ASSIGNMENT – 2

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

Ans: b) 1 and 2

2. Sentiment Analysis is an example of:

i) Regression ii) Classification iii) Clustering iv) Reinforcement

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

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: b) 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: a) 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.

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) 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.

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 :

A cluster refer to the set of data points collected together due to their similarities. We consider the number 'K' as the centroid of that given dataset. In K-means algorithm, centroid is defined by simple mean of the cluster which in terms leads to high variation due to outliers. Thus 'K' in K-means is sensitive to outliers. This can be explained with following example:

Data set: [1,1,1,2,2,2,3,3,3,50]

Mean =6.8

Median = 2

Mode = 3

Here we can observe that, mean is highly affected by the outlier which is '50'

13. Why is K means better?

Ans:

K-means algorithms are simple and easy to understand. Time consumed for clustering the K-means datapoints varies linear to number of data points. If outliers are removed, no other clustering algorithm performs efficient than this.

14.Is K means a deterministic algorithm?

Ans: The k-means clustering is a non-deterministic algorithm. This means that running the algorithm many times on the same datapoints, will give different results. The non-deterministic nature of K-Means is due to its random selection of data points as centroids of the clusters.