Machine-Learning Assignment

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

Solution : a) 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

Solution : d) 1,2 and 4

3. Can decision trees be used for performing clustering?

a) True b) False

Solution : a) 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

Solution : a) 1 only

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

a) 0 b) 1 c) 2 d) 3

Solution : b) 1

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

a) Yes b) No

Solution : a) Yes

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

Solution : 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 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

Solution : d) 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

Solution : 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.

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

Solution : 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

Solution : d) All of the above.

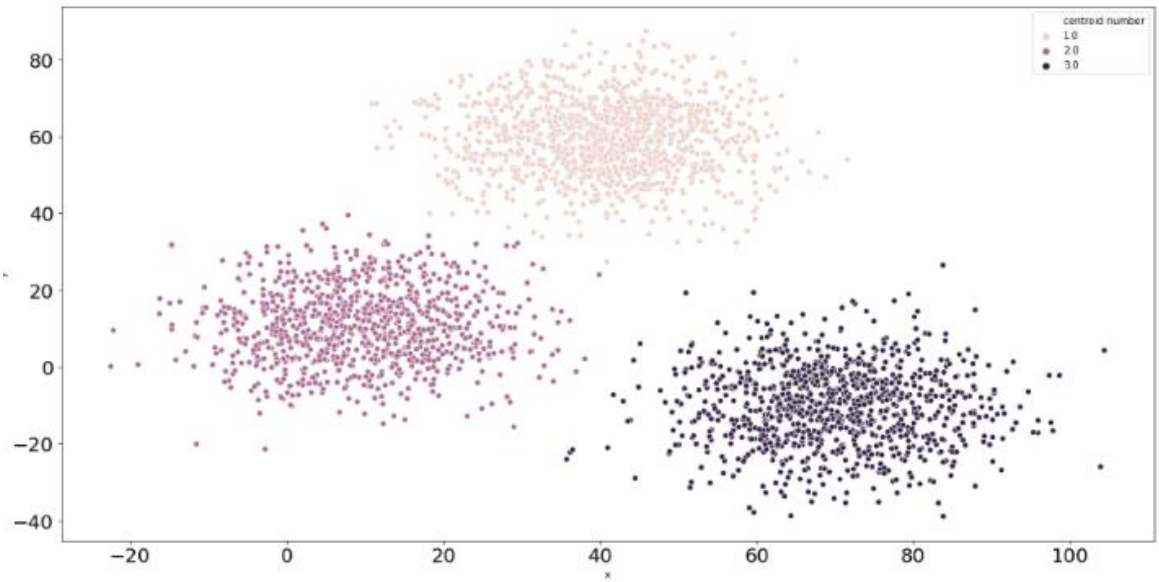
12. Is K sensitive to outliers?

Solution : K-Means clustering is an unsupervised learning algorithm which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest centroid. The algorithm aims to minimize the squared Euclidean distances between the observation and the centroid of cluster to which it belongs.

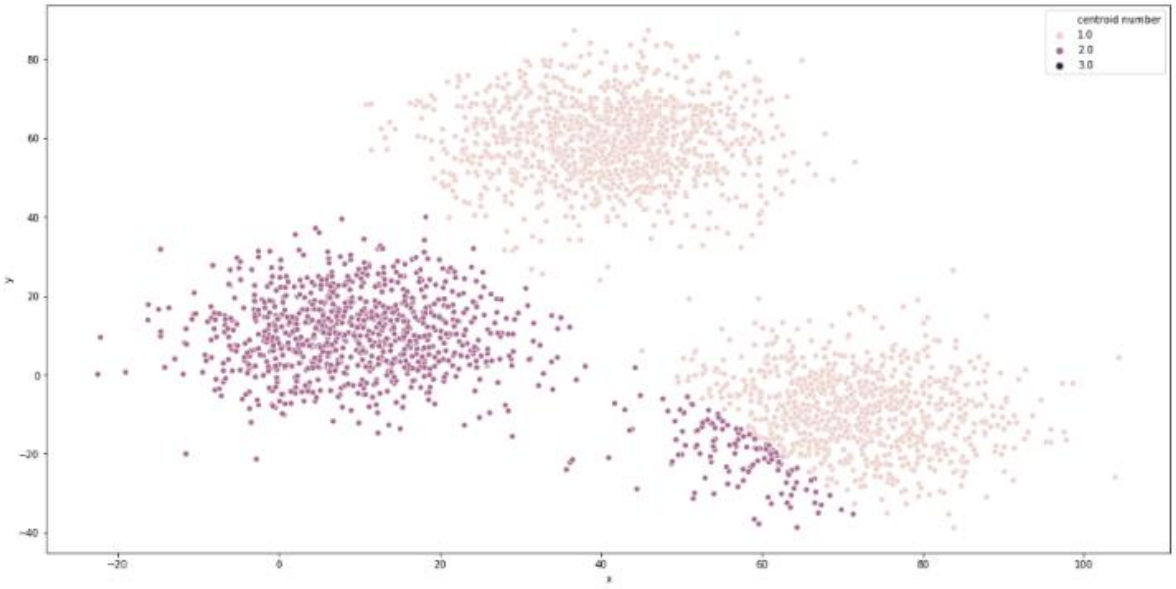
But sometime K-Means algorithm does not give best results. It is sensitive to outliers. An outlier is a point which is different from the rest of data points. So If one applies K-Means before removing the outliers then, they tend to form a separate cluster and sometime they get merged with non-outliers and which is cause of less of accuracy sometimes.

Hence it is better to identify and remove outliers before applying K-means clustering algorithm.

Scatter-plot when K-Means is applied after removal of outliers:



b) Scatter-plot when K-Means is applied before removal of outliers:



13. Why is K means better?

Solution :K-means clustering is a very famous and powerful unsupervised machine learning algorithm. It is used to solve many complex unsupervised machine learning problems.

It is very simple to implement.

It is scalable to a huge data set and also faster to large datasets.

it adapts the new examples very frequently.

Generalization of clusters for different shapes and sizes.

14. Is K means a deterministic algorithm?

Solution : 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.

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.