Flip Robo Technologies | Internship- 31

-: MACHINE LEARNING WORKSHEET- 2 :-

Submitted By: *TAMALI SAHA (tamali428@gmail.com)*

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 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 (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.

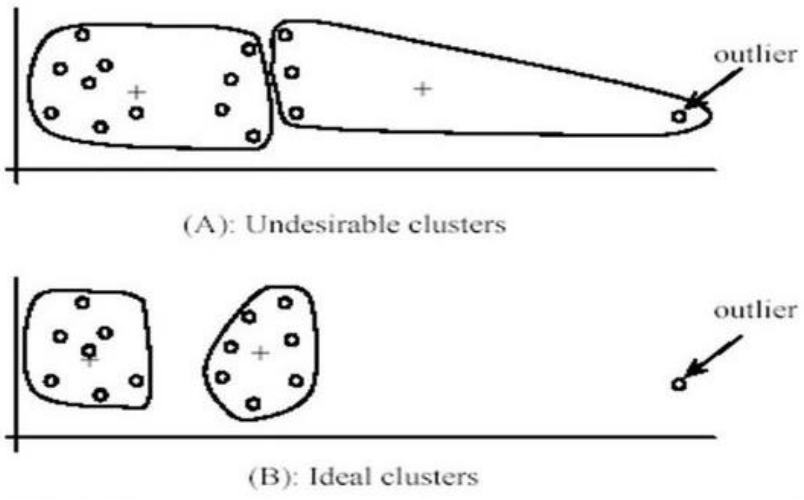
Ans: d) All of the above

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. It uses the average of cluster data points to find the cluster centre of the given dataset. For a given set of observation, K means algorithm partitioning the observations into certain number of clusters (say k). Here each observations are belongs to a cluster with the nearest centroid. Now outliers is such a data point which deviates significantly from the rest of the data points (whole observations). Now outliers exists, it actually influence the mean by increasing or decreasing the value of the mean. It also can skew our actual cluster to very large extent.



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: K-means is an unsupervised learning algorithm. 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.