MACHINE LEARNING

ASSIGNMENT – 2

ANSWER SHEET

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. Sentiment Analysis is an example of:

i) Regression

ii) Classification

iii) Clustering

iv) Reinforcement

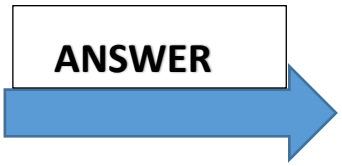
Options:

a) 1 Only

b) 1 and 2

c) 1 and 3

d) 1, 2 and 4

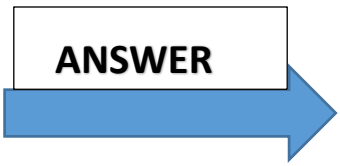


(D)

3. Can decision trees be used for performing clustering?

a) True

b) False



(A)

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

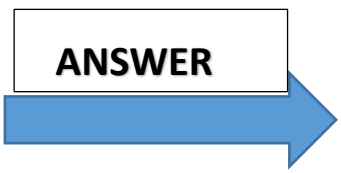
ii) Removal of outliers

Options:

a) 1 only

b) 2 only

c) 1 and 2



(C)

d) None of the above

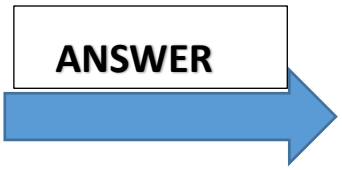
5. What is the minimum no. of variables/ features required to perform clustering?

a) 0

b) 1

c) 2

d) 3

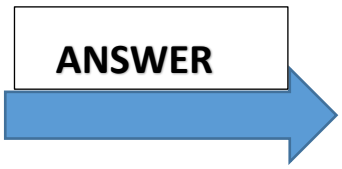


(B)

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

a) Yes

b) No

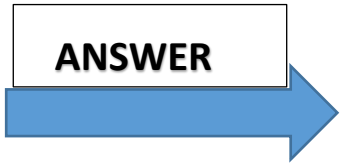


(B)

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 **(C)**

d) None of these

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.

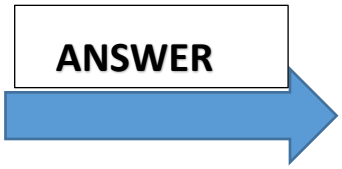
Options:

a) 1, 3 and 4

b) 1, 2 and 3

c) 1, 2 and 4

d) All of the above



(D)

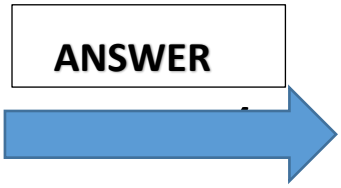
9. Which of the following algorithms is most sensitive to outliers?

a) K-means clustering algorithm

b) K-medians clustering algorithm

c) K-modes clustering algorithm

d) K-medoids clustering algorithm



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.

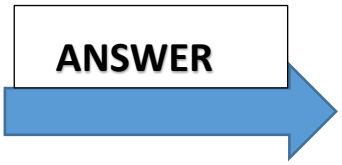
Options:

a) 1 only

b) 2 only

c) 3 and 4

d) All of the above



(D)

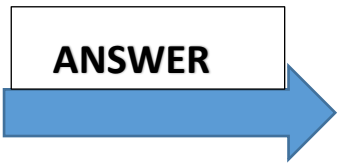
11. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?

a) Proximity function used

b) of data points used

c) of variables used

d) All of the above



(C)

12. Is K sensitive to outliers?

Ans - 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 that is greater than (Q3 + 1.5*IQR) or lesser than (Q1–1.5*IQR).

13. Why is K means better?

Ans - k-means is one of the simplest algorithm which uses unsupervised learning method to solve known clustering issues. It works really well with large datasets. It has some features where we can say the K-means is better than other algorithms-

1) Relatively simple to implement.

2) Scales to large data sets.

3) Guarantees convergence.

4) Can warm-start the positions of centroids.

5) Easily adapts to new examples.

6) Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

However, there are also drawbacks of K-Means which are:

- Strong sensitivity to outliers and noise.

- Doesn't work well with non-circular cluster shape -- number of cluster and initial seed value need to be specified beforehand.
- Low capability to pass the local optimum.

14. Is K means a deterministic algorithm?

Ans - No, K-means is not a deterministic algorithm. The random initialization step causes the k-means algorithm to be nondeterministic, meaning that cluster assignments will vary if you run the same algorithm twice on the same dataset. Researchers commonly run several initializations of the entire k-means algorithm and choose the cluster assignments from the initialization with the lowest SSE(sum of suqare error).