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

   Ans : Classification

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

   Ans : 1, 2 and 4

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

   Ans : True

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

   Ans : 1 only

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

   Ans : 1

MACHINE LEARNING

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

   Ans : 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

   Ans : Yes

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) 1, 3 and 4
   b) 1, 2 and 3
   c) 1, 2 and 4
   d) All of the above

   Ans : 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

   Ans : K-means 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

    Ans : 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

    Ans : All of the above

MACHINE LEARNING

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

12. Is K sensitive to outliers?

Ans :

The k means algorithm updates the cluster centers by taking the average of all the data points that are closer to each cluster center. When all the points are packed nicely together, the average makes sense.

However, when you have outliers, this can affect the average calculation of the whole cluster. As a result, this will push your cluster center closer to the outlier.

For ex :

Data set point are 1, 2, 3, 7, 8, 80
Now 80 is outlier.
If I take K value=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

13. Why is K means better?

Ans :

K means algorithms are deployed to subdivide data points of a dataset into clusters based on nearest mean values. To determine the optimal division of your data points into clusters, such that the distance between points in each cluster is minimized, one can use the k means clustering algorithm.

The procedure follows a simple and easy way to classify a given data set through a certain number of clusters. The main idea is to define k centers, one for each cluster. These centers should be placed with subtlety because a different location causes a different result. The next step is to take each point belonging to a given data set and associate it with the nearest center.

When no data is pending, the first step is completed, and an early group age is done. At this point, we need to re-calculate new k-centroids as the barycenter of the clusters resulting from the previous step. After we have these new k-centroids, a contemporary binding must be done between the same data set points and the nearest new center. A loop is generated. As a result of this loop, we may notice that the k centers change their location step-by-step until no more changes are done.

Below are the advantages mentioned:

- It is fast and Robust
- Comparatively efficient
- If data sets are distinct, then gives the best results
- Produce tighter clusters
- When centroids are recomputed, the cluster changes.
- Better computational cost and Enhances Accuracy
- Generalize clusters of different shapes and sizes, such as elliptical clusters
- Relatively simple to implement
- Scales to large data sets and Guarantees convergence

MACHINE LEARNING

As we mentioned above advantages it also have some disadvantages like Manual Selection, Dependent on initial values ,Clustering data of varying sizes and densities, Clustering outliers and Scaling with a number of dimensions.

K means gives more weight to the bigger clusters. K means assumes spherical shapes of clusters (with radius equal to the distance between the centroid and the furthest data point) and doesn't work well when clusters are in different shapes such as elliptical clusters.

There is no such thing as a good method or good algorithm without the context of a problem it is used in. Thus we do use k-means because there are problems for which k-means is an optimal solution.

14. Is K means a deterministic algorithm?

Ans :

No, Due to its Random selection of the data points as a initial centroids. Running the algorithm several times it gives different results in each occurrence so it is not a deterministic algorithm.

MACHINE LEARNING