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: d) 2 and 3

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) 1,2 and 4

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 only

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

Ans: b) No

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

MACHINE LEARNING

a) Yes
b) No
c) Can't say
d) None of these

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

Ans: 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

Ans: a) and c)

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 continuousvariable.

Options:

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

Ans: d) 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: d) 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?

Ans: K means clustering is an unsupervised learning algorithm which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest centroid. The algorithm aims to minimize the squared Euclidean distances between the observation and the centroid of cluster to which it belongs. It is sensitive to outliers. An outlier is a point which is different from the rest of data points.

13. Why is K means better?

Ans: We observe that the outliers show up as a separate cluster and also cause other clusters to merge which suggests clustering was not efficient when outliers were included in data set.

MACHINE LEARNING

Even though the outliers were about 2 percent of non-outliers which is common in real world data sets, they had a significant impact on clustering. Hence it is better to identify and remove outliers before applying K-means clustering algorithm.

14. Is K means a deterministic algorithm?

Ans: Hierarchical Agglomerative Clustering is deterministic except for tied distances when not using single-linkage. DBSCAN is deterministic, except for permutation of the data set in rare cases. k-means is deterministic except for initialization. You can initialize with the first k objects, then it is deterministic, too.