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

2. Sentiment Analysis is an example of:
   i) Regression
   ii) Classification
   iii) Clustering
   iv) Reinforcement
   Options:

   a) 1, 2 and 4

3. Can decision trees be used for performing clustering?

   a) 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

5. What is the minimum no. of variables/ features required to perform clustering?

   a) 1

6. For two runs of K-Mean clustering is it expected to get same clustering results?

   a) No

7. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means?

   a) Yes

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) All of the above

9. Which of the following algorithms is most sensitive to outliers?

   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) 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) 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?

    a) K-Means clustering does not give best results when outliers are present. This is because in this algorithm, Mean is used and mean varies depending on the values. For eg: the mean of values 2,4,6,8,10 is 6. If we have an outlier let's say, 200 then mean of 2,4,6,8,10,200 becomes 38.333 and we understand that an outlier causes the mean to be increased by about 32 units. This shows that K-Means are sensitive to Outliers.

    We can use K-median or K-medoid in case there are many outliers to reduce the risk of sensitivity or remove the outliers if the data is large or use capping and flooring method.

13. Why is K means better?

    a) K means is an easy and simple algorithm which can be used in market segmentation, document clustering, image segmentation, etc. it can be used in a large dataset and guarantees convergence.it easily adapts to new examples and generalizes to clusters of different shapes and sizes, such as elliptical clusters. While k means is better it has few drawbacks as well. Before using k means, I prefer to use the elbow method and find the good k number of clusters.

MACHINE LEARNING

14. Is K means a deterministic algorithm?

a) The basic kmeans is non-deterministic algorithm as it gives different results on running the algorithm several times. K-means starts with randomly chosen cluster centroids so to find optimal ones. Except for initialization, kmeans is deterministic in all other steps.