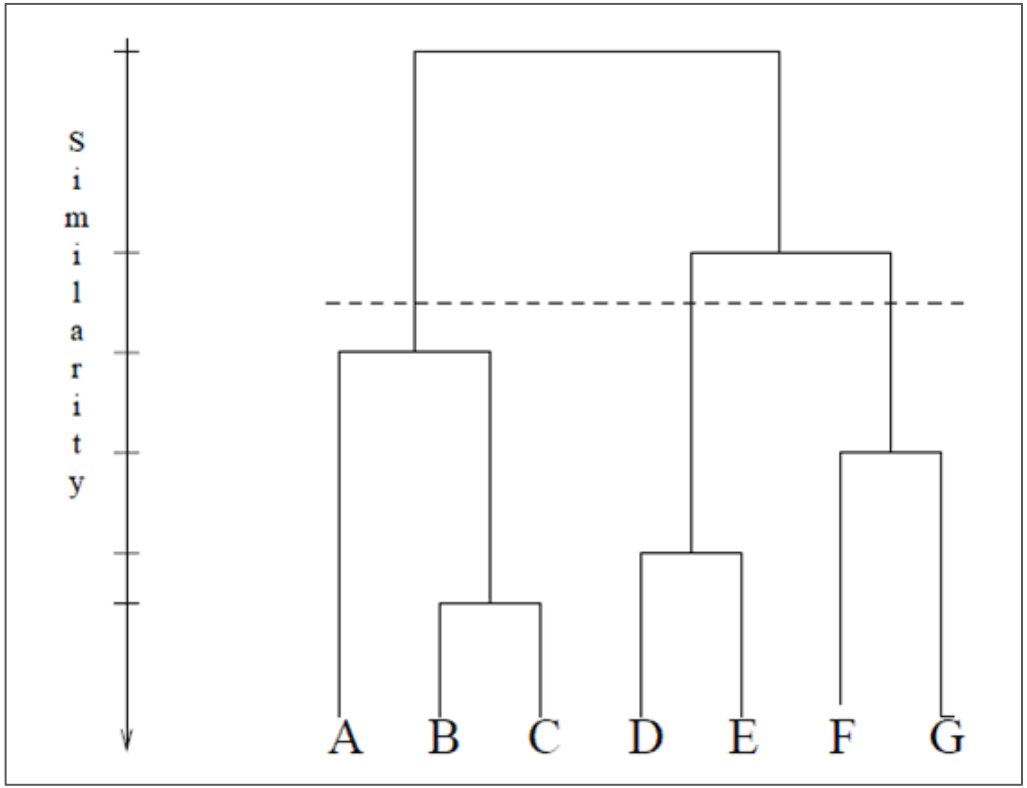




Dendrogram

- Dendrograms are used to represent the distances at which the different clusters meet.
- They provide us an idea as to how the clustering looks like diagrammatically
- Dendrograms for the same dataset change based on the method chosen to calculate distance between the clusters (linkage function) and distance functions

How to interpret a dendrogram



- The y-axis is a measure of closeness of either individual data points or clusters.
- Nearest clusters/data points are merged to make a bigger custer (B and C and near to each other)
- The vertical position of the split, shown by dashed gives the distance (dissimilarity) between the two clusters.
- Horizontal dashed lines at any point on x-axis gives the no. of clusters based on the number of cuts it makes.
  C1 = {A,B,C}
  C2 = {D,E}
  C3 = {F,G}

Hierarchical Divisive Clustering

- It’s a top-down approach to build clusters.
- It is inverse of agglomerative clustering
- It begins with root, in which all the objects are in one single cluster
- At each step of iteration, the most heterogeneous cluster is divided into two.
- The process is iterated until all objects are in their own cluster
- Same linkage methods are used to find similarity

Key Points

- Domain knowledge helps in selecting inter cluster distance metric
- Complete and average linkage are better choice if the clusters are likely to be spherical