**MACHINE LEARNING**

1. Movie Recommendation system are an example of:
2. Classification
3. Clustering
4. Regression

Ans:- A) 2 Only

1. Sentiment Analysis is an example of:
2. Regression
3. Classification
4. Clustering
5. Reinforcement

Ans:- D) 1,2 and4

1. Can decision trees be used for performing clustering?

Ans:- A) True

1. Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points:
2. Capping and flooring of variables
3. Removal of outliers

Ans:- A) 1 only

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

Ans:- B) 1

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

Ans:- B) No

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

Ans:- A) Yes

1. Which of the following can act as possible termination conditions in K-Means?
2. For a fixed number of iterations.
3. Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.
4. Centroids do not change between successive iterations.
5. Terminate when RSS falls below a threshold.

Ans:- D) All of the above

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

Ans:- A) K-means clustering algorithm

1. How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):
2. Creating different models for different cluster groups.
3. Creating an input feature for cluster ids as an ordinal variable.
4. Creating an input feature for cluster centroids as a continuous variable.
5. Creating an input feature for cluster size as a continuous variable.

Ans:- D) All of the above

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

Ans:- D) All of the above

1. 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 noise and outliers. Instead of using the mean point as the center of a cluster, K-medoids uses an actual point in the cluster to represent it. Medoid is the most centrally located object of the cluster with minimum sum of distances to other points. The group of points in the right form a cluster while the rightmost point is an outlier. Mean is greatly influenced by the outlier and thus cannot represent the correct cluster center while medoid is robust to the outlier and correctly represents the cluster center.

1. Why is K means better?

Ans:-

* Relatively simple to implement.
* Scales to large data sets.
* Guarantees convergence.
* Can warm – Strat the positions of centrods.
* Easily adapts to new example.
* Generalizes to clusters of different shapes and sizes. Such as elliptical clusters.

1. Is K means a deterministic algorithm?

Ans:-

K-means is a non- deterministic algorithm. This means that a comiler 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
2. Assignment

* One 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. Now all the data points x(i)s (1,m) are assigned to clusters in accordance to which cluster means 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 are now assigned to the mean of the data points in the cluster. This step is repeated a couple of times.