Machine Learning Worksheet

1. Movie Recommendation systems are an example of:
   i) Classification
ii) Clustering
iii) Regression

   Ans: a) 2 Only

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?
   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:
   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?
   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 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?
   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?
    Ans: d) All of the above

12. Is K sensitive to outliers?
    Ans: The K-means clustering algorithm is sensitive 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.

13. Why is K means better?
    Ans: High Performance. It is easy to use, Unlabeled data, Result Interpretation

14. Is K means a deterministic algorithm?
    Ans: K-means is non-deterministic in nature, because of its random selection of data points as initial centroids.

We propose an improved, density based version of K-means, which involves a novel and systematic method for selecting initial centroids.