ASSIGNMENT -1

MACHINE LEARNING

1. The most appropriate no. of clusters for the data pointes represented by the following dendrogram is
   c) 6
2. K-Means clustering will fail to give good results when
   - Data points with outliers
   - Data points with different densities
   - Data points with non-convex shapes

   d) 1,2 and 4
3. The most important part of _____ is selecting the variables on which clustering is based
   d) Formulating the clustering problems
4. The most commonly used measure of similarity is the ____ or its square
   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
   b) Divisive clustering
6. K-means clustering require
   - Defined distance metric
   - Number of clusters
   - Initial guess as to cluster centroids

   d)All of above
7. The goal of clustering is to
   - Divide data points into groups
   - Classify the data point into different groups
   - Predict the output values of input data points

   c) All of the above
8. Clustering is
   b) unsupervised learning
9. a) K-means clustering algorithm suffers from the problem of convergence at local optima
10. K-means clustering algorithm is most sensitive to outliers.
11. Bad characteristics of a dataset for clustering analysis are
    - Data points with outliers
    - Data points with different densities
    - Data points with non-convex shapes

    d) All of the above
12. For clustering we do not require
    a) Labeled Data

13. How is cluster analysis calculated?

Answer:-

Cluster analysis is a multivariate data mining technique whose goal is to group objects(eg.products, respondents, or other entities) based on a set of user selected characteristics or attributes. It is the basic and most important step of data mining and common technique for statistical data analysis.

Steps to calculate clusters:

- Prepare Data :- In this step we have to normalize scale and transform feature data and additionally calculate the similarity between the data.
- Create similarity matric :- Here we have to create similarity for the algorithm to know how similar pairs of examples are.
- Run clustering algorithm :- Here similarity metric is used to cluster data. Here different types of clustering algorithm is used.
- Interpret Result and Adjust :- Here the output is verified against the expectations. This result can be improved by experimenting with the previous steps.

14. How is cluster quality measured?

    Answer:-

    Cluster quality can be measured by

    - Dissimilarity/Similarity metric: Similarity is expressed in terms of a distance function.
    - Cluster completeness: Cluster completeness is the essential parameter for good clustering, if any two data objects are having similar characteristics then they are assigned to the same category of the cluster according to ground truth. Cluster completeness is high if the objects are of the same category.
    - Ragbag: In some situations, there can be a few categories in which the objects of those categories cannot be merged with other objects. Then the quality of those cluster categories is measured by the Rag Bag method. According to the rag bag method, we should put the heterogeneous object into a rag bag category.
    - Small cluster preservation: If a small category of clustering is further split into small pieces, then those small pieces of cluster become noise to the entire clustering and thus it becomes difficult to identify that small category from the clustering. The small cluster preservation criterion states that are splitting a small category into pieces is not advisable and it further decreases the quality of clusters as the pieces of clusters are distinctive.

15. What is cluster analysis and its types?

    Answer:-

    - A collection of data objects is known as clusters. They are similar to one another within the same cluster and dissimilar to the objects in other clusters
    - Cluster analysis means grouping a set of data objects into clusters
    - Clustering is unsupervised classification that is no predefined classes.

Types of clusters are :-

- Centroid-based clustering organizes the data into non-hierarchical clusters.
- Density-based clustering connects areas of high example density into clusters. This allows for arbitrary-shaped distributions as long as dense areas can be connected. These algorithms have difficulty with data of varying densities and high dimensions.
- Distribution based Clustering assumes data is composed of distributions, such as Gaussian distributions.
- Hierarchical clustering creates a tree of clusters. Hierarchical clustering, not surprisingly, is well suited to hierarchical data, such as taxonomies.