MACHINE LEARNING WORKSHEET-2

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

Answer : 2 only (i.e Clustering)

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

Answer : 1,2 and 4

3. **Can decision trees be used for performing clustering?**

a) True
b) False

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:

a) 1 only

b) 2 only

c) 1 and 2

d) None of the above

Answer : 1 Only

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

a) 0

b) 1

c) 2

d) 3

Answer : 1

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

a) Yes

b) No

Answer : No

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

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

Answer : All of the above

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

Answer : 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.

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

Answer : 1 Only

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

Answer : All of the above

12. Is K sensitive to outliers?

Answer : Yes , it is sensitive to outliers, because In K-Means clustering outliers are found by distance based approach and cluster based approach. In case of hierarchical clustering, by using dendrogram outliers are found. The goal of the project is to detect the outlier and remove the outliers to make the clustering more reliable. Also , Mean is easily influenced by extreme values. K-medoids clustering is a variant of K-means that is more robust to noises and outliers. Instead of using the mean point as the center of a cluster, K-medoids uses an actual point in the cluster to represent it.

We observe that the outlier **increases the mean of data by about 10 unit.**

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.

13. Why is K means better?

Answer : 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 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.

According to some users, K-means is very simple and easy to implement. However, it is unlikely to be the state-of-the-art, but for straightforward clustering, it is also a part of a larger data-processing pipeline, K-means is a reasonable default choice, at least until you figure out that the clustering step is your bottleneck in terms of overall performance.

K-means is used to learn feature representations for images (use k-means to cluster small patches of pixels from natural images, then represent images in the basis of cluster centres; repeat this several times to form a “deep” network of feature representations) gives image classification results that are competitive with much more complex / intimidating deep neural network models. In fact, a lot of k-means applications are now done using support vector machines.

It gives good results:

1. It is already implemented in the software
2. Number of clusters has to be fixed before
3. Dependent of the initialisation parameters and the chosen distance

14. Is K means a deterministic algorithm?

Answer : No. K-means is not a deterministic algorithm . The basic k-means clustering is based on a non-deterministic algorithm. This means that running the algorithm several times on the same data, could give different results.

This means that a compiler cannot solve the problem in polynomial time and doesn’t clearly know the next step. This is because some problems have a great degree of randomness to them. These algorithms usually have 2 steps —

1)Guessing step

2)Assignment step.

On similar lines is the K-means algorithm. The K-Means algorithm divides the data space into K clusters such that the total variance of all data points with respect to the cluster mean is minimized. However, the approach that compiler takes does not involve Multivariate Calculus as it seems. Rather, the approach taken is iterative.