MACHINE LEARNING

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

2. Ans = d. All of the above

3. On which data type, we cannot perform cluster analysis? a. Time series data b. Text data c. Multimedia data d. None

Ans = d. None

4. Netflix's movie recommendation system usesa. 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 isa. 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 representing how close the data points are to each other

Q5 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

Ans = 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

Ans = 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 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?

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 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:

Ans = a

12. Given, six points with the following attributes:

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 a Unsupervised learning is an important concept in machine learning. It saves data analysts' time by providing algorithms that enhance the grouping and investigation of data. It's also important in well-defined network models. Many analysts prefer using unsupervised learning in network traffic analysis because of frequent data changes and scarcity of labels. it's needed when creating better forecasting, especially in the area of threat detection. This can be achieved by developing network logs that enhance threat visibility. This category of machine learning is also resourceful in the reduction of data dimensionality. We need dimensionality reduction in datasets that have many features. Unsupervised learning can analyze complex data to establish less relevant features. The model can then be simplified by dropping these features with insignificant effects on valuable insights. Clustering is the process of dividing uncategorized data into similar groups or clusters. This process ensures that similar data points are identified and grouped. Clustering algorithms is key in the processing of data and identification of groups (natural clusters).

1) Through the use of clusters, attributes of unique entities can be profiled easier. This can subsequently enable users to sort data and analyze specific groups.
2) Clustering enables businesses to approach customer segments differently based on their attributes and similarities. This helps in maximizing profits.

3) It can help in dimensionality reduction if the dataset is comprised of too many variables. Irrelevant clusters can be identified easier and removed from the dataset.

14. How can I improve my clustering performance?

Ans = k-means is a very simple and ubiquitous clustering algorithm. But quite often it does not work on your problem, for example because the initialization is bad. I ran into a similar problem recently, where I applied k-means to a smaller number of files in my data sets and everything worked fine, but when I ran it on many more samples it just wasn't reliably getting good results. As we have seen in my previous post, vanilla k-means suffers from its random initialization. Depending on which points get chosen as the start centers the solution can be a very bad local minimum. And since k-means has a strictly concave loss function, it has no way of escaping from this local minimum during training.