MACHINE LEARNING 2

- 1. Movie Recommendation systems are an example of:
- i) Classification ii) Clustering iii) Regression
- b) 1 and 2
- 2. Sentiment Analysis is an example of:
- i) Regression ii) Classification iii) Clustering iv) Reinforcement
- d) 1, 2 and 4
- 3. Can decision trees be used for performing clustering?
- 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
- a) 1 only
- 5. What is the minimum no. of variables/ features required to perform clustering?
- b) 1
- 6. For two runs of K-Mean clustering is it expected to get same clustering results?
- b) No
- 7. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means?
- 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.
- d) All of the above
- 9. Which of the following algorithms is most sensitive to outliers?
- 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.
- 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?

d) All of the above

12. Is K sensitive to outliers?

Yes, K sensitive to outliers because the outlier increases the mean of data by about 10 units. This is a significant increase in all data points range from 0 to 1. This shows that the mean is influenced by outliers. Since K-Means algorithm is about finding mean of clusters, the algorithm is influenced by outliers and a mean is easily influenced by extreme values. The group of points in the right form a cluster, while the rightmost point is an outlier.

13. Why is K means better?

- a) If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering
- b) if k is small, then K-Means produce tighter clusters
- c) K-means attempts to minimize the total squared error

14. Is K means a deterministic algorithm?

The number of clusters identified from data by algorithm is represented by 'K' in K-means. In this algorithm, the data points are assigned to a cluster in such a manner that the sum of the squared distance between the data points and centroid would be minimum.

No, K means is not a deterministic algorithm. It is based on a non-deterministic algorithm. This means that running the algorithm several times on the same data, could give different results. The non-deterministic nature of K-Means is due to its random selection of data points as initial centroids. The key idea of the algorithm is to select data points which belong to dense regions and which are adequately separated in feature space as the initial centroids.