MACHINE LEARNING ASSIGNMENT

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) 4

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

Ans. a) 1, 2 and 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 of the above

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. d) All of the above

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

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 calculated follows three basic steps:

1) calculate the distances,

2) link the clusters

3) choose a solution by selecting the right number of clusters.

14. How is cluster quality measured?

Ans. Clustering Evaluation: Evaluating the goodness of clustering results

No commonly recognized best suitable measure in practice

Three categorization of measures: External, internal, and relative

External: Supervised, employ criteria not inherent to the dataset

Compare a clustering against prior or expert-specified knowledge (i.e., the ground truth) using certain clustering quality measure

Internal: Unsupervised, criteria derived from data itself

Evaluate the goodness of a clustering by considering how well the clusters are separated and how compact the clusters are, e.g., silhouette coefficient

Relative: Directly compare different clusterings, usually those obtained via different parameter

settings for the same algorithm

15. **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 selected characteristics 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.

Type of cluster analysis

- Hierarchical Cluster Analysis
- Centroid-base Clustering
- Disribution-base Clustering
- Density-base Clustering