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:

a) 2 b) 4 c) 6 d) 8

ANS:- **(d)**

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

Options: a) 1 and 2 b) 2 and 3 c) 2 and 4 d) 1, 2 and 4

ANS:- **(d)**

3. The most important part of **interpreting and profiling clusters** 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

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

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

MACHINE LEARNING ASSIGNMENT – 1

5. **Agglomerative clustering** 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

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)**

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:- **(d)**

8. Clustering is a **Unsupervised learning**

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

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)**

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

ANS:- **(a)**

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)**

12. For clustering, we do not require: **Labeled Data**

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

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

13. How is cluster analysis calculated?

ANS:- **Clustering is main method used in Unspervised learning techniques for analysis of statistical data.Cluster analysis is calculated on raw data where row is the object and column is a quantity of objects. They are further then refered to as Clustering Variables. Some methods used in machine learning area: k-means clustering , density methods, grid-based methods, heirarchial bases method etc.**

14. How is cluster quality measured?

ANS:- **Some of the methods used for measuring good quality clusters:**

1) **Dissimilarity/Similarity metric: the similarity between clusters is expressed in terms of a distance function represented by 'd'(i,j) . Distance function measures are expressed in Euclidean distance, Mahalanobis and Cosine Distance for different types of data.**
2) **Cluster Completeness: If any two data are similar in nature then they are clubbed together into one good cluster under same category.**
3) **RagBag Category: If under any circumstance it occurs that a data doesn't fit into a particular type of cluster then it is entered into separate RAGBAG cluster.**
4) **Small Cluster Preservation: Sometimes it happens that unser one big cluster there occur many small sections of cluster, therefore to prevent this occurences of lot of different clusters this method is used.**

15. What is cluster analysis and its types?

ANS:- **Cluster Analysis is when we group data of similar type into one cluster. It consists of data having similar characteristics. Its application is mainly in fields of data compression, machine learning, pattern recognition, information retrieval etc. For example: if we want to know the number of students enrolled into one course 'Machine Learning' then we can find all the students enrolled for that course under the main cluster 'Machine Learning'. We don't have to search for each and every student file.**

Types of Cluster Analysis:

1) **HEIRARCHIAL CLUSTER ANALYSIS**
   A cluster is made and added to one main cluster, this step is repeated until all the similar clusters are added to one big cluster. This method is also known as Agglomerative Method.
2) **CENTROID BASED CLUSTERING**
   Clusters in this type are represented by a single central entity and the one which is the nearest to this measure of central entity is then selected. K-Means method is also used for representation of this method.
3) **DISTRIBUTION BASED CLUSTERING**
   As the name suggests 'Distribution', this method is closely related to statistics based modals of distribution. This type of clustering requires some properties like correlation and dependence between attributes.
4) **DENSITY BASED CLUSTERING**
   This type of clustering is represented by the high density areas, than the remaining data set. This type is usually used in Population ratio, Elections etc.