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

Answers: a) 2 Only

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

Answers: d) 1, 2 and 4

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

Answers: b) 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

Answers: a) 1 only

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

Answers: b) 1

MACHINE LEARNING

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

Answers: 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

Answers: a) 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

Answers: 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

Answers: a) 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

Answers: a) 1 only

MACHINE LEARNING

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

Answers: d) All of the above

12. Is K sensitive to outliers?

Answers: Yes, K is sensitive to outliers as it uses the mean of cluster data points to find the cluster center. When all the points are making clusters together, the mean makes sense. However, when we have outliers, this can affect the mean calculation of the whole cluster. As a result, this will push your cluster center closer to the outlier. The algorithm aims to minimize the squared Euclidean distances between the observation and the centroid of cluster to which it belongs. But sometime K-Means algorithm does not give best results. It is sensitive to outliers. An outlier is a point which is different from the data set makes next observation totally different.

13. Why is K means better?

Answers: K means clustering is better that clustering algorithm like Hierarchical and DBScan clustering because it can handle big data well also the time complexity of K means is linear ie $O(n)$ while that hierarchical clustering is quadratic ie $O(n2)$. And by the use of elbow method we can easily determine the perfect cluster number of the model.

14. Is K means a deterministic algorithm?

Answers: K means is a non a deterministic nature algorithm. It has so many drawbacks that it starts with a random set of data points as initial centroids. This random selection influences the quality of the resulting of clusters. It means that running the algorithm several times on the same data, could give different results. The key idea of the algorithm is to select data points which belong to dense regions and which are adequately separated in feature space as the initial centroids.

MACHINE LEARNING