

MACHINE LEARNING (With Answers)

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?

- a. Biological network analysis
- b. Market trend prediction
- c. Topic modeling
- d. All of the above

Ans: 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

Ans: 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

Ans: c. Reinforcement learning and Unsupervised learning

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

Ans: b. The tree represents how close the data points are to each other

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

- a. A distance metric
- b. An initial number of clusters
- c. Initial guess as to cluster centroids
- d. None

Ans: d. None

6. Which of 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 neighbor is same as k-means
- d. None

Ans: c. k-nearest neighbor is the 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 Options:
 - a. 1 and 2
 - b. 1 and 3
 - c. 2 and 3
 - d. 1, 2 and 3

Ans: 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

Ans: 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 b. 4 c. 3 d. 5

Ans: a. 2

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 the stock price of a company will increase tomorrow.
- d. Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

Ans: 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:

Ans: 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.

Ans: 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 is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters).

Importance of clustering:

- Having clustering methods helps in restarting the local search procedure and removing the inefficiency. In addition, clustering helps to determine the internal structure of the data.
- This clustering analysis has been used for model analysis, and vector region of attraction.
- 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.
- They play a wide role in applications like marketing economic research and weblogs to identify similarity measures, Image processing, and spatial research.
- They are used in outlier detections to detect credit card fraudulence.

14. How can I improve my clustering performance?

Ans: Clustering is an unsupervised machine learning methodology that aims to partition data into distinct groups, or clusters. There are a few different forms including hierarchical, density, and similarity based. Each have a few different algorithms associated with it as well.

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 reconstruction cost (RICA) and sparse filtering (SFT), improves clustering performance.

K-means is the most popular and partition based clustering algorithm. But it is computationally expensive and the quality of resulting clusters heavily depends on the selection of initial centroid and the dimension of the data. Several methods have been proposed in the literature for improving performance of the k-means clustering algorithm. Principal Component Analysis (PCA) is an important approach to unsupervised dimensionality reduction technique.

THE END