MACHINE LEARNING ASSIGNMENT 3

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

a. Biological network analysis

b. Market trend prediction

c. Topic modeling

d. All of the above

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

a. Time series data

b. Text data

c. Multimedia data

d. None

3. Netflix's movie recommendation system uses

a. Supervised learning

b. Unsupervised learning

c. Reinforcement learning and Unsupervised learning

d. All of the above

4. The final output of Hierarchical clustering is

a. The number of cluster centroids

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

c. A map defining the similar data points into individual groups

d. All of the above

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

a. A distance metric

b. Initial number of clusters

c. Initial guess as to cluster centroids

d. None

6. Which is the following is wrong?

a. k-means clustering is a vector quantization method

b. k-means clustering tries to group n observations into k clusters

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

d. None

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

Options:

a.1 and 2

b. 1 and 3

c. 2 and 3

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

Options:

a. 1 only

b. 2 only

c. 1 and 2

d. None of them

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

b. 4

c. 3

d. 5

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

a.Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

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

c. Predicting whether stock price of a company will increase tomorrow.

d. Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

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:

Answer : 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.

Answer : A

13. What is the importance of clustering?

Answer:

Clustering is used to find structure in unlabelled data. This is the most common form of unsupervised learning. Given a dataset you don't know anything about, a clustering algorithm can discover groups of objects where the average distances between the members of each cluster are closer than to members in other clusters.

- Increased resource availability: If one Intelligence Server in a cluster fails, the other Intelligence Servers in the cluster can pick up the workload. This prevents the loss of valuable time and information if a server fails.
- Strategic resource usage: You can distribute projects across nodes in whatever configuration you prefer. This reduces overhead because not all machines need to be running all projects, and allows you to use your resources flexibly.
- Increased performance: Multiple machines provide greater processing power.
- Greater scalability: As your user base grows and report complexity increases, your resources can grow.
- Simplified management: Clustering simplifies the management of large or rapidly growing systems.

14. How can I improve my clustering performance?

Answer:

- Graph-based clustering performance can easily be improved by applying ICA blind source separation during the graph Laplacian embedding step.
- Applying unsupervised feature learning to input data using either RICA or SFT, improves clustering performance.
- 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.