Hierarchical Clustering (Agglomerative)

Prerequisite-

- Unsupervised learning
- Clustering

Objectives-

- Understanding hierarchical clustering
- Proximity matrix and Dendrogram
- Understanding the similarity measures for hierarchical clustering

Introduction to hierarchical clustering-

Hierarchical clustering method is based on hierarchy representation of clusters where parent cluster node is connected to further to child cluster node. A node represents collection of data points to one cluster. It is further divided into two types:

- Agglomerative Clustering
- Divisive Clustering

The **agglomerative clustering** is the most popular and common hierarchical clustering also known as Agglomerative Nesting (AGNES). The methods starts by considering each data point as a single cluster. In the next step the singleton clusters are merged into a bog cluster based on the similarity between them. The procedure is repeated until all the datapoints are merged into one big cluster. The procedure can be represented as hierarchy/tree of clusters.

The **divisive clustering** works completely opposite to agglomerative clustering and also known as Divisive Analysis (DIANA). The method starts from one big cluster considering all data points within it. In the next the big cluster is divided into the most heterogeneous two clusters. The procedure is repeated until each data point is in its own cluster. Figure on the next page is describe the hierarchical clustering methods.

[source: https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/]

Note- It should be noted that agglomerative clustering is good choice to identify small sized clusters whereas divisive clustering is more effective in case of big size clusters.

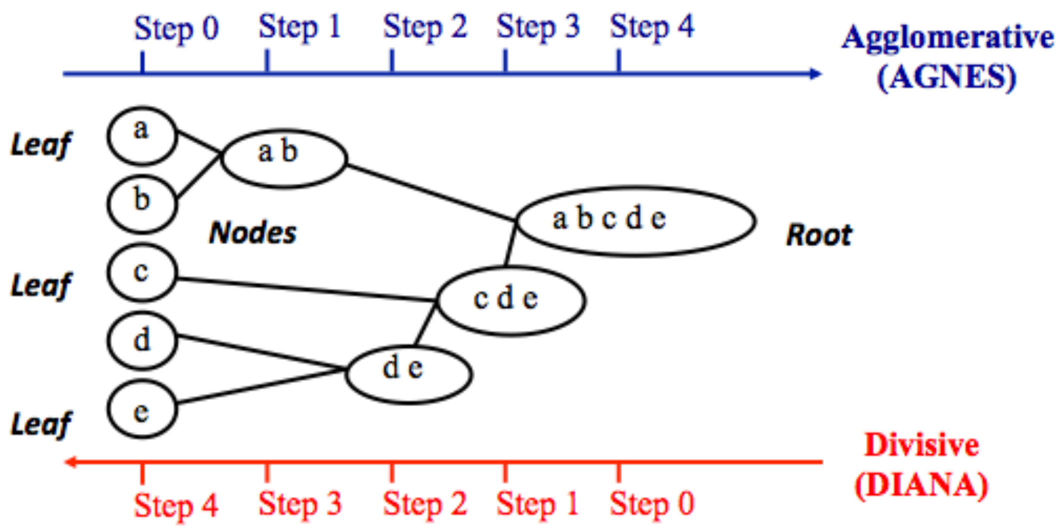
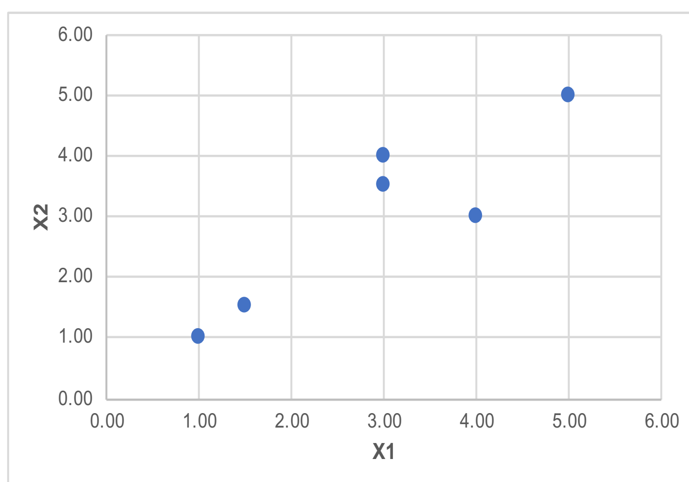


Figure 1 Hierarchical Clustering

Proximity Matrix-

Proximity matrix is defined as a square matrix whose each cell (i, j) is a some kind of distance/similarity measure between the entities shown by row i and column j. For example if we have given with following data sample with points {A, B, C, D, E, F} of two dimensional.

	X1	X2
A	1.00	1.00
B	1.50	1.50
C	5.00	5.00
D	3.00	4.00
E	4.00	3.00
F	3.00	3.50



The proximity between the data points can be calculated using any distance matric. Let the distance matric is Euclidean distance:

$$D_{euclidiean} = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2}$$

$$D_{CD} = \sqrt{(5-3)^2 + (5-4)^2} = 2.24$$

and the complete proximity matrix is:

	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.22	1.41	2.50
D	3.61	2.92	2.22	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Dendrogram-

A dendrogram is a pictorial way to visualize hierarchical clustering. It is mainly used to show the outcome of hierarchical clustering a tree like diagram that records the sequences of merges and splits.

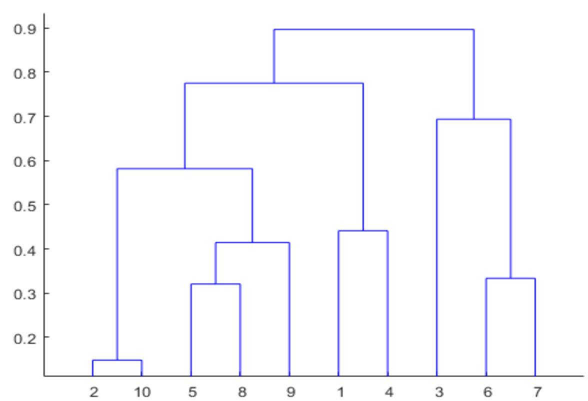


Figure 2 Dendrogram

The complete procedure of agglomerative clustering can be understood by example shown in the figure 1. The algorithm computes in following steps:

1. Consider each data points as cluster.
2. Calculate the proximity matrix.
3. Join the two closet clusters based on proximity value and recompute the proximity matrix.
4. Repeat step 3 until all points are merged into one cluster.

Let's we have five data points { a, b, c, d, e}

Step 0- In the first step, we consider each point into a individual cluster. Next step is to calculate the proximity matrix between the points.

Step 1- In step 1 let suppose point a and b closer to each other as compared to other points and formed one cluster by merging them together. Now we left with points { ab, c, d, e}

Step 2- The step 2 again we calculate proximity matrix with left out points. Let suppose again d and e are much closer than others so to merge into a cluster and left with the points { ab, c, de}.

Step 3- In step 3 once again calculate proximity matrix with remaining clusters. Let suppose cluster c and de are closer to each other and formed a new cluster cde and left with { ab, cde} clusters.

Step 4- Now we have left with only two cluster { ab, cde } so to merge them into one cluster as { abcde }.

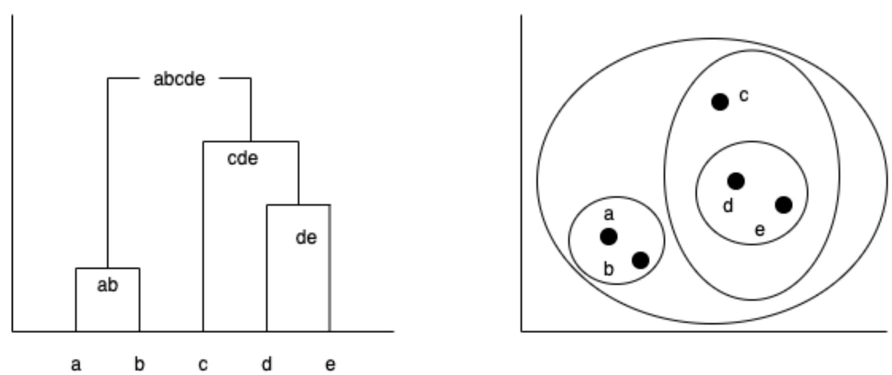


Figure 3 Agglomerative Clustering

Linkage-

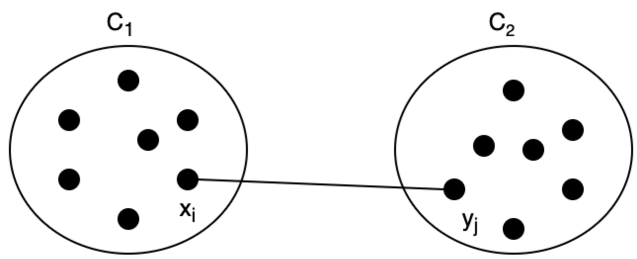
Linkage process merges two clusters into one cluster based on the distance or similarity between them. The similarity between two clusters is very important parameter for merging and dividing of cluster. Following methods are popularly used to calculate similarity between two clusters.

- Minimum or single linkage
- Maximum or complete linkage
- Mean or average linkage
- Centroid linkage
- Ward's method or minimum variance method

Minimum or single linkage method-

Minimum linkage between two clusters is defined as the minimum distance between all pairs of points within two clusters. This can be written as:

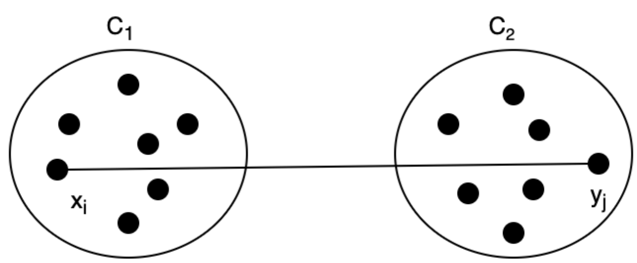
$$link = Min\{Dist(x_i, y_j)\} \quad where\ x_i \in C_1\ and\ y_j \in C_2$$



Maximum or complete linkage method-

Maximum linkage between two clusters is defined as the maximum distance between all pairs of points within two clusters. This can be written as:

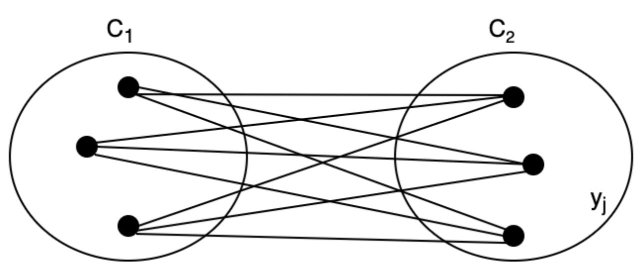
$$link = Max\{Dist(x_i, y_j)\} \quad where\ x_i \in C_1\ and\ y_j \in C_2$$



Mean or Average linkage method-

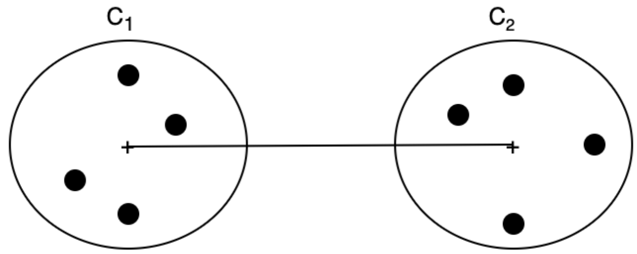
The distance between two clusters is calculated by taking mean of similarity among all pair of points.

$$link = \frac{\sum Dist(x_i, y_j)}{|C_1||C_2|} \quad where\ x_i \in C_1\ and\ y_j \in C_2$$



Centroid linkage method-

The distance between two clusters is measured as the distance between the centroids of the two clusters.



Ward's method-

The distance between two clusters is based on the similarity calculated as the sum of square of the of the distances x_i *and* y_j. This is very much similar to average method except it works on the sum of squares.

$$link = \frac{\sum Dist(\, x_i, y_j)^2}{|C_1||C_2|} \quad where\ x_i \in C_1\ and\ y_j \in C_2$$

Challenges of agglomerative clustering-

- The identification of distance measure.
- Identify the number of clusters.
- To calculate the similarity between clusters has its disadvantages.
- High space and time complexity for Hierarchical clustering. Hence this clustering algorithm cannot be used when we have huge data.

****