MACHINE LEARNING

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

Ans -> Clustering

2. Sentiment Analysis is an example of:

i) Regression

ii) Classification

iii) Clustering

iv) Reinforcement

Ans - > d) 1, 2 and 4

3. Can decision trees be used for performing clustering?

a) True

b) False

Ans- > True

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

Ans -> i) Capping and flooring of variables

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

Ans-> b) 1

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

Ans -> No

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

Ans-> Yes

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 witha bad local minimum.

iii) Centroids do not change between successive iterations.

iv) Terminate when RSS falls below a threshold.

Ans -> All of the above

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

Ans - a) K-means 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.

Ans -> . d) All of the above

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

Ans -> a) Proximity function used

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

12. Is K sensitive to outliers?

Ans -> Yes , K is vey much sensitive to outliers . Because ,

First See With Example then understand with pure laymen's term. With figure.

EXAMPLE

Suppose ,

We created one array x=[1 2 3 4 100],

So, Here 100 is Outliers .

Then Calculate the Statistics of the x .

We get ,

Mean -> 22

Median -> 3

Mode -> 2

Lets check which statistical parameter . which are huge effected by outliers,

Median -> 3 which are in the data points.

Mode ->-> 2 which are in the data points.

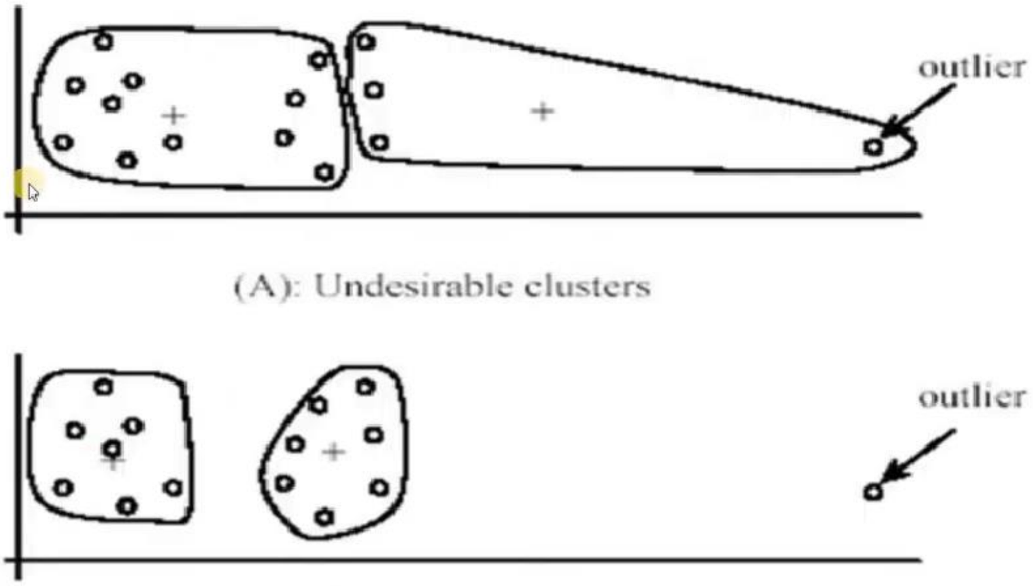
Mead -> Mean is 22 is not representing any data points from our dataset. Either 1,2,3,4 is very small for our mean and 100 is very high from our mean.

So Basically , I want to show here . outliers are how much effect on our mean. .and if mean is effected then our then our accuracy will also effected.

A- Undesirable Cluster -> If we have a dataset with outliers . Then our prediction must be go wrong . Because outliers are creating too much variance between groups. .

  If we see in this figure (A) Undesirable Cluster Then we clearly see there are 2 groups of data and 1 outlier . and one outliers can how much effect on the clustering of the data . We easily create two groups

By removing outliers .Lets See on the figure (B).



(A): Undesirable clusters

(B): Ideal clusters

Lets See on the figure (B).

If we can see in the figure (B). They show if we ignore the outliers then we easily create 2 groups and make our accuracy good .So I want to show you how outliers can effect K.

13. Why is K means better?

Ans -> 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 whereother clustering algorithms can be applied. K-means is the simplest.

14. Is K means a deterministic algorithm ?

Ans -> . The basic k-means clustering is based on a non-deterministic algorithm. This means that running thealgorithm several times on the same data, could give different results. However, to ensure consistentresults, FCS Express performs k-means clustering using a deterministic method