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?
13. Why is K means better?
14. Is K means a deterministic algorithm?

12. In statistics, the word "k" is not frequently used. It is not clear which particular statistic is being discussed as a result. Though generally speaking, some statistics are more susceptible to outliers than others. For instance, the median and mode are less impacted by outliers than the mean, which is sensitive to outliers because extreme values can push the median in their direction.

It is crucial to carefully consider whether there are any outliers in the data and to select statistical methods that are resistant to their effects, such as using the median rather than the mean or non-parametric tests.

FLIP ROBO

MACHINE LEARNING

13.K-means is a clustering algorithm that groups data points that are similar together. It is preferred for a variety of reasons. To begin with, K-means is computationally efficient and scalable, making it appropriate for large datasets. It is also adaptable, allowing the user to specify the number of clusters.

K-means has been demonstrated to perform well in a wide range of clustering tasks and to produce meaningful and useful results. It is simple to comprehend and explain, making it suitable for a wide range of users. K-means can also be used to identify groups of similar data points, which can aid in decision making in a variety of fields.
K-means is a flexible and powerful clustering algorithm that can be applied in a variety of situations to combine related data points based on their similarity.

14. Yes, K-means is a deterministic algorithm. This means that for a given dataset and set of initial conditions (such as the number of clusters and the starting position of centroids), the algorithm will always produce the same clustering result. However, the algorithm is sensitive to the initial conditions, and different initial conditions can lead to different results. Therefore, it is often necessary to run the algorithm multiple times with different initial conditions to ensure that the optimal clustering solution has been found.