**ML worksheet 2**

1. Movie Recommendation systems are an example of:

Ans: Clustering

2. Sentiment Analysis is an example of:

Ans : Regression, classification and reinforcement learning

3. Can decision trees be used for performing clustering?

Ans: 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:

Ans : Capping and flouring of variables

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

Ans: 1

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

Ans: No

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

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.

Ans: All of the above

9. Which of the following can act as possible termination conditions in K-Means?

i) K- Means clustering algorithm

ii) Agglomerative clustering algorithm

iii) Expectation-Maximization clustering algorithm

iv) Diverse clustering algorithm

Ans: 1 and 3

10. 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

11. 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: All the above

12. 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 of the above

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly

13. Is K sensitive to outliers?

Ans: The K-means clustering algorithm is sensitve to outliers, because a mean is easily influenced by extreme values. K-medoids clustering is a variant of K-means that is more robust to noises and outliers. ... The group of points in the right form a cluster, while the rightmost point is an outlier.

14. Why is K means better?

Ans: K-means is like the Exchange Sort algorithm. Easy to understand, helps one get into the topic, but should never be used for anything real, ever. In the case of Exchange Sort, even Bubble Sort is better because it can stop early if the array is partially sorted.

15. 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 deterministic except for initialization. You can initialize with the first k objects, then it is deterministic, too.