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

1. Movie Recommendation systems are an example of:
   1. Classification
   2. Clustering
   3. Regression Options:
      1. 2 Only
      2. 1 and 2
      3. 1 and 3
      4. 2 and 3

**Answer** - d

1. Sentiment Analysis is an example of:
   1. Regression
   2. Classification
   3. Clustering
   4. Reinforcement Options:
      1. 1 Only
      2. 1 and 2
      3. 1 and 3
      4. 1, 2 and 4

**Answer** - d

1. Can decision trees be used for performing clustering?
2. True
3. False

**Answer** - a

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:
   1. Capping and flooring of variables
   2. Removal of outliers Options:
      1. 1 only
      2. 2 only
      3. 1 and 2
      4. None of the above

**Answer** - a

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

a) 0

b) 1

c) 2

d) 3

**Answer** -b

1. For two runs of K-Mean clustering is it expected to get same clustering results?
2. Yes
3. No

**Answer** -b

1. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means?
2. Yes
3. No
4. Can't say
5. None of these

**Answer** - a

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

**Answer** - d

1. Which of the following algorithms is most sensitive to outliers?
2. K-means clustering algorithm
3. K-medians clustering algorithm
4. K-modes clustering algorithm
5. K-medoids clustering algorithm

**Answer** - a

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

**Answer** - d

1. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?
2. Proximity function used
3. of data points used
4. of variables used
5. All of the above

**Answer** - d

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

1. Is K sensitive to outliers?
2. Why is K means better?
3. Is K means a deterministic algorithm?

**13) Answer** – 1) K-means starts by selecting k random data points as the initial set of centroids, which is then improve by two subsequent steps.in the assignment steps,every point is put into the cluster of the nearest centroid.

2) When the data has overlapping clusters k-means can improve the results of the initialization technique.

3) when the data has well separated clusters,the performs of the k-means depends completely on goodness of the initialization

4) initialization usig simple furthest point heuristic reduces the clustering error of k-means from 15% to 6% , on average

**12) Answer** – yes, the means algorithm update the cluster centers by taking the average of all the data points thata are closer to each cluster center, when all the points are packd nicely together, the averge make sence. However, when you have outliers, this can affect the average calculations of whole cluster, as results this will push your cluster centers closer to the outlier. The main disadvantage of K-mean algoritham is that it only gives ua convex clusters

**14) Answer** – This is 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 algoriths.