MACHINE LEARNING

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

1. Movie Recommendation systems are an exampleof:
   i) Classification
   ii) Clustering
   iii) Regression

   Options:
   a) 2 Only
   b) 1 and2
   c) 1 and3
   d) 2 and3

ANS- 2 and 3

2. Sentiment Analysis is an exampleof:
   i) Regression
   ii) Classification
   iii) Clustering
   iv) Reinforcement
   v)

   Options:
   a) 1 Only
   b) 1 and2
   c) 1 and3
   d) 1, 2 and4

ANS- 1 ,2 and 4

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

ANS- True

4. Whichofthefollowingisthemostappropriatestrategyfordatacleaningbeforeperformingclustering analysis, given less than desirable number of data points:
   i) Capping and flooring ofvariables
   ii) Removal of outliers

   Options:
   a) 1only
   b) 2only
   c) 1 and 2
   d) None of theabove

ANS- 1 only

5. What is the minimum no. of variables/ features required to performclustering?
   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 clusteringresults?
   a) Yes
   b) No

ANS- No

7. Isitpossiblethat Assignmentofobservationstoclustersdoesnotchangebetweensuccessive iterations inK-Means?
   a) Yes
   b) No
   c) Can'tsay
   d) None ofthese

ANS- Yes

8. Which of the following can act as possible termination conditions inK-Means?
   i) For a fixed number ofiterations.
   ii) Assignment of observations to clusters does not change between iterations. Except for cases witha bad localminimum.
   iii) Centroids do not change between successiveiterations.
   iv) Terminate when RSS falls below a threshold.

   Options:
   a) 1, 3 and4
   b) 1, 2 and3
   c) 1, 2 and4
   d) All of theabove

ANS- All of the above

9. Which of the following algorithms is most sensitive tooutliers?
   a) K-means clusteringalgorithm
   b) K-medians clusteringalgorithm
   c) K-modes clusteringalgorithm
   d) K-medoids clusteringalgorithm

ANS- K-means clusteringalgorithm

10. HowcanClustering(UnsupervisedLearning)beusedtoimprovetheaccuracyofLinearRegression model (SupervisedLearning):
    i) Creating different models for different clustergroups.
    ii) Creating an input feature for cluster ids as an ordinalvariable.
    iii) Creating an input feature for cluster centroids as a continuousvariable.
    iv) Creating an input feature for cluster size as a continuousvariable.

    Options:
    a) 1only
    b) 2only
    c) 3 and 4
    d) All of theabove

ANS- All of theabove

11. Whatcouldbethepossiblereason(s)forproducingtwodifferentdendrogramsusingagglomerative clustering algorithms for the samedataset?
    a) Proximity functionused
    b) of data pointsused
    c) of variablesused
    d) All of theabove

ANS- All of theabove

MACHINE LEARNING

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

12. Is K sensitive tooutliers?

ANS-

K-means can be quite sensitive to outliers.

So if you think you need to remove them, I would rather remove them first, or use an algorithm that is more robust to noise. For example k medians is more robust and very similar to k-means, or you use DBSCAN.

K-means can be used as outlier detection. BUT, more attention needs to be given for the definition of outliers. In K-means, using the symmetric distance measure is the key component to define the samples that belonging to the same cluster. symmetric distance measurement gives similar weight to each dimension (feature) this may not always be the case for defining outliers.

13. Why is K meansbetter?

ANS– K-means is an unsupervised clustering algorithm designed to partition unlabelled data into a certain number (thats the " *K*") of distinct groupings. In other words, k-means finds observations that share important characteristics and classifies them together into clusters. A good clustering solution is one that finds clusters such that the observations within each cluster are more similar than the clusters themselves.

K means is better because of following reasons….

Relatively simple to implement.

Scales to large data sets.

Guarantees convergence.

Can warm-start the positions of centroids.

Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

14. Is K means a deterministicalgorithm?

ANS- K-Means is a non-deterministic algorithm. This means that a compiler cannot solve the problem in polynomial time and doesn't clearly know the next step. This is because some problems have a great degree of randomness to them. These algorithms usually have 2 steps — 1)Guessing step 2)Assignment step. On similar lines is the K-means algorithm. The K-Means algorithm divides the data space into K clusters such that the total variance of all data points with respect to the cluster mean is minimized.