MACHINE LEARNING ASSIGNMENT – 3

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following is an application of clustering?

d. All of the above

2. On which data type, we cannot perform cluster analysis?

d. None

3. Netflix's movie recommendation system uses-

a. Supervised learning

4. The final output of Hierarchical clustering is-

b. The tree representing how close the data points are to each other

5. Which of the step is not required for K-means clustering?

d. None

6. Which is the following is wrong?

c. k-nearest neighbour is same as k-means

7. Which of the following metrics, do we have for finding dissimilarity between two clusters in

hierarchical clustering?   i. Single-link ii.   Complete-link   iii. Average-link

d. 1, 2 and 3

8. Which of the following are true?

i. Clustering analysis is negatively affected by multicollinearity of features

ii. Clustering analysis is negatively affected by heteroscedasticity

a. 1 only

9. In the figure above, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?

a. 2

10. For which of the following tasks might clustering be a suitable approach?

b. Given a database of information about your users, automatically group them into different market segments.

11. Given, six points with the following attributes:

Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:

a.

12. Given, six points with the following attributes:

Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering?

b.

Q13 to Q14 are subjective answers type questions, Answers them in their own words briefly

13. What is the importance of clustering?

Ans: Clustering helps in understanding the natural grouping in a dataset. Their purpose is to make sense to partition the data into some group of logical groupings. Clustering quality depends on the methods and the identification of hidden patterns.

14. How can I improve my clustering performance?

Ans:

1. Graph-based clustering performance can easily be improved by applying ICA blind source separation during the graph Laplacian embedding step.
2. Applying unsupervised feature learning to input data using either RICA or SFT, improves clustering performance.
3. Surprisingly for some cases, high clustering performance can be achieved by simply performing K-means clustering on the ICA components after PCA dimension reduction on the input data. However, the number of PCA and ICA signals/components needs to be limited to the number of unique classes.