UNIT V

CLUSTER ANALYSIS

Introduction, Types of clustering, Correlations and distances, clustering by partitioning methods hierarchical clustering, overlapping clustering, K-means Clustering - Profiling and Interpreting Clusters.

Introduction:

Cluster analysis is a data exploration (mining) tool for dividing a multivariate dataset into "natural" clusters (groups). We use the methods to explore whether previously undefined clusters (groups) exist in the dataset.

For instance, a marketing department may wish to use survey results to sort its customers into categories (perhaps those likely to be most receptive to buying a product, those most likely to be against buying a product and so forth).

Cluster analysis is used when we believe that the sample units come from an unknown number of distinct populations or sub-populations. We also assume that the samples ~~that~~ # units come from a number of distinct populations but there is no apriori definition of those populations.

Our objective is to describe those populations with the observed data.

Cluster analysis:

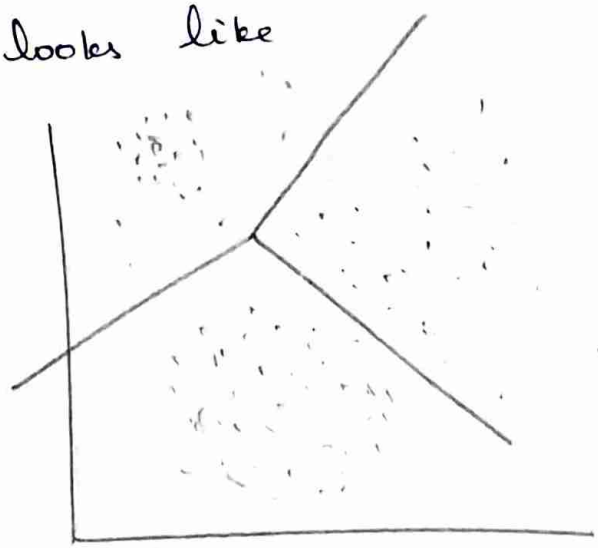
Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters.

Types of clustering:

- Partitioning methods
- Hierarchical ~~clustering~~ clustering
- Fuzzy clustering
- Density-Based clustering
- Model-Based clustering.

1. Partitioning clustering:

It is a type of clustering technique that divides the data set into a set number of groups.
It can be also called as a centroid based method.
In this approach cluster center is formed such that the distance of data points in that cluster is minimum when calculated with other cluster centroids.
A most popular example of this algorithm is the KNN algorithm. This is how partitioning clustering algorithm looks like

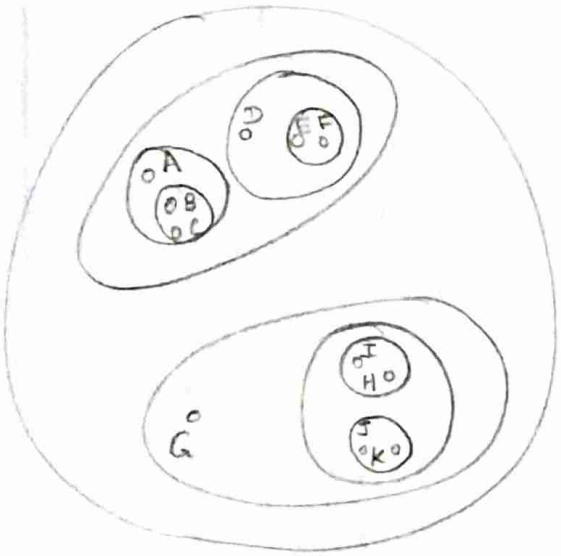


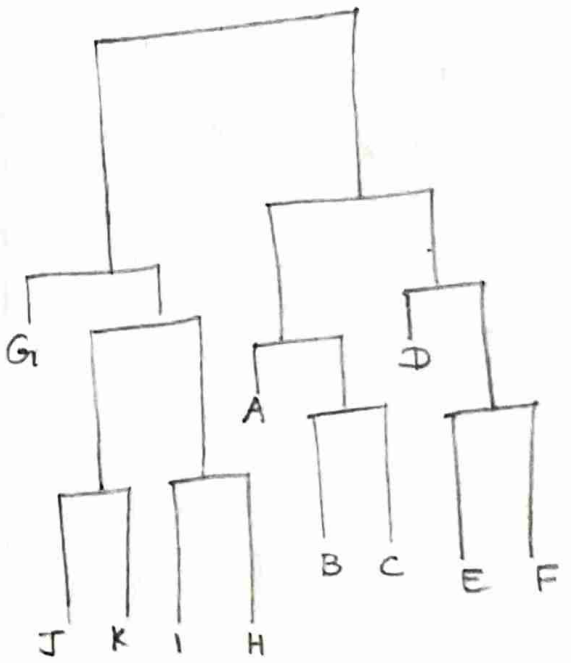
2. Hierarchical Clustering

It is a type of clustering technique that divides that data set into a number of clusters where the user dosen't specify the no. of clusters to be generated before training the model. This type of clustering technique is also known as connectivity based methods.
In this method, simple partitioning of the data set will not be done, whereas it provides us with

the hierarchy of the clusters that merge with each other after a certain distance.

After the hierarchical clustering is done on the dataset the result will be a tree based representation of data points [Dendogram], which are divided into clusters. This is how a hierarchical clustering looks like after training is done.





3. Fuzzy clustering:

Belongs to a branch of soft method clustering techniques, whereas all the above-mentioned clustering techniques belong to hard method clustering techniques. In this type of clustering technique points close to the centre may be a part of the other cluster to a higher degree than points at the edge of the same cluster. The probability of a point belonging to a given cluster is a value that lies between 0 & 1.

The most popular algorithm in this type of technique is FCM (Fuzzy-C-means Algorithm). Here the centroid of a cluster is calculated as the mean of all points, weighted by their probability of belonging to the cluster.

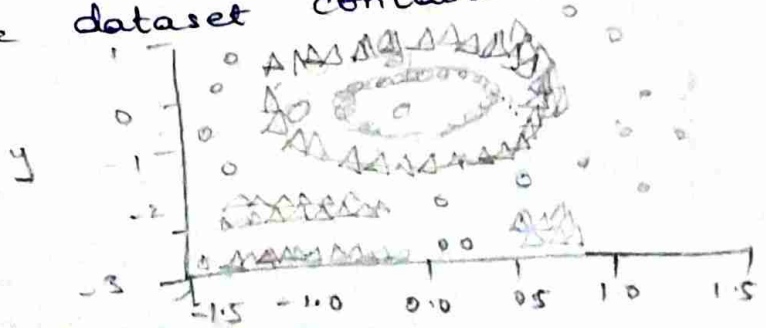
4. Density Based Clustering:

In this clustering, technique clusters will be formed by segregation of various density regions based on different densities in the data plot.

Density-Based Spatial clustering & Application with Noise (DBSCAN) is the most used algorithm in this type of technique. The main idea behind this algorithm is there should be a minimum no. of points that contain in the neighborhood of a given radius of each point in the cluster.

We can notice one common thing in all the techniques that are the shape of cluster formed are either spherical or oval or concave shaped.

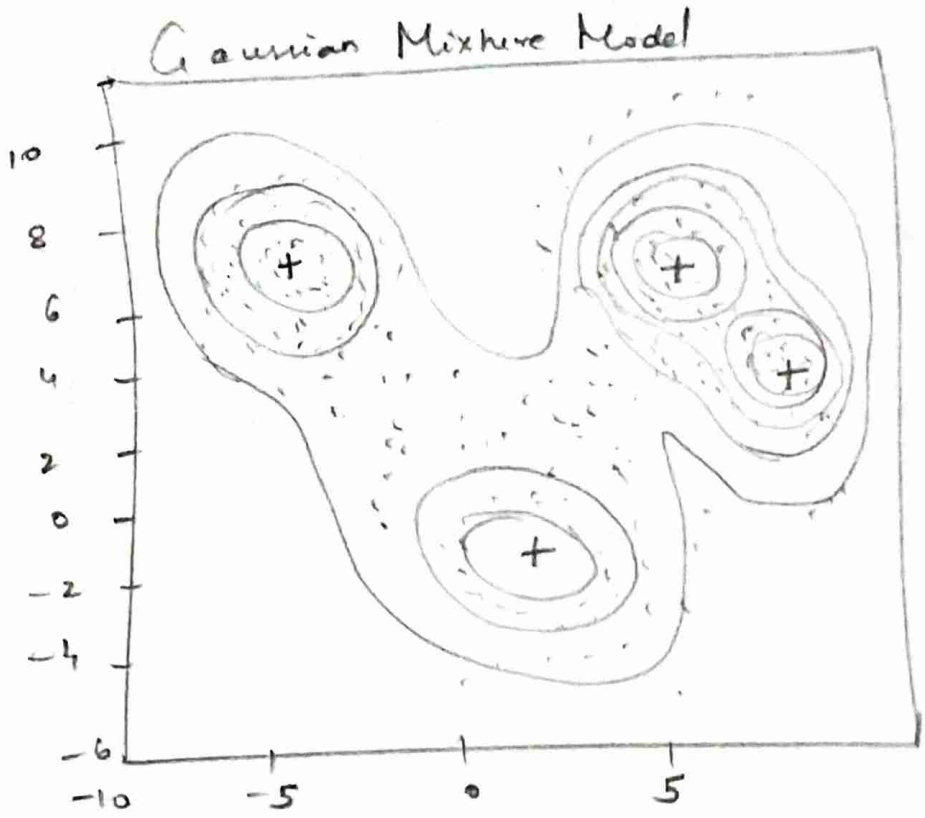
DBSCAN can form clusters in different shapes, this type of algorithm is most suitable when the dataset contains noise or outliers.



5. Distribution Model-Based Clustering:

In this type of clustering, technique clusters are formed by identifying by the probability of all data points in the cluster come from the same distribution (Normal, Gaussian). The most popular algorithm in this type of technique is Expectation-Maximization (EM) clustering using Gaussian Mixture models (GMM).

Normal clustering techniques like Hierarchical clustering and Partitioning clustering are not based on formal models, KNN in partitioning clustering yields different results with different k-values. As KNN and KMN consider mean for the cluster center it is not suitable in some cases with Gaussian mixture models we presume that data points are Gaussian distributed, this way we have two parameters to describe the shape of the clusters mean and the standard deviation. In this way for each cluster one Gaussian distribution is assigned, to get the optimum values of these parameters (mean & St. deviation) an optimization algorithm called Expectation Maximization is being used.



Correlation Clustering (data mining)

It relates to different task, where correlations among attributes of feature vectors in a high-dimensional space are assumed to exist guiding the clustering process. These correlations may be different in different clusters, thus a global decorrelation can not reduce this to traditional clustering.

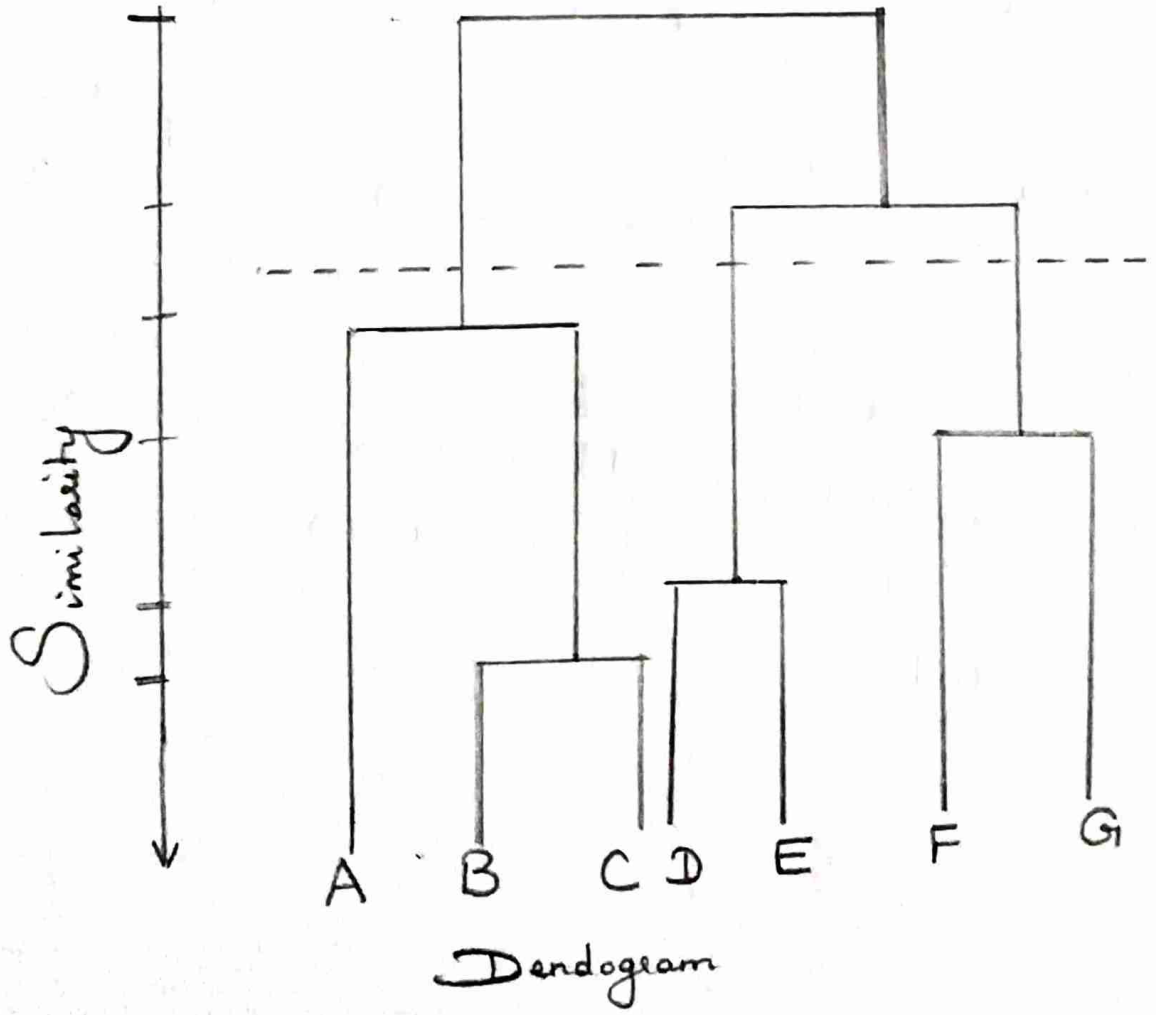
Correlations among subsets of attributes result in different spatial shapes of clusters. Hence, the similarity between cluster objects is defined by taking into account the local correlation patterns.

Hierarchical Clustering

Clustering is a data mining technique to group a set of objects in a way such that objects in the same cluster are more similar to each other than to those in other clusters.

In hierarchical clustering, we assign each object (data point) to a seperate cluster. Then compute the distance (similarity) between each of the clusters and join the two most similar clusters.

Let's understand further by solving an example.

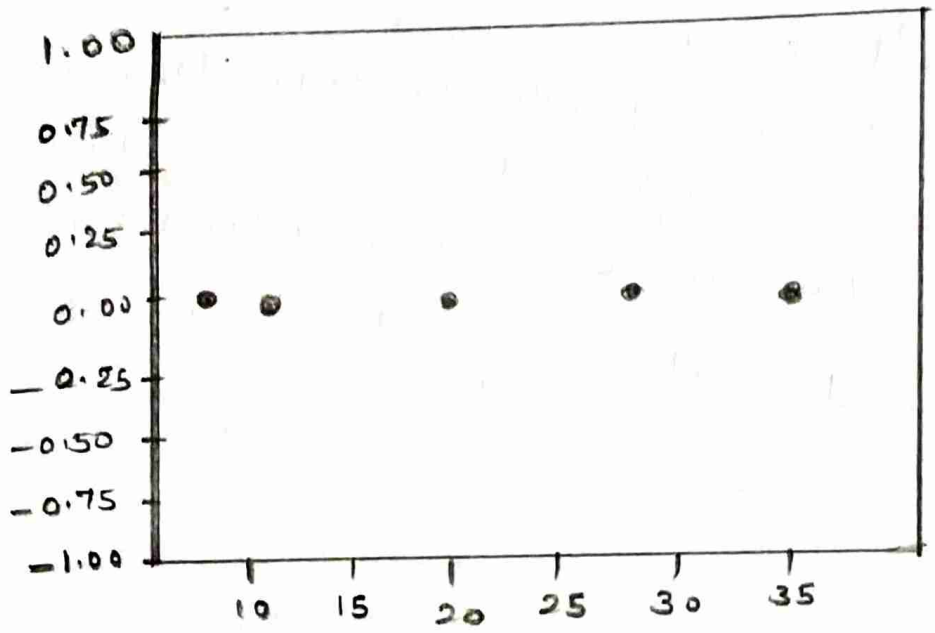


Dendogram

Objective:

For the one dimensional data set {7, 10, 20, 28, 35} perform hierarchical clustering & plot the dendogram to visualize it.

Solution: First let's visualize the data.



Observing the plot above, we can intuitively obelude that

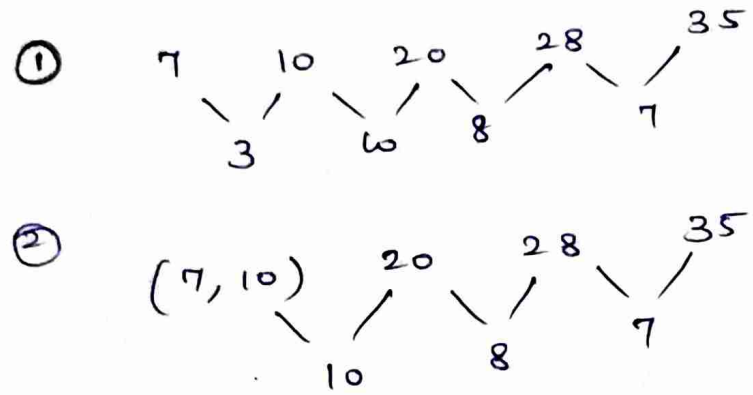
1. The first two points (7 and 10) are close to each other and should be in the same cluster.
2. Also the last two points (28 and 35) are close to each other and should be in the same cluster.
3. Cluster to the center point (20) is not easy to conclude.

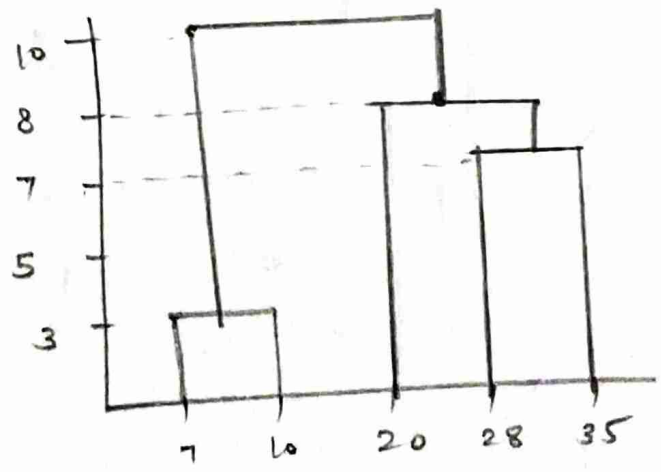
Let's solve the problem by hand using both the types of agglomerative hierarchical clustering:

1. Single Linkage:

Here we merge in each step the two clusters, whose two closest members have smallest distance.

Single Linkage





Dendogram

Using single linkage two clusters are formed:

Cluster 1: (7, 10)

Cluster 2: (20, 28, 35).

2. Complete linkage:

Here we merge in the members of the clusters in each step, which provide the smallest maximum pairwise distance.

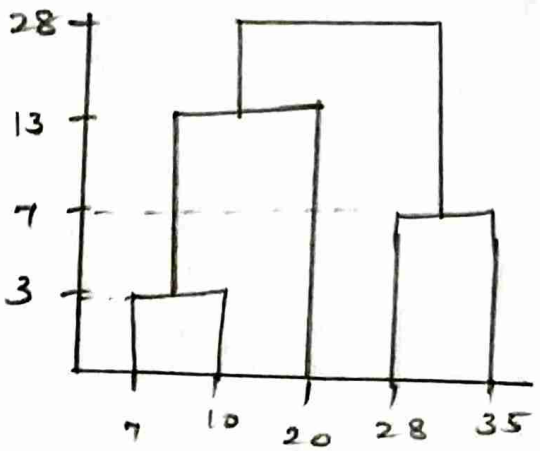
Complete linkage

1. 7 10 20 28 35
   3 10 8 7

2. (7,10) 20 28 35
   13 8 7

3. (7,10) 20 (28,35)
   13 15

4. (7,10,20) (28,35)
   28



Dendogram

Using Complete linkage two clusters are formed:

Cluster 1: (7, 10, 20)
Cluster 2: (28, 35)

Conclusion: Hierarchical clustering is mostly used when the application requires a hierarchy, e.g. creation of a taxonomy. However, they are expensive in terms of their computational & storage requirements.