MACHINE LEARNING ASSIGNMENT – 1

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:

Ans: a)2

2. In which of the following cases will K-Means clustering fail to give good results?

1. Data points with outliers
2. Data points with different densities
3. Data points with round shapes
4. Data points with non-convex shapes

Ans: d)1,2,4

3. The most important part of is selecting the variables on which clustering is based.

a) interpreting and profiling clusters
b) selecting a clustering procedure
c) assessing the validity of clustering
d) formulating the clustering problem

Ans: d)formulating the clustering problem

4. The most commonly used measure of similarity is the or its square.

a) Euclidean distance
b) city-block distance
c) Chebyshev’s distance
d) Manhattan distance

Ans: a)Euclidean distance

5. is a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters

a) Non-hierarchical clustering
b) Divisive clustering
c) Agglomerative clustering
d) K-means clustering

Ans: c) Agglomerative clustering

6. Which of the following is required by K-means clustering?

a) Defined distance metric
b) Number of clusters
c) Initial guess as to cluster centroids
d) All answers are correct

Ans: d) All answers are correct

7. The goal of clustering is to

a) Divide the data points into groups
b) Classify the data point into different classes
c) Predict the output values of input data points
d) All of the above

Ans: a) Divide the data points into groups

8. Clustering is a

a) Supervised learning
b) Unsupervised learning
c) Reinforcement learning
d) None

Ans: b) Unsupervised learning

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?

a) K- Means clustering
b) Hierarchical clustering
c) Diverse clustering
d) All of the above

Ans: d) All of the above

10. Which version of the clustering algorithm is most sensitive to outliers?

a) K-means clustering algorithm
b) K-modes clustering algorithm
c) K-medians clustering algorithm
d) None 11.

Ans: a) K-means clustering algorithm

11.Which of the following is a bad characteristic of a dataset for clustering analysis

a) Data points with outliers
b) Data points with different densities
c) Data points with non-convex shapes
d) All of the above

Ans: d) All of the above

12. For clustering, we do not require

a) Labeled data
b) Unlabeled data
c) Numerical data
d) Categorical data

Ans: a) Labeled data

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly.

13. How is cluster analysis calculated?

Ans: Cluster analysis is a data analysis technique that explores the naturally occurring groups within a data set known as clusters. Cluster analysis doesn't need to group data points into any predefined groups, which means that it is an unsupervised learning method. Clustering is a broad set of techniques for finding subgroups of observations within a data set. When we cluster observations, we want observations in the same group to be similar and observations in different groups to be dissimilar. Because there isn't a response variable, this is an unsupervised method, which implies that it seeks to find relationships between the n� observations without being trained by a response variable. Clustering allows us to identify which observations are alike, and potentially categorize them therein. K-means clustering is the simplest and the most commonly used clustering method for splitting a dataset into a set of k groups.

This tutorial serves as an introduction to the k-means clustering method.

1. **Replication Requirements:** What you'll need to reproduce the analysis in this tutorial
2. **Data Preparation:** Preparing our data for cluster analysis
3. **Clustering Distance Measures:** Understanding how to measure differences in observations
4. **K-Means Clustering**: Calculations and methods for creating K subgroups of the data

5. **Determining Optimal Clusters**: Identifying the right number of clusters to group your data.

14. How is cluster quality measured?

Ans: To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set.

We have a few methods to choose from for measuring the quality of a clustering. In general, these methods can be categorized into two groups according to whether ground truth is available. Here, *ground truth* is the ideal clustering that is often built using human experts.
If ground truth is available, it can be used by **extrinsic methods**, which compare the clustering against the group truth and measure. If the ground truth is unavailable, we can use **intrinsic methods**, which evaluate the goodness of a clustering by considering how well the clusters are separated. Ground truth can be considered as supervision in the form of “cluster labels.” Hence, extrinsic methods are also known as *supervised methods*, while intrinsic methods are *unsupervised methods*.

15. What is cluster analysis and its types?

Ans**: Types of Cluster Analysis**

The clustering algorithm needs to be chosen experimentally unless there is a mathematical reason to choose one cluster method over another.It should be noted that an algorithm that works on a particular set of data will not work on another set of data. There are a number of different methods to perform cluster analysis. Some of them are,

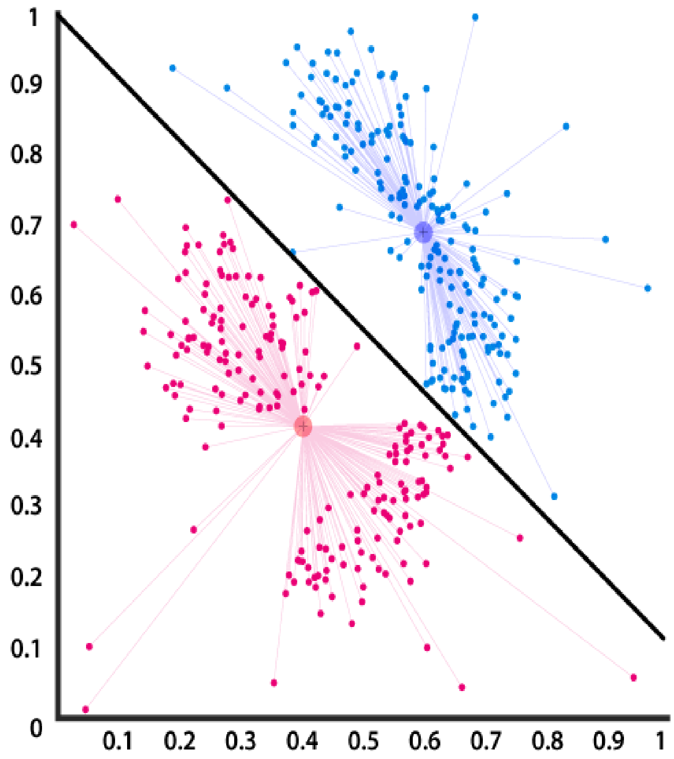
Hierarchical Cluster Analysis

In this method, first, a cluster is made and then added to another cluster (the most similar and closest one) to form one single cluster. This process is repeated until all subjects are in one cluster. This particular method is known as **Agglomerative method**. Agglomerative clustering starts with single objects and starts grouping them into clusters.

The divisive method is another kind of Hierarchical method in which clustering starts with the complete data set and then starts dividing into partitions.

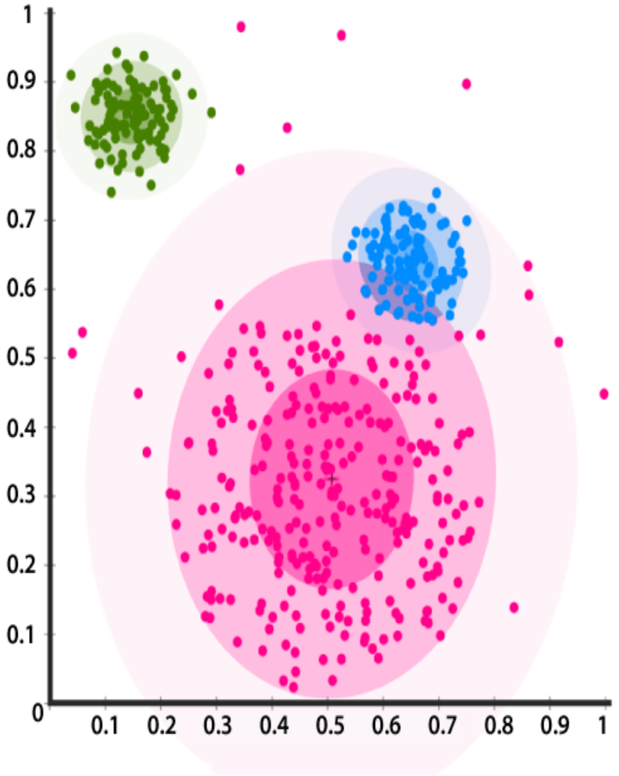
Centroid-based Clustering

In this type of clustering, clusters are represented by a central entity, which may or may not be a part of the given data set. K-Means method of clustering is used in this method, where k are the cluster centers and objects are assigned to the nearest cluster centres.



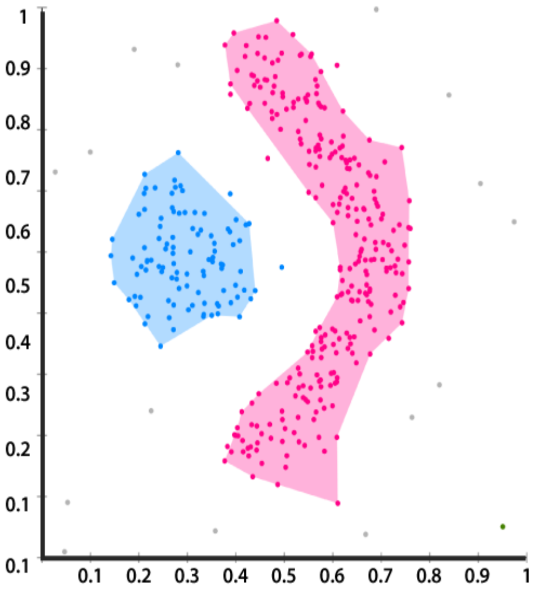
Distribution-based Clustering

It is a type of clustering model closely related to statistics based on the modals of distribution. Objects that belong to the same distribution are put into a single cluster.This type of clustering can capture some complex properties of objects like correlation and dependence between attributes.



Density-based Clustering

In this type of clustering, clusters are defined by the areas of density that are higher than the remaining of the data set. Objects in sparse areas are usually required to separate clusters.The objects in these sparse points are usually noise and border points in the graph.The most popular method in this type of clustering is DBSCAN.


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1