Machine Learning (Assignment one)

Q1 =What is the most appropriate no. of clusters for the data points represented by the following?

Ans = 4

Q2 =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 = 1, 2 and 4

Q3 =The most important part is selecting the variables on which clustering is based.

Ans = d) formulating the clustering problem

Q4 = The most commonly used measure of similarity is the or it's square.

Ans = a) Euclidean distance

Q5 ______ 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.

Ans = b) Divisive clustering

Q6 . Which of the following is required by K-means clustering?

Ans d)) All answers are correct

Q7 The goal of clustering is to_

Ans d) All of the above

Q8 . Clustering is a_

Ans b) Unsupervised learning

Q9 . Which of the following clustering algorithms suffers from the problem of convergence at local optima?

Ans a) K- Means clustering

Q10 . Which version of the clustering algorithm is most sensitive to outliers?

Ans a) K-means clustering algorithm

Q11 Which of the following is a bad characteristic of a dataset for clustering analysis

Ans d) All of the above

Q12. For clustering, we do not require_

Ans a) Labeled data

Q13. How is cluster analysis calculated?

Ans SPSS(Statistical Package for Social Sciences)offers three methods for the cluster analysis: K-Means Cluster, Hierarchical Cluster, and Two-Step Cluster.

- K-means cluster is a method to quickly cluster large data sets. The researcher define the number of clusters in advance. This is useful to test different models with a different assumed number of clusters.
- Hierarchical cluster is the most common method. It generates a series of models with cluster solutions from 1 (all cases in one cluster) to n (each case is an individual cluster). Hierarchical cluster also works with variables as opposed to cases; it can cluster variables together in a manner somewhat similar to factor analysis. In addition, hierarchical cluster analysis can handle nominal, ordinal, and scale data; however it is not recommended to mix different levels of measurement.
- Two-step cluster analysis identifies groupings by running pre-clustering first and then by running hierarchical methods. Because it uses a quick cluster algorithm upfront, it can handle large data sets that would take a long time to compute with hierarchical cluster methods. In this respect, it is a combination of the previous two approaches. Two-step clustering can handle scale and ordinal data in the same model, and it automatically selects the number of clusters.

Q14. How is cluster quality measured?

Ans Intrinsic Methods

When the ground truth of a data set is not available, we have to use an intrinsic method to assess the clustering quality. In general, intrinsic methods evaluate a clustering by examining how well the clusters are separated and how compact the clusters are. Many intrinsic methods have the advantage of a similarity metric between objects in the data set.

The silhouette coefficient is such a measure. For a data set, D, of n objects, suppose D is partitioned into k clusters, C1, …, Ck. For each object o ∈ D, we calculate a (i)as the average distance between o and all other objects in the cluster to which o belongs. Similarly, b(i)is the minimum average distance from o to all clusters to which o does not belong. Formally, suppose (i) ∈Ci (1 ≤ i ≤ k); then

.

$$a(i) = \frac{1}{n_c - 1} \sum_{i,j \in C_c, i \neq j} d(i,j) \qquad (1)$$

$$b(i) = \min_{p, p \neq c}\left[\frac{1}{n_p} \sum_{i \in C_c, j \in C_p} d(i,j)\right] \qquad (2)$$

$$s_i = \frac{b(i) - a(i)}{\max[a(i), b(i)]} \qquad (3)$$

The silhouette coefficient of (i) is then defined as

The value of the silhouette coefficient is between −1 and 1. The value of a(i) reflects the compactness of the cluster to which o belongs. The smaller the value, the more compact the cluster. The value of b(i) captures the degree to which o is separated from other clusters. The larger b(i) is, the more separated o is from other clusters. Therefore, when the silhouette coefficient value of o approaches 1, the cluster containing o is compact and (i) is far away from other clusters, which is the preferable case. However, when the silhouette coefficient value is negative (i.e., b(i) < a(i)), this means that, in expectation, (i) is closer to the objects in another cluster than to the objects in the same cluster as (i). In many cases, this is a bad situation and should be avoided.

To measure a cluster's fitness within a clustering, we can compute the average silhouette coefficient value of all objects in the cluster. To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set. The silhouette coefficient and other intrinsic measures can also be used in the elbow method to heuristically derive the number of clusters in a data set by replacing the sum of within-cluster variances.

Q15. What is cluster analysis and its types?

Ans Cluster analysis is a multivariate data mining technique whose goal is to groups objects (eg., products, respondents, or other entities) based on a set of user-selectedtics or attributes. It is the basic and most important step of data mining and a common technique for statistical data analysis, and it is used in many fields such as data compression, machine learning, pattern recognition, information retrieval etc

- 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,

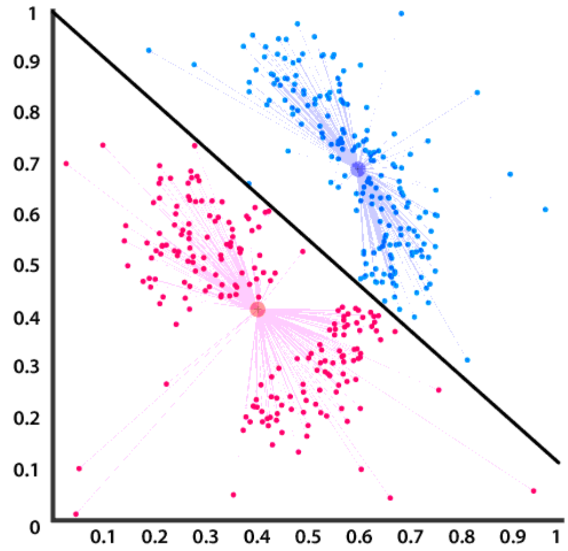
- Hierarchical Cluster Analysis

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.

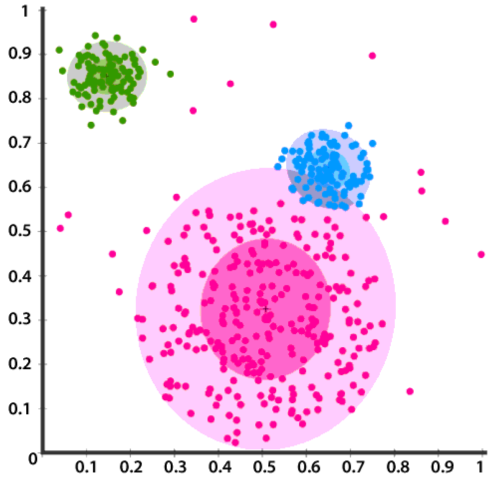
- Centroid-based Clustering

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.



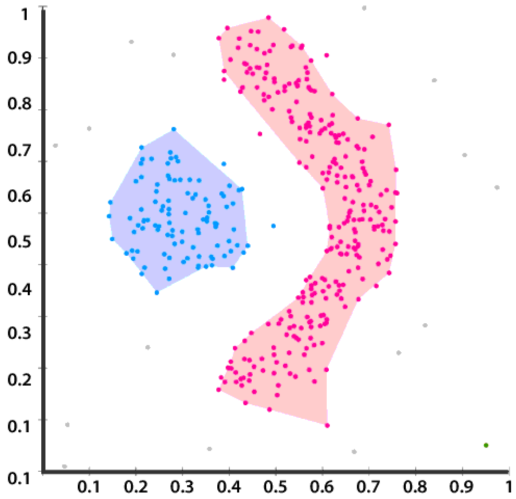
- Distribution-based Clustering

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