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-:b) 4

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.

Ans-:d)) formulating the clustering problem

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

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.

Ans-:b) Divisive clustering

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

Ans-:d) All answers are correct

7. The goal of clustering is to-

Ans-:a) Divide the data points into groups

8. Clustering is a-

Ans-:b) Unsupervised learning

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

Ans-:a) K- Means clustering

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

Ans-:a) K-means clustering algorithm

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

Ans-:d) All of the above

12. For clustering, we do not require-

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 an exploratory analysis that tries to identify structures within the data. Cluster analysis is also called segmentation analysis or taxonomy analysis. More specifically, it tries to identify homogenous groups of cases if the grouping is not previously known. Because it is exploratory, it does not make any distinction between dependent and independent variables. The different cluster analysis methods that SPSS offers can handle binary, nominal, ordinal, and scale data.
The researcher must be able to interpret the cluster analysis based on their understanding of the data to determine if the results produced by the analysis are actually meaningful.

Calculations-:By varying k from 1 to 10 clusters. For each k, calculate the total within-cluster sum of square. Plot the curve of was according to the number of clusters k. The location of a bend in the plot is generally considered as an indicator of the appropriate number of clusters.

14. How is cluster quality measured?

Ans-: Usually clustering algorithms are non-supervised which often means that the correct partition into classes is not know and the number of classes is also unknown. In these cases there are measures of cluster validity based in maximizing inter-cluster distances and minimizing intra-cluster distances. There is a vast scientific literature on cluster validity. If in your case you know the correct partition then you can use a distance measure between partitions to measure how similar to the correct clustering is the clustering obtained by the algorithm. The choice of metric rather depends on what you consider the purpose of clustering to be. Personally I think clustering ought to be about identifying different groups of observations that were each generated by a different data generating process. So I would test the quality of a clustering by generating data from known data generating processes and then calculate how often patterns are misclassified by the clustering. Of course this involved making assumptions about the distribution of patterns from each generating process, but you can use datasets designed for supervised classification. Here you have a couple of measures, but there are many more:

SSE-: Sum of the square error from the items of each cluster. Inter cluster distance: sum of the square distance between each cluster centroid. Intra cluster distance for each cluster: sum of the square distance from the items of each cluster to its centroid.

Maximum Radius-: Largest distance from an instance to its cluster centroid. Average Radius: sum of the largest distance from an instance to its cluster centroid divided by the number of clusters.

To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set.In depth the calculation method used they are,

1.There are few well known measures like silhouette width (SW)-: **silhouette width** of the observation i is defined by the **formula**: Si=(bi−ai)/max(ai,bi). **Silhouette** Coefficient = (x-y)/ max(x,y)
where, y is the mean intra cluster distance: mean distance to the other instances in the same cluster. x depicts mean nearest cluster distance i.e. mean distance to the instances of the next closest cluster. The coefficient varies between -1 and 1.

2. The Davies- Bouldin index (DB)-: Davies–Bouldin index , introduced by David L. Davies and Donald W. Bouldin in 1979, is a metric for evaluating clustering algorithms. This is an internal evaluation scheme, where the validation of how well the clustering has been done is made using quantities and features inherent to the dataset

3.The Calinski-Harabasz index (CH)-: The Calinski-Harabasz index also known as the Variance Ratio Criterion, is the ratio of the sum of between-clusters dispersion and of inter-**cluster** dispersion for all clusters, the higher the score , the better the performances.

4.and, The Dunn index -: The Dunn index is a metric for evaluating clustering algorithms. This is part of a group of validity indices including the Davies–Bouldin index or Silhouette index, in that it is an internal evaluation scheme, where the result is based on the clustered data itself.

15. What is cluster analysis and its types?

Ans-: Clustering is a type of unsupervised learning method of machine learning. In the unsupervised learning method, the inferences are drawn from the data sets which do not contain labelled output variable. It is an exploratory data analysis technique that allows us to analyse the multivariate data sets. Clustering is a task of dividing the data sets into a certain number of clusters in such a manner that the data points belonging to a cluster have similar characteristics. Clusters are nothing but the grouping of data points such that the distance between the data points within the clusters is minimal. In other words, the clusters are regions where the density of similar data points is high. It is generally used for the analysis of the data set, to find insightful data among huge data sets and draw inferences from it. Clustering itself can be categorized into two types i.e. Hard Clustering and Soft Clustering. In hard clustering, one data point can belong to one cluster only. But in soft clustering, the output provided is a probability likelihood of a data point belonging to each of the pre-defined numbers of clusters. Various types of Clustering Methods are as follows:

1. Connectivity-based Clustering (Hierarchical clustering)

2. Centroids-based Clustering (Partitioning methods)

3. Distribution-based Clustering

4. Density-based Clustering (Model-based methods)

5. Fuzzy Clustering

6. Constraint-based (Supervised Clustering)

#Hierarchical Clustering -:Hierarchical Clustering groups (Agglomerative or also called as Bottom-Up Approach) or divides (Divisive or also called as Top-Down Approach) the clusters based on the distance metrics. In Agglomerative clustering, each data point acts as a cluster initially, and then it groups the clusters one by one. Divisive is the opposite of Agglomerative, it starts off with all the points into one cluster and divides them to create more clusters. These algorithms create a distance matrix of all the existing clusters and perform the linkage between the clusters depending on the criteria of the linkage. The clustering of the data points is represented by using a dendrogram.

 # Partitioning Clustering -:This method is one of the most popular choices for analysts to create clusters. In partitioning clustering, the clusters are partitioned based upon the characteristics of the data points. We need to specify the number of clusters to be created for this clustering method. These clustering algorithms follow an iterative process to reassign the data points between clusters based upon the distance.

#Distribution-based Clustering -: It is closely related to statistics as it very closely relates to the way how datasets are generated and arranged using random sampling principles i.e., to fetch data points from one form of distribution. Clusters can then be easily be defined as objects that are most likely to belong to the same distribution. A major drawback of density and boundary-based approaches is in specifying the clusters apriori to some of the algorithms and mostly the definition of the shape of the clusters for most of the algorithms. There is at least one tuning or hyper-parameter which needs to be selected and not only that is trivial but also any inconsistency in that would lead to unwanted results. Distribution based clustering has a vivid advantage over the proximity and centroid based clustering methods in terms of flexibility, correctness and shape of the clusters formed. The major problem however is that these clustering methods work well only with synthetic or simulated data or with data where most of the data points most certainly belong to a predefined distribution, if not, the results will overfit.

 # Density-Based Clustering-: In this method, the clusters are created based upon the density of the data points which are represented in the data space. The regions that become dense due to the huge number of data points residing in that region are considered as clusters. The data points in the sparse region (the region where the data points are very less) are considered as noise or outliers. The clusters created in these methods can be of arbitrary shape.

 # Fuzzy Clustering -:In fuzzy clustering, the assignment of the data points in any of the clusters is not decisive. Here, one data point can belong to more than one cluster. It provides the outcome as the probability of the data point belonging to each of the clusters. One of the algorithms used in fuzzy clustering is Fuzzy c-means clustering.

#Constraint-Based Clustering -: The clustering process, in general, is based on the approach that the data can be divided into an optimal number of “Unknown” groups. The underlying stages of all the clustering algorithms to find those hidden patterns and similarities, without any intervention or predefined conditions. However, in certain business

scenarios, we might be required to partition the data based on certain constraints. Here is where a supervised version of clustering machine learning techniques come into play. A constraint is defined as the desired properties of the clustering results, or a user's expectation on the clusters so formed – this can be in terms of a fixed number of clusters, or, the cluster size, or, important dimensions (variables) that are required for the clustering process