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 – 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
Ans- d)1,2and4

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

Ans- 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
b) 2 only
c) 1 and 2
d) None of the above

ans- a) 1 only

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

Ans-b)1

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
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

ans:-All 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- d)all 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 the above

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

12. Is K sensitive to outliers?

The k-means algorithm update cluster centers by taking average of all data points that are closer to each cluster center. When all the points are packed nicely together, the average makes sense. However, when we have outlier, this can affect the average calculation of the whole cluster. As a result, this will push our cluster center closer to the outlier.
An example:
10+50+40+60 = average is 160/4= 40
But if we add an outlier in the same data like
10+50+40+600 = average is 700/4 = 175
Note that two averages are widely different from one another.

13. Why is K means better?

k-means is an unsupervised algorithm which is used to find discover the patterns in the data. In other means to find the clusters in the data such as grouping customers by purchasing behavior.K-Means for Clustering is one of the popular algorithms for this approach. Where K means the number of clustering and means implies the statistics mean a problem. The algorithm clusters into k groups and here k is the input parameter. In this procedure, a dataset is classified through a certain number of clusters, commonly known as k clusters and the main idea is to define k centres, one for each cluster. These centres should be placed in a way since different location causes different results. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a
given data set and associate it to the nearest centre. When no point is pending, the first step is completed and an early group age is done. However, the main disadvantage is one has to specify the number of clusters as an input in the algorithm. In short:
1. Select the k values.

2. Initialize the centroids.
3. Select the group and find the average

Advantages of K-means

1. It is very simple to implement.
2. It handles huge and large dataset efficiently.
3. It generalizes the clusters for different shapes and sizes

14. Is K means a deterministic algorithm?

No, rather K-means is a non-deterministic algorithm because of its random selection of data points as initial centroids. And because of this randomness it give different results on different execution for the same dataset. This non deterministic nature of K-means algorithm limits its applicability in areas such as cancer subtype prediction. To handle this issue, we can use density based version of K-Means which involves a novel and systematic method for selecting initial centroids.