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

Answer 1 to 11

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

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

3. Can decision trees be used for performing clustering?
   a)
   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)
   b) 2 only
   c)
   d)

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

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

7. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means?
   a) Yes
   b)
   c)
   d)

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 with a bad local minimum.
   iii) Centroids do not change between successive iterations.
   iv) Terminate when RSS falls below a threshold.
   Options:
   a)
   b)
   c)
   d) All of the above

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

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.
    vi) Creating an input feature for cluster size as a continuous variable.

Options:
   a)
   b)
   c)
   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)
   b)
   c)
   d) All of the above

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

12. Is K-means sensitive to outliers?

Ans. Yes, K-means clustering algorithm is sensitive to outliers. Outliers can significantly affect the centroid calculation and the assignment of data points to clusters. Outliers are data points that are significantly different from the other data points in the dataset, and they can distort the distribution of the data. K-means algorithm tries to minimize the sum of squared distances between the data points and the centroids of the clusters. Outliers can have a large influence on this sum, and they can cause the centroids to shift towards the outliers. As a result, the clusters may not be representative of the underlying structure of the data, and the clustering results may be inaccurate.

13. Why is K means better?

Ans. K-means clustering algorithm is a popular and efficient clustering method because of its simplicity and computational efficiency. It works by iteratively partitioning the data into K clusters based on the similarity between data points, using a distance metric such as Euclidean distance. K-means is fast and scalable, making it suitable for large datasets. It is also easy to implement and interpret, and it can handle a variety of data types, including continuous and categorical variables. Overall, K-means is a powerful and versatile clustering algorithm that can be used in a wide range of applications.

14. Is K means a deterministic algorithm?

Ans. Yes, K-means clustering algorithm is a deterministic algorithm. This means that if we run the algorithm on the same dataset multiple times, with the same initial conditions (such as the number of clusters, the initialization method, and the distance metric), it will always produce the same clustering results.

The reason for this determinism is that K-means algorithm uses a fixed set of rules to assign data points to clusters and update the cluster centroids iteratively.