MACHINE LEARNING

Q1 to Q11 have only one correct answer. Choose the correct option to answer your question.

1. Movie Recommendation systems are an example of:
   i) Classification
   ii) Clustering
   iii) Regression
   Options:
   a) 2 Only
   b) 1 and 2
   c) 1 and 3
   d) 2 and 3

2. Sentiment Analysis is an example of:
   i) Regression
   ii) Classification
   iii) Clustering
   iv) Reinforcement
   Options:
   a) 1 Only
   b) 1 and 2
   c) 1 and 3
   d) 1, 2 and 4

3. Can decision trees be used for performing clustering?
   a) True
   b) False

4. Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points:
   i) Capping and flooring of variables
   ii) Removal of outliers
   Options:
   a) 1 only
   b) 2 only
   c) 1 and 2
   d) None of the above

5. What is the minimum no. of variables/ features required to perform clustering?
   a) 0
   b) 1
   c) 2
   d) 3

6. For two runs of K-Mean clustering is it expected to get same clustering results?
   a) Yes
   b) No

7. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means?
   a) Yes
   b) No
   c) Can't say
   d) None of these

MACHINE LEARNING

8. Which of the following can act as possible termination conditions in K-Means?
   i) For a fixed number of iterations.
   ii) Assignment of observations to clusters does not change between iterations. Except for cases witha bad local minimum.
   iii) Centroids do not change between successive iterations.
   iv) Terminate when RSS falls below a threshold.
   Options:
   a) 1, 3 and 4
   b) 1, 2 and 3
   c) 1, 2 and 4
   d) All of the above

9. Which of the following algorithms is most sensitive to outliers?
   a) K-means clustering algorithm
   b) K-medians clustering algorithm
   c) K-modes clustering algorithm
   d) K-medoids clustering algorithm

10. How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):
    i) Creating different models for different cluster groups.
    ii) Creating an input feature for cluster ids as an ordinal variable.
    iii) Creating an input feature for cluster centroids as a continuous variable.
    iv) Creating an input feature for cluster size as a continuous variable.
    Options:
    a) 1 only
    b) 2 only
    c) 3 and 4
    d) All of the above

11. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?
    a) Proximity function used
    b) of data points used
    c) of variables used
    d) All of the above

Q12 to Q14 are subjective answers type questions, Answers them in their own words briefly

12. Is K sensitive to outliers?
    - Yes. K means is sensitive to outliers.

      For e.g. Data set point are 1 2 3 7 8 80

      Now 80 is outlier.

      K=2

      C1=1 C2=7

      After first iteration

      C1=2 C2=31.67

      As 80 data point which is outlier comes in cluster 2.

      Cluster 2 centroid changes to accommodate 80 .

      Therefore K means is sensitive to outliers

MACHINE LEARNING

13. Why is K means better?
    - It is relatively simple to implement, it Scales to large data sets. Guarantees convergence, it Can warm-start the positions of centroids. It easily adapts to new examples. Also Generalizes to clusters of different shapes and sizes, such as elliptical clusters.
    - It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.
    - It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.
    - It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.
    - The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

14. Is K means a deterministic algorithm?
    - K-Means has many drawbacks too. One of the significant drawbacks of K-Means is its **non-deterministic nature**. K-Means starts with a random set of data points as initial centroids. This random selection influences the quality of the resulting clusters.
    - The basic k-means clustering is based on a non-deterministic algorithm
    - This means that running the algorithm several times on the same data, could give different results