MACHINE LEARNING WORKSHEET -2

Q1 TO 11

1. Movie Recommendation systems are an example of: i) Classification.
ii) Clustering
iii) Regression Options:
a) 2 Only
b) 1 and 2
c) 1 and 3
d) 2 and 3

Ans-A

2. Sentiment Analysis is an example of.
i) Regression
ii) Classification
iii) Clustering
iv) Reinforcement Options:
a) 1 Only
b) 1 and 2
c) 1 and 3
d) 1, 2 and 4

Ans-D

3. Can decision trees be used for performing clustering.
a) True
b) False

Ans-A

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
a) 1 only
b) 2 only
c) 1 and 2
d) None of the above

Ans-A

5. What is the minimum no. of variables/ features required to perform clustering.
a) 0
b) 1
c) 2
d) 3

Ans-B

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

Ans-B

7. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means.
a) Yes
b) No
c) Can't say
d) None of these

Ans-A

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. Options
      a) 1, 3 and 4
      b) 1, 2 and 3
      c) 1, 2 and 4
      d) All of the above

Ans-D

9. 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

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:
      a) 1 only
      b) 2 only
      c) 3 and 4
      d) All of the above

Ans-D

11. 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-D

Q12 TO 14

12. Is K sensitive to outliers

ANS-We observe that the outlier increases the mean of data by about 10 units. This is a significant increase considering the fact that 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. Let us take an example to understand how outliers affect the K-Means algorithm using python. We have a 2 dimensional data set called 'cluster' consisting of 3000 points with no outliers. We get the following scatter plot after K-means algorithm is applied.

13. Why is K means better

ANS-K-Means for Clustering is one of the popular algorithms for this approach. Where K means the number of clustering and means implies the statistics mean a problem. It is used to calculate code-vectors (the centroids of different clusters). According to a tutorial, for any word/value/key that needs to be 'vector quantized', it is by calculating the distance from all the code vectors and assign the index of the code vector with the minimum distance to this value. For example, clustering can be applied to MP3 files, cellular phones are the general areas that use this technique.

14. Is K means a deterministic algorithm

ANS-A deterministic algorithm is simply an algorithm that has a predefined output. For instance if you are sorting elements that are strictly ordered(no equal elements) the output is well defined and so the algorithm is deterministic. In fact most of the computer algorithms are deterministic.