MACHINE LEARNING ASSIGNMENT

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

Answer a) 2 only.
Explanation : If we visit imdb.com, we notice people in the same cluster are made similar recommendations.

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 d) 1,2 and 4
Explanation :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.

Another way of looking at sentiment analysis is to consider it using a reinforcement learning perspective where the algorithm constantly learns from the accuracy of past sentiment analysis performed to improve the future performance.

3. Can decision trees be used for performing clustering?
a) True
b) False
Answer a) True
Explanation :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
Options:
a) 1 only
b) 2 only
c) 1 and 2
d) None of the above

Answer a) 1 only
Explanation : 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?
a) 0
b) 1
c) 2
d) 3

Answer b) 1
Explanation : 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?
a) Yes
b) No

Answer b) No
Explanation : 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?
a) Yes
b) No
c) Can't say
d) None of these

Answer A) Yes.
Explanation : 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 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 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

Answer a)K-means clustering algorithm
Explanation: 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.
Options:
a) 1 only
b) 2 only

c) 3 and 4
d) All of the above

Answer d) All of the above.
Explanation : 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?
a) Proximity function used
b) of data points used
c) of variables used
d) All of the above

Answer d) All of the above.
Explanation : Change in either of Proximity function, no. of data points or no. of variables will lead to different clustering results and hence different dendrograms.

Q12 to Q14 are subjective answers type questions, Answers them in their own words briefly
12. Is K sensitive to outliers?
13. Why is K means better?
14. Is K means a deterministic algorithm

Note: Sample working of K-Means is also attached in Worksheet2 folder

12. Is K sensitive to outliers?

Answer 12) K-Means Clustering is most widely used Unsupervised Machine Learning Algorithm.

Unsupervised Machine learning is nothing but clustering techniques. Clustering in laymen term is group(s) of data.

K-Means is a clustering approach in which the data is grouped into K distinct non-overlapping clusters based on their distances from K centers. The value of K needs to be specified first and then the algorithm assigns the points to exactly one clusters.

In the k-means based outlier detection technique, the data are partitioned into k groups by assigning them to the closest cluster centers.

Once assigned we can compute the distance or dissimilarity between each object and its cluster center,

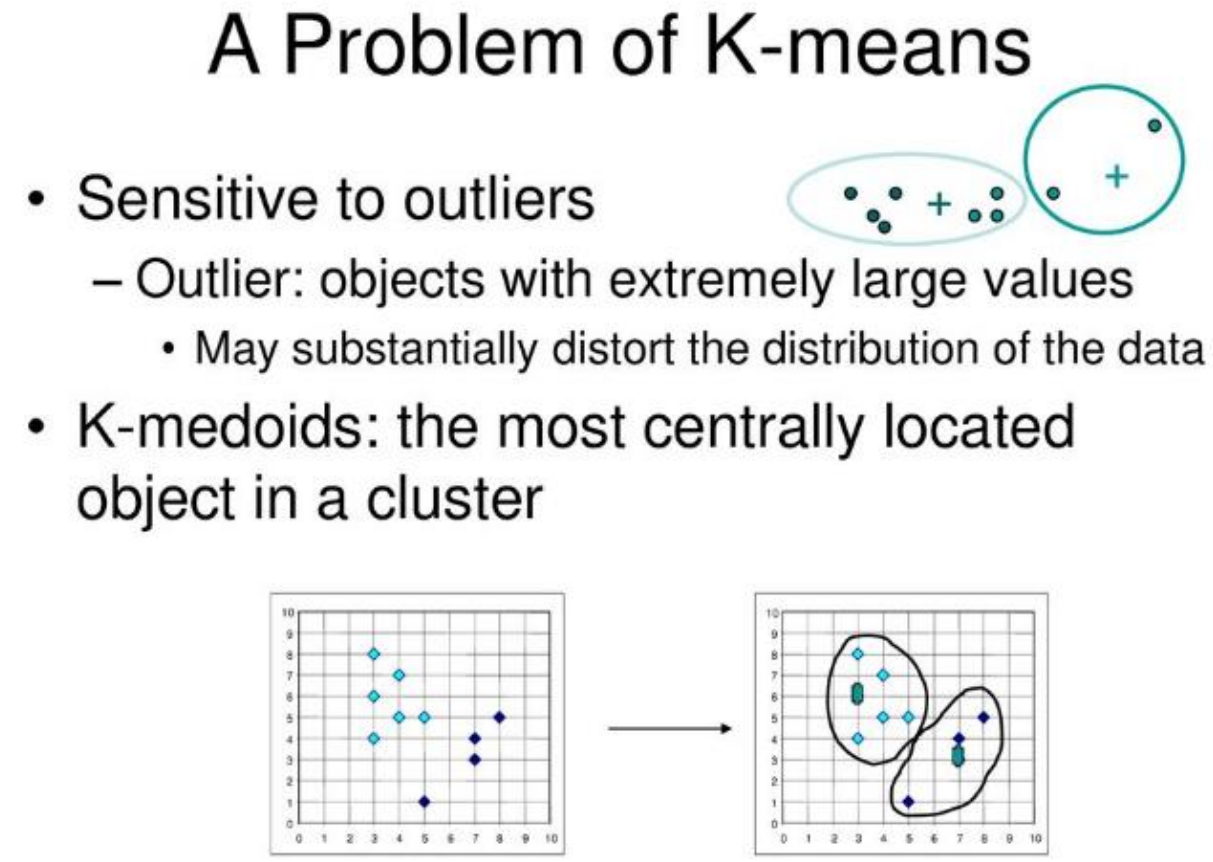
and pick those with largest distances as outliers.

Example We have the below set of value stored in x

[1,2,3,4,100]

mean(x): 22
median(x):3
mode(x):1
So it is sensitive to outliers.

But K-Means is sensitive to Outliers.



13. Why is K means better?

k-means is one of the simplest algorithm which uses unsupervised learning method to solve known clustering issues. Based on the research, it works really well with large datasets.
Other clustering algorithms with better features tend to be more expensive. In this case, k-means becomes a great solution for pre-clustering, reducing the space into disjoint smaller sub-spaces where other clustering algorithms can be applied.

K Means is a part of divisive approach. It first considers all points to be a part of one big cluster and in the subsequent steps tries to find out the points/clusters which are least similar to each other and then breaks the bigger cluster into smaller ones. This continues until there are as many clusters as there are data points. This is also called top-down approach.

K-means follows the below Steps:

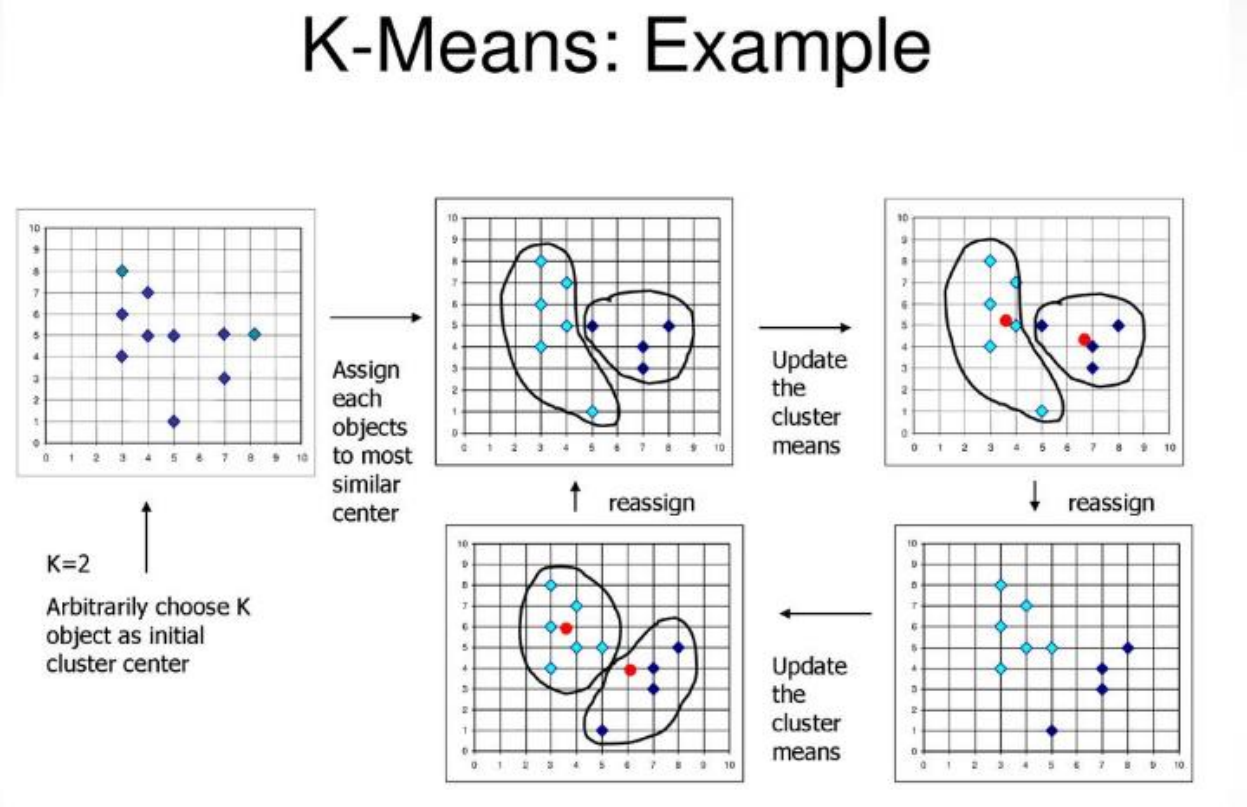
Step1 K-means uses Euclidian distance to find the distance between Centroid and corresponding data points

Step2 Then we get mean distance in that cluster (Add all distance divided by the number of data points in that cluster)

Step3 Again it calculates the mean distance between Centroid and corresponding data points

Step4 The process continues until there is no change in Average distance or there is no further movement of Centroids

Step5 Then we can call it as final cluster



Here are the **advantages**:

Unlabeled Data Sets

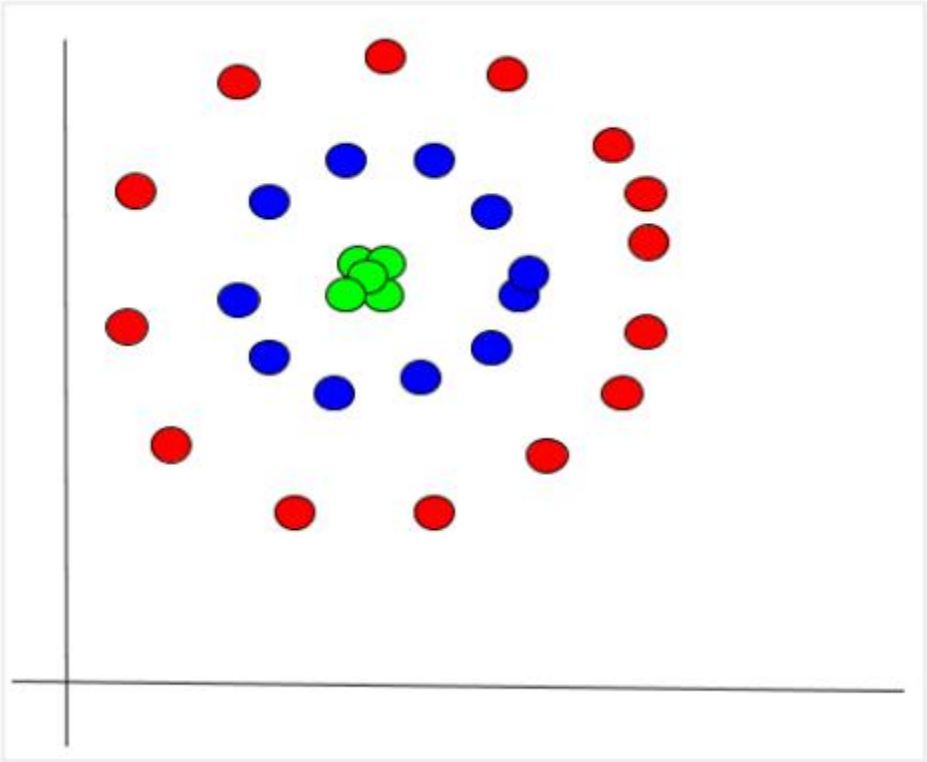
A lot of real-world data comes unlabeled, without any particular class. The benefit of using an algorithm like K-means clustering is that we often do not know how instances in a data set should be grouped.

For example, consider the problem of trying to group viewers of Netflix into clusters based on similar viewing behavior. We know that there are clusters, but we do not know what those clusters are. Linear models will not help us at all with these sorts of issues.

Nonlinearly Separable Data

Consider the data set below containing a set of three concentric circles. It is nonlinearly separable. In other words, there is no straight line or plane that we could draw on the graph below that can easily discriminate the colored classes red, blue, and green. 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.

Below figure illustrates the same.



Simplicity

The meat of the K-means clustering algorithm is just two steps, the **cluster assignment** step and the **move centroid** step. If we're looking for an 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.

Speed

Based on my experience using K-means clustering, the algorithm does its work quickly, even for really big data sets.

14. Is K means 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.

The *K*-means clustering algorithm uses iterative refinement to produce a final result. The algorithm inputs are the number of clusters *K* and the data set. The data set is a collection of features for each data point. The algorithms start with initial estimates for the *K* centroids, which can either be randomly generated or randomly selected from the data set. The algorithm then iterates between two steps:

1) Data assignment step

2) Centroid update step

Choosing K

The algorithm described above finds the clusters and data set labels for a particular pre-chosen K. To find the number of clusters in the data, the user needs to run the K-means clustering algorithm for a range of K values and compare the results. In general, there is no method for determining exact value of K, but an accurate estimate can be obtained using the following techniques.

One of the metrics that is commonly used to compare results across different values of K is the mean distance between data points and their cluster centroid. Since increasing the number of clusters will always reduce the distance to data points, increasing K will always decrease this metric, to the extreme of reaching zero when K is the same as the number of data points. Thus, this metric cannot be used as the sole target. Instead, mean distance to the centroid as a function of K is plotted and the "elbow point," where the rate of decrease sharply shifts, can be used to roughly determine K.

A number of other techniques exist for validating K, including cross-validation, information criteria, the information theoretic jump method, the silhouette method, and the G-means algorithm. In addition,

monitoring the distribution of data points across groups provides insight into how the algorithm is splitting the data for each K.