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:

ANS: a) 2 Only

2. Sentiment Analysis is an example of:

ANS: d) 1, 2 and 4

3. Can decision trees be used for performing clustering?

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:

ANS: a) 1 only

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

ANS: b) 1

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

ANS: b) No

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

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 with a 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?

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. Options:

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?

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 is highly sensitive to outliers as means varies drastically as the value of the outliers increases. K-means algorithm updates the cluster centers by taking the average of all the data points that are closer to each cluster center. When all the points are packed together, the average makes sense. However, when there are outliers, it will affect the average calculation of the whole cluster, thus pushing the cluster center closer to the outlier.

13. Why is K means better?

ANS: K-means is better because its relatively simple to implement. I can be used/implemented on large data sets and it guarantees convergence. With varying widths across clusters, we can improvise the result. It also generalizes to clusters of different shapes and sizes, such as elliptical clusters.

14. Is K means a deterministic algorithm?

ANS: K-means is a non-deterministic algorithm because of its random selection of data points as initial centroids. This random selection influences the quality of the resulting clusters. Besides, each run of the algorithm for the same dataset may yield a different output