

WORKSHEET 2

Machine Learning(Answer Sheet)

Ans. 1. (A) Clustering

Ans. 2. (D) Regression, Classification & Reinforcement

Ans. 3. (A) True

Ans. 4. (A) 1 only

Ans. 5. (B) 1

Ans. 6. (B) No

Ans. 7. (A) Yes

Ans. 8. (D) All of the above

Ans. 9. (A) K-means clustering algorithm

Ans. 10. (D) All of the above

Ans. 11. (D) All of the above

Ans. 12. Yes, K-means is sensitive to outliers. As we know that, K-means Algorithm is an Unsupervised Learning Algorithm, which groups the unlabeled dataset into different clusters. We also know that it is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distance between the data point and their corresponding clusters.

For an example:

Lets take a list of salary.

$X = [2000, 4000, 5000, 3000, 20000]$

Now we calculate Mean of salary,

Mean of $x = \text{sum of all items} / \text{total no of items}$

$$= (2000 + 4000 + 5000 + 3000 + 20000) / 5$$

$$= 34000 / 5$$

$$= 6800$$

Now we exclude 20000 from the list and calculate mean of salary again,

Mean of $x = 3500$

This shows that this mean is near to our values into the list, but on the above we get the value far from the data values.

That is why the K-Means is sensitive to the outliers.

Ans. 13. K-Means for clustering is one of the popular algorithm for Unsupervised data. K-means is a clustering algorithm whose main goal is to group similar elements or data points into a cluster. Here, 'K' in K-Means represent the number of clusters. It is better than other clustering algorithms like density-based, expectation-maximization. It is one of the most robust methods, especially for image segmentation and image annotation projects.

This algorithm makes clusters into K groups and here k is the input parameter. In this procedure, a data set is classified through a certain numbers of clusters, commonly known as k cluster and the main idea is to define k centre, one of each cluster. These centers should be placed in a way since different location causes different results. So, the better choice is to place them as much as possible far away from each other.

The next step is to take each point belonging to a given data set and associate it to the nearest centre. When no points is pending, the first step is completed and an early group age is done. However, the main disadvantage is one has to specify the number of clusters as an input in the algorithm.

Ans. 14. No, K-Means is a non-deterministic algorithm. K-Means starts with a random set of data points as initial centroids. This random selection influence the quality of the resulting clusters. Besides, each run of the

algorithm for the same data set may yield a different output.