Machine Learning Assignment-3

Q.1. Answer- D) All of the above

Q.2. Answer- D) None

Q.3. Answer- C) Reinforcement learning and Unsupervised learning

Q.4. Answer- B) The tree representing how close the data points are to each other

Q.5. Answer- D) None

Q.6. Answer- C) k-nearest neighbour is same as k-means

Q.7. Answer- D) 1, 2 and 3

Q.8. Answer- A) 1 only

Q.9. Answer- A) 2

Q.10. Answer- B) Given a database of information about your users, automatically group them into different market segments.

Q.11. Answer- A)

Q.12. Answer- B)

Q.13. Answer- 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.

Clustering is important in data analysis and data mining applications. It 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.

It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

A good clustering algorithm is able to identity clusters irrespective of their shapes. The stages involved in clustering algorithm are:

Raw Data-> Clustering Algorithms-> Clusters of Data

Q.14. Answer- The most common ways of measuring the performance of clustering models are to either measure the distinctiveness or the similarity between the created groups. Given this, there are three common metrics to use, these are:

1. **Silhouette Score:** It is calculated using the mean intra-cluster distance and the mean nearest-cluster distance. This score is between -1 and 1, where the higher the score the more well-defined and distinct clusters are.

2. **Calinski-Harabaz Index:** It is calculated using the between-cluster dispersion and within-cluster dispersion in order to measure the distinctiveness between groups. This score has no bound, meaning that there is no ‘acceptable’ or ‘good’ value.

3. **Davies-Bouldin Index**: It is the average similarity of each cluster with its most similar cluster. This score measures the similarity of your clusters, meaning that the lower the score the better separation there is between your clusters.

So, to improve the performance of clustering methods, we need to use metrics which have an upper and lower bound. The most commonly used metric for measuring the performance of a clustering algorithm is the Silhouette Score and it can therefore be used for comparison as it's bounded between -1 and 1.