ASSIGNMENT – 2 **MACHINE LEARNING**

1. Movie Recommendation systems are an example of:

   i) Classification
   ii) Clustering
   iii) Regression

   Answer: d) 2 and 3

   - The movie recommendation system can be viewed as a reinforcement learning problem where it learns by its previous recommendations and improves the future recommendations.

2. Sentiment Analysis is an example of:

   i) Regression
   ii) Classification
   iii) Clustering
   iv) Reinforcement

   Answer: d) 1, 2 and 4

   ♦ Sentiment analysis at the fundamental level is the task of classifying the sentiments represented in an image, text or speech into a set of defined sentiment classes like happy, sad, excited, positive, negative, etc. It can also be viewed as a regression problem for assigning a sentiment score of say 1 to 10 for a corresponding image, text or speech.

3. Can decision trees be used for performing clustering?

   Answer: a) True

   ➢ Decision trees can also be used to for clusters in the data but clustering often generates natural clusters and is not dependent on any objective function.

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

   Answer: a) 1 only

   - Removal of outliers is not recommended if the data points are few in number. In this scenario, capping and flouring of variables is the most appropriate strategy.

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

   Answer: b) 1

   - At least a single variable is required to perform clustering analysis. Clustering analysis with a single variable can be visualized with the help of a histogram.

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

   Answer: b) No

   - K-Means clustering algorithm instead converses on local minima which might also correspond to the global minima in some cases but not always. Therefore, it's advised to run the K-Means algorithm multiple times before drawing inferences about the clusters.

   ♦ However, note that it's possible to receive same clustering results from K-means by setting the same seed value for each run. But that is done by simply making the algorithm choose the set of same random no. for each run.

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

Answer: a) Yes

- When the K-Means algorithm has reached the local or global minima, it will not alter the assignment of data points to clusters for two successive iterations.

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 bad local minimum.
iii) Centroids do not change between successive iterations.
iv) Terminate when RSS falls below a threshold

Answer: d) All of the above

All four conditions can be used as possible termination condition in K-Means clustering:

1. This condition limits the runtime of the clustering algorithm, but in some cases the quality of the clustering will be poor because of an insufficient number of iterations.
2. Except for cases with a bad local minimum, this produces a good clustering, but runtimes may be unacceptably long.
3. This also ensures that the algorithm has converged at the minima.
4. Terminate when RSS falls below a threshold. This criterion ensures that the clustering is of a desired quality after termination. Practically, it's a good practice to combine it with a bound on the number of iterations to guarantee termination.

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

Answer : a) K-means clustering algorithm

- Out of all the options, K-Means clustering algorithm is most sensitive to outliers as it uses the mean of cluster data points to find the cluster center.

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

Answer: d)All of the above

- Creating an input feature for cluster ids as ordinal variable or creating an input feature for cluster centroids as a continuous variable might not convey any relevant information to the regression model for multidimensional data. But for clustering in a single dimension, all of the given methods are expected to convey meaningful information to the regression model. For example, to cluster people in two groups based on their hair length, storing clustering ID as ordinal variable and cluster centroids as continuous variables will convey meaningful information.

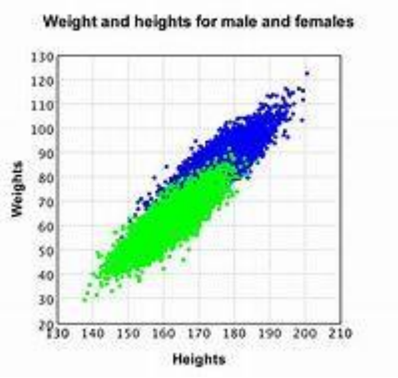
11. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?

Answer: d) All of the above

- Change in either of Proximity function, no. of data points or no. of variables will lead to different clustering results and hence different dendrograms.

12. Is K sensitive to outliers?

- K-means can be quite sensitive to outliers. So if you think you need to remove them, I would rather remove them first, or use an algorithm that is more robust to noise.
- K-means, using the symmetric distance measure is the key component to define the samples that belonging to the same cluster. symmetric distance measurement gives similar weight to each dimension (feature) this may not always be the case for defining outliers.
- For example :k medians are more robust and very similar to k-means, or you use DBSCAN.



13. Why is K means better?

- Relatively simple to implement.
- Scales to large data sets.
- Guarantees convergence.
- Can warm-start the positions of centroids.
- Easily adapts to new examples.
- Generalizes to clusters of different shapes and sizes, such as elliptical clusters**.**

k-means Generalization:

Compare the intuitive clusters on the left side with the clusters actually found by k-means on the right side. The comparison shows how k-means can stumble on certain datasets.

Ungeneralized k-means :

To cluster naturally imbalanced clusters like the ones shown in Figure 1, you can adapt (generalize) k-means. In Figure 2, the lines show the cluster boundaries after generalizing k-means as:

Left plot: No generalization, resulting in a non-intuitive cluster boundary.

Center plot: Allow different cluster widths, resulting in more intuitive clusters of different sizes.

Right plot: Besides different cluster widths, allow different widths per dimension, resulting in elliptical instead of spherical clusters, improving the result.

14. Is K means a deterministic algorithm?

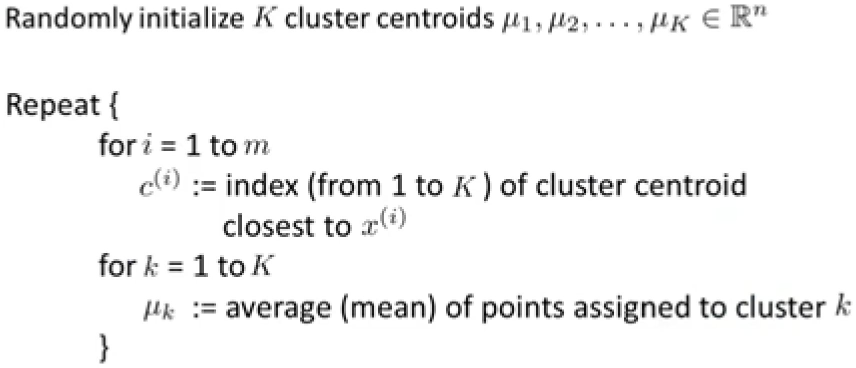
- K-Means is a non-deterministic algorithm. This means that a compiler cannot solve the problem in polynomial time and doesn’t clearly know the next step. This is because some problems have a great degree of randomness to them. These algorithms usually have 2 steps

  1)Guessing step

  2)Assignment step.

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.

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$



```
Randomly initialize $K$ cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$

Repeat {
        for $i$ = 1 to $m$
            $c^{(i)}$ := index (from 1 to $K$ ) of cluster centroid
                    closest to $x^{(i)}$
        for $k$ = 1 to $K$
            $\mu_k$ := average (mean) of points assigned to cluster $k$
        }
```