MACHINE LEARNING

WORKSHEET (CLUSTERING)

Q1 to Q12 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. Reinforcement Learning 4. Regression

Options:

a. 2 Only

b. 1 and 2

c. 1 and 3

d. 2 and 3

e. 1, 2 and 3

f. 1, 2, 3 and 4

2. Sentiment Analysis is an example of:

1. Regression 2. Classification 3. Clustering 4. Reinforcement Learning

Options:

a. 1 Only

b. 1 and 2

c. 1 and 3

d. 1, 2 and 3

e. 1, 2 and 4

f. 1, 2, 3 and 4

3. Can decision trees be used for performing clustering?

a. True

b. False

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:

a. Capping and flooring of variables b. Removal of outliers

Options:

a. 1 only

b. 2 only

c. 1 and 2

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

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

a. Yes 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

8. 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 with a bad local minimum.

3. Centroids do not change between successive iterations.

4. 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

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?

1. K- Means clustering algorithm

2. Agglomerative clustering algorithm

3. Expectation-Maximization clustering algorithm

4. Diverse clustering algorithm

Options:

a. 1 only

b. 2 and 3

c. 2 and 4

d. 1 and 3

e. 1,2 and 4

f. All of the above

10. 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

WORKSHEET

11. 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:

a. 1 only

b. 1 and 2

c. 1 and 4

d. 3 only

e. 2 and 4

f. All of the above

12. 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. B and c only

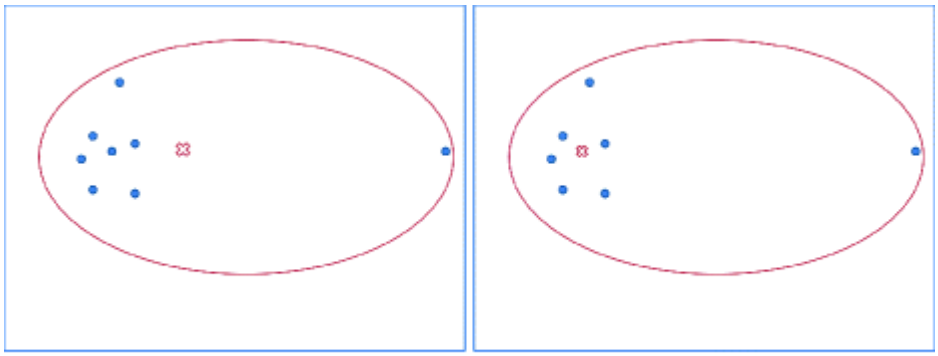
e. All of the above

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly

13. Is K sensitive to outliers?

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 noises and outliers. Instead of using the mean point as the centre 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.



(a) Mean (b) Medoid

14. Why is K means better?

Many argue that in the field of data science, one should primarily use simple, self-learning algorithms. And clustering algorithm, the most commonly used unsupervised learning algorithm is self-improving and one doesn't need to set parameters. In fact, most data science teams rely on simple algorithms like regression and completely because they solved all normal business problems with simple algorithms like XG Boost.

Another key upside of K-means, the standard data mining tool is that as opposed to conventional statistical methods, the clustering algorithms do not depend on statistical distributions of data and can be used with little prior knowledge. In fact, a lot of k-means applications are now done using support vector machines.

- It gives good results
- It is already implemented in the software
- Number of clusters has to be fixed before
- Dependent of the initialisation parameters and the chosen distance

15. Is K means a deterministic algorithm?

A deterministic algorithm is that in which output does not change on different runs. K means would give the different result if we run again and hence we can say that K means is not a deterministic algorithm. K-means is not deterministic and it also consists of number of iterations. The **non-deterministic** nature of **K-Means** is due to its random selection of data points as initial centroids.