MACHINE LEARNING ASSIGNMENT 2

1. Movie recommendation systems are an example of:
   i) Classification
   ii) Clustering
   iii) Regression
   Options:
   - 2 only
   - 1 and 2
   - 1 and 3
   - 2 and 3

Answer = Classification and Clustering, options (i) and (ii)

2. Sentiment Analysis is an example of:
   i) Regression
   ii) Classification
   iii) Clustering
   iv) Reinforcement
   Options:
   - 1 only
   - 1 and 2
   - 1 and 3
   - 1,2 and 4

Answer = Regression, Classification and Reinforcement, options (i), (ii) and (iv)

3. Can decision trees be used for performing clustering?

Answer = 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:
   - 1 only
   - 2 only
   - 1 and 2
   - None of the above

Answer = Capping and flooring of variables, option (i)

5. What is the minimum no. of variables/features required to perform clustering?
   i) 0
   ii) 1
   iii) 2
   iv) 3

Answer = 1 option, (ii)

6. For two runs of K-Mean clustering is it expected to get same clustering results?
   i) Yes
   ii) No

Answer = No, option (ii)

7. Is it possible that assignment of observations to cluster does not change between successive iterations in K-Mean?
   i) Yes
   ii) No
   iii) Can't say
   iv) None of these

Answer = No, option (ii)

8. Which of the following can act as possible termination conditions in K-Mean?
   i) For a fixed number of iterations
   ii) Assignment of observations to cluster does not change between iteration. Except for the cases with a bad local minimum.
   iii) Centroids do not change between successive iterations.
   iv) Terminates when RSS falls below a threshold.
   Options:
   - 1,3 and 4
   - 1,2 and 3
   - 1,2 and 4
   - All of the above

Answer = All of the above, options: (i), (ii), (iii) and (iv)

9. Which of the following algorithms is most sensitive to outliers?
   i) K-means clustering algorithms
   ii) K-medians clustering algorithms
   iii) K-modes clustering algorithms
   iv) K-medoids clustering algorithms

Answer = K-means clustering algorithms, option (i)

10. How can clustering (Unsupervised Learning) be used to improve the accuracy of Liner 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:
    - 1 only
    - 2 only
    - 3 and 4
    - All of the above

Answer = Creating different models for different cluster groups, option (i)

11. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?
    i) Proximity function used
    ii) Of data points used
    iii) Of variables used
    iv) All of the above

Answer = All of the above, option (iv)

12. Is K sensitive to outliers?

Answer = The *K*-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. The algorithm aims to minimize the distances between the observation and the centroid of cluster to which it belongs. But sometime K-Means algorithm does not give best results. It is sensitive to outliers. An outlier is a point which is different from the rest of data points.

13. Why is K-mean better?

Answer = The K-means is used to find groups which have not been explicitly labelled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets. It is one of the most robust methods, especially for image segmentation and image annotation projects. If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls. K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

14. Is K-means a deterministic algorithm?

Answer = The k-means clustering is a non-deterministic algorithm which means that running the algorithm several times on the same data, could give different results.