FLIP ROBO

ASSIGNMENT – 2

MACHINE LEARNING

Q1 to Q11 have only one correct answer. Choose the correct option to answer your question.

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 Only

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) 1, 2 and 4

3. Can decision trees be used for performing clustering?

a. True
b. False

Ans: 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

Ans: a) 1 Only

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

   Ans: b) 1

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

   Ans: b) 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

   Ans: 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

   Ans: 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

Ans: 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

    Ans: d) All of the above

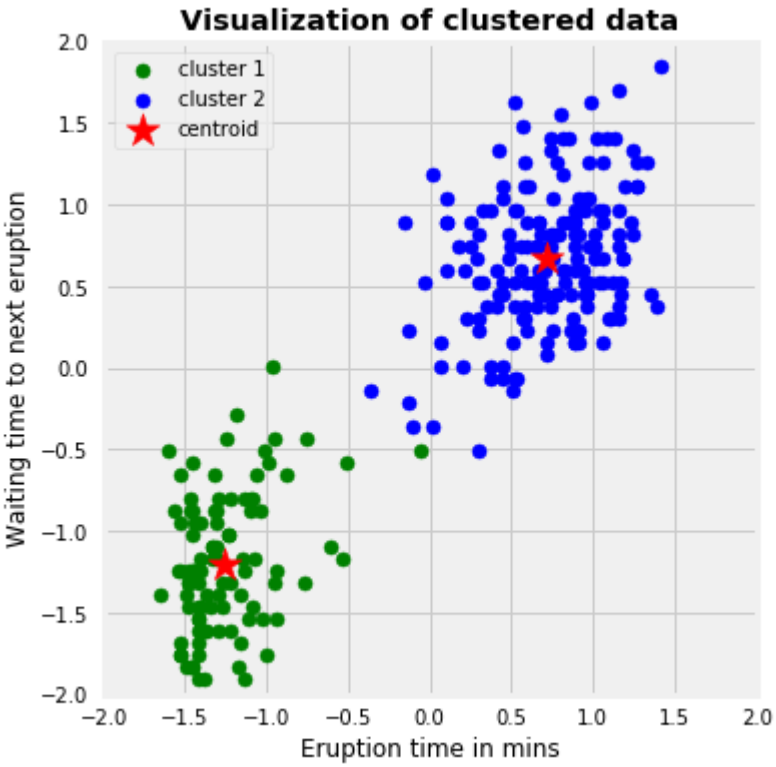
11. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?
    a. Proximity function used
    b. of data points used
    c. of variables used
    d. All of the above

    Ans: d) All of the above

Q12 to Q14 are subjective answers type questions, Answers them in their own words briefly

12. Is K sensitive to outliers?

- K-means Clustering is sensitive to outliers because the mean can be easily influence by outliers.
- **K-means** algorithm is an iterative algorithm that tries to partition the dataset into *K* pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.
- The approach k-means follows to solve the problem is called **Expectation-Maximization**.



There are two main types of clustering — **K-means Clustering** and **Hierarchical Agglomerative Clustering**.

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabelled dataset into different clusters.

13. **Why is K means better?**
    Below are the advantage of K means.

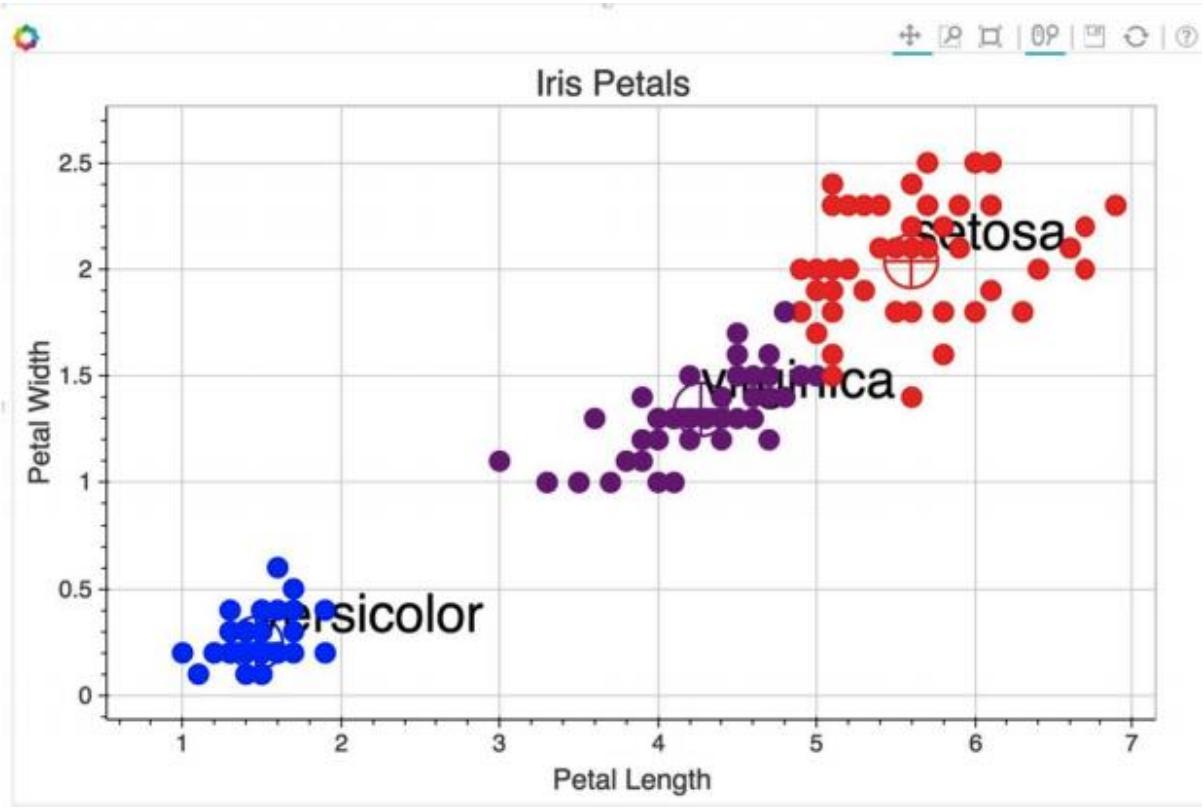
- Relatively simple to implement.
- Scales to large data sets.
- Guarantees convergence.
- Can warm-start the positions of centroids.
- Easily adapts to new examples.
- Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

- Unlabelled Data Sets
  A lot of real-world data comes unlabelled, without any particular class. The benefit of using an algorithm like K-means clustering is that we often do not know how instances in a data set should be grouped.

- Nonlinearly Separable Data
  Using K-means clustering and converting the coordinate system below from Cartesian coordinates to Polar coordinates, we could use the information about the radius to create concentric clusters.

- Simplicity.
  Unsupervised learning algorithm that is easy to implement and can handle large data sets, K-means clustering is a good starting point.

- Availability
  Most of the popular machine learning packages contain an implementation of K-means clustering.



14.Is K means a deterministic algorithm?

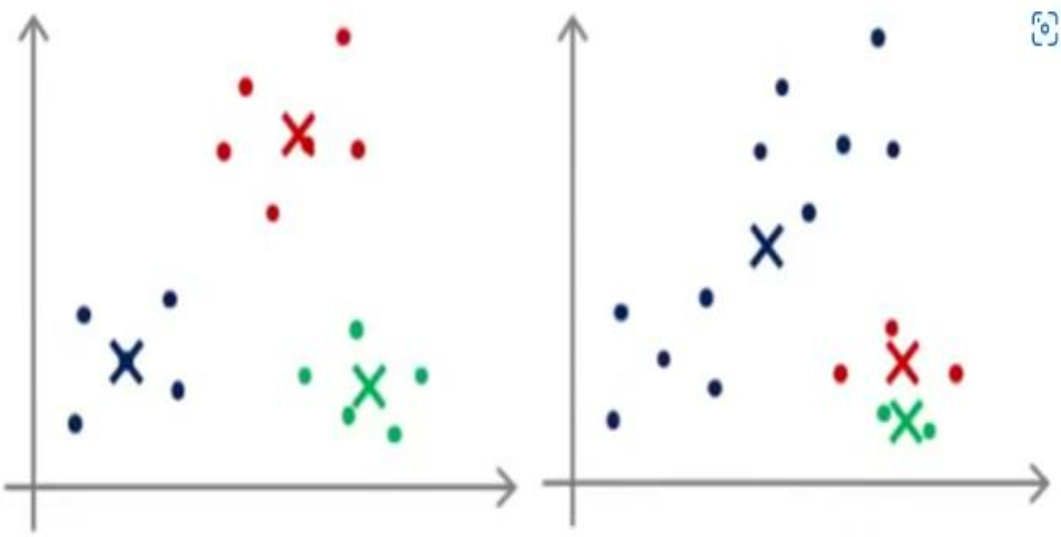
- K-Means is **a non-deterministic** algorithm. This means that running the algorithm several times on the same data, could give different results.
- Compiler cannot solve the problem in polynomial time and doesn't clearly know the next step. This is because some problems have a great degree of randomness 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.
- Any deterministic algorithm it has 2 phases. Guessing phase: Randomly initializing k means in the data space (Mu(k)s).
- Mathematically, this step tries to minimize the within cluster variance. Hence, every point is now assigned a cluster.
- 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.



Two different results

- Clustering algorithms with steps involving randomness usually give different results on different executions for the same dataset.
- This non-deterministic nature of algorithms such as the K-Means clustering algorithm limits their applicability in areas such as cancer subtype prediction using gene expression data.
- It is hard to sensibly compare the results of such algorithms with those of other algorithms.
- The non-deterministic nature of K-Means is due to its random selection of data points as initial centroids.