MACHINE LEARNING – WORKSHEET SET 2

FLIP ROBO TECHNOLOGIES : INTERNSHIP 32

SUBMITTED BY : SHAHLA M

1. Movie Recommendation systems are an example of:

i) Classification

ii) Clustering

iii) Regression

Ans: b) 1 and 2

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?

Ans: a) 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: a) 1 only

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: b) No

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

Ans: a) 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 with a bad local minimum.

iii) Centroids do not change between successive iterations.

iv) Terminate when RSS falls below a threshold.

Ans: d) 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 g (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

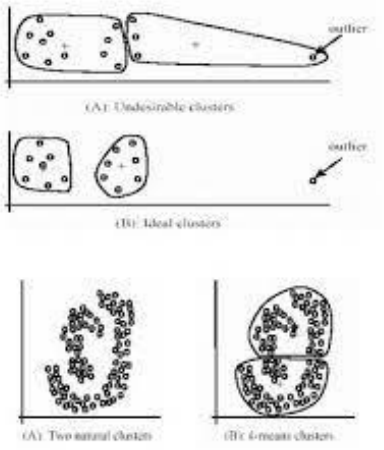
11. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?

Ans: d) All of the above

12. Is K sensitive to outliers?

Ans : Yes. K means is sensitive to outliers. Sometime K-Means algorithm does not give best results. It is sensitive to outliers. An outlier is a point which is different from the rest of data points.

The k-means algorithm updates the cluster centers by taking the average of all the data points that are closer to each cluster centre. When all the points are packed nicely together, the average makes sense. However, when you have outliers, this can affect the average calculation of the whole cluster. As a result, this will push your cluster centre closer to the outlier.



Here we can see the deviation of mean point for the presence of outliers. In the above figure, fig (B) is the actual clusters without considering outlier. But, after considering outlier, it will push the original cluster centre closer to the outlier and fig (A) is generated. To counter this we use algorithms like K-medoids

13. Why is K means better?

Ans:

It's ideal to choose K-means when you have no idea on what basis you are classifying the data. Since k-means is an unsupervised learning algorithm it doest have any attribute based on which it will learn to classify, rather it all group all similar data points and form clusters. It is efficient in terms of computing than rest of the algorithms which have better features. We can ensure definite converge using this algorithm. It is also simple and highly flexible. It is easy to explain the results in contrast to Neural Networks.

14. Is K means a deterministic algorithm?

Ans: No. K means is a non-deterministic algorithm. Deterministic algorithms are a type of algorithms which gives the similar outputs after every time execution on same data. In other hand, non-deterministic algorithms are a type of algorithms which gives different results on same data after every time execution. Actually, every time it randomly selected the data points as initial centroids. This random selection influenced the final result and each run of the algorithm for the same dataset may give different output.