MACHINE LEARNING ASSIGNMENT - 2

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. 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)

3. Can decision trees be used for performing clustering?

a) True

b) False

Ans: (a)

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)

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

a) 0

b) 1

c) 2

d) 3

Ans: (b)

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

a) Yes

b) No

Ans: (b)

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: (a)

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

Ans: (d)

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)

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)

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)

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

12. Is K sensitive to outliers?

Ans: It totally depends on the user, which algorithm is being used to calculate the K.

There are two K based methods i.e K-means and KNN which can be sensitive to outliers because these methods depends on the distance between data points to group and that distance can be affected by outliers.

13. Why is K means better?

Ans: It is most widely used unsupervised machine learning algorithm that used to seperate the dataset into number of clusters. For clustering, K-means is considered to be a better algorithm due to below following reasons: -

Simple and Easy to Implement - In comparison of other algorithms, K-means algorithm is simple and easy to implement because It only requires the number of clusters and the data points to be clustered.

Cluster Interpretability -The clusters produced by K-means algorithm are easily interpretable. And in each cluster there is similar data points are available that's make easier to understand the underlying patterns in the data.

Versatility: It can be apply on wide range of applications K-means algorithm like customer segmentation, image segmentation, text clustering etc.

14. Is K means a deterministic algorithm?

Ans: No, K-means is not a deterministic algorithm because it can produce different result on each execution with the same input data and parameters.

According to the assignment of data points centroid also get updated that can affect the final clustering results. So, to bed rid from this randomness, K-means is often run multiple times with different random initializations. And after analysing the results of each run, the result having the lowest total sum of squared distances between points is the selected and their assigned centroid is the final result.