Worksheet Set 2

Machine Learning Worksheet 2

Question 1. (B) 1 and 2

Question 2. (D) 1,2 and 4

Question 3. (A) True

Question 4. (A) Capping and flooring of variables

Question 5. (B) 1

Question 6. (B) No

Question 7. (A) Yes

Question 8. (D) All of the above

Question 9. (A) K-means clustering algorithm

Question 10. (D) All of the above

Question 11. (D) All of the above

Question 12. Is K sensitive to outliers?

Answer 12. Yes, K is sensitive to outliers. K means clustering is an unsupervised learning algorithm which aims to partition 'n' observations into 'k' clusters in which each observation belongs to the cluster with the nearest centroid. The algorithm aims to minimize the squared Euclidean distances between the observation and the centroid of cluster to which it belongs. But sometimes K-Means algorithm does not give best results because it is sensitive to outliers. An outlier is a point which is far away from the rest of data points. We can observe that the outlier increases the mean of data. Hence the mean is influenced by outliers. Since K-Means algorithm is about finding mean of clusters, the algorithm is also influenced by outliers. Hence it is better to identify and remove outliers before applying K-means clustering algorithm.

Question 13. Why is K means better?

Answer 13. Some of the common clustering algorithms are Hierarchical clustering, Gaussian mixture models and K-means clustering.

The K-means clustering is considered a better unsupervised learning algorithm, wherein data is split into ‘k’ distinct clusters based on distance to the centroid of a cluster. The algorithm clusters into ‘k’ groups and here ‘k’ is the input parameter.

Reasons Why It is better:

1. K-means is very simple and easy to implement.

2. It scales to large data sets**.**

3. It guarantees convergence**.**

4. It easily adapts to new examples.

5. It generalizes to clusters of different shapes and sizes such as elliptical clusters.

6. It is one of the most robust methods, especially for image segmentation and image annotation projects.

Question 14. Is K means a deterministic algorithm?

Answer 14. K-Means is non-deterministic in nature.

K-Means starts with a random set of data points as initial centroids. This random selection influences the quality of the resulting clusters. Besides, each run of the algorithm for the same dataset may yield a different output.