Machine learning assignment 2

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

Ans : 1 and 2

2)Sentiment Analysis is an example of: i) Regression ii) Classification iii) Clustering iv) Reinforcement

Ans : 1,2 and 4

3) 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 : Capping and flooring of variables

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

Ans : 1

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

Ans : no

7) Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means?

Ans : 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 : 1,2,3,4

9) Which of the following algorithms is most sensitive to outliers?

Ans : 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 : 1,2,3,4

11)What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?

Ans : Proximity function used , of data points used , of variables used

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. The k-means algorithm updates the cluster centres by taking the average of all the data points that are closer to each cluster centre. When all the points are packed nicely together, the average makes sense. However, when you have outliers, this can affect the average calculation of the whole cluster. As a result, this will push your cluster centre closer to the outlier.

13) Why is K means better?

14) Is K means a deterministic algorithm?

Ans : K-Means is a non-deterministic algorithm. 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. Now, like any deterministic algorithm it has 2 phases. Guessing phase: Randomly initializing k means in the data space(Mu(k)s). Now, all the data points X(i)s (1,m) are assigned to clusters in accordance to which cluster mean they are closer to. Mathematically, this step tries to minimize the within cluster variance. Hence, every point is now assigned a cluster. Next is the assignment step. All the cluster means (Mu(k)s)are now assigned to the mean of the data points in the cluster. This step is repeated a couple of times.

The K Means Algorithm

Now similar to the most of non-deterministic algorithms, K-Means has a bad habit. Which is that every time you run a K-Means clustering it would give you different results. The situation gets even worsened when you are unsure if the any modification to the K-Means would improve the results. Refer to the image below to see how bad sometimes can a K-Means Algorithm's result get.

We should understand that we cannot do much about this issue as it is almost impossible to analytically imagine a data-space so huge. Although, we do have a couple of suggestions to follow.

How to choose the value of K?

One should rely on the problem statement for this. For example in a tree specie classification problem, if one know the number of possible specie and given that all of them appear in significant numbers in the data-set, one can assign the number of species to K.

2. How to be sure if the solution obtained is appropriate?

There is no versatile approach for this issue. Rather one should focus on initializing the cluster means with the best possible estimate. There could be any statistical approach for this. For example — One could assign the calculated mean of a few data points which one is sure would fall in the same cluster. Secondly, one can iterate 10–100 times a K-Means Algorithm and decide the one which best minimizes the Mathematical function given above.