

WORKSHEET-1

Machine Learning Assignment

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

Q1. (b) → 4

Q2. (d) → 1,2 &4

Q3. (d) → formulating the clustering problem

Q4. (a) → Euclidean distance

Q5. (b) → Divisive clustering

Q6. (d) → All answers are correct

Q7. (a) → Divide the data points into groups

Q8. (b) → Unsupervised learning

Q9. (d) → All of the above

Q10. (a) → K-means clustering algorithm

Q11. (d) → All of the above

Q12. (a) → Labelled data

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly.

Q13. How is cluster analysis calculated?

Ans. Steps for cluster analysis:

1.Choose the Right Variable – The concept involves identifying what is the right attribute and how much is it worth it. Here, one must select a variable that one feels may be important for identifying and understanding differences among groups of observation within the data.

2.Scaling the Data – In this, the data samples from different sources may be grouped in different scales. Here, the variables in the analysis vary in range; the variable with the largest range will have the greatest impact on the results.

3.Calculate Distances- Here, if the variables in the analysis vary in range, the variable with the largest range will have the greatest impact on the results.

Calculation of Distance between Points in a Cluster:

- 1) One way is that we can take the centre of the cluster and find out the centre of the next group and calculate distance between the centres.
- 2) Or take the closest point and find distance between closest points.
- 3) Or take the largest distance points and find out the distant between them.

Q14. How is cluster quality measured?

Ans. A cluster quality is measured through its performance, the capability to find out hidden patterns, and the similarity measure utilized by the methodology.

Q15. What is cluster analysis and its types?

Ans. Cluster analysis is a multivariate data mining technique whose goal is to groups objects based on a set of user selected characteristics or attributes. It is the basic and most important step of data mining and a common technique for statistical data analysis, and it is used in many fields such as data compression, machine learning, pattern recognition, information retrieval etc.

Types of Cluster analysis:

1. Hierarchical Cluster Analysis

In this method, first, a cluster is made and then added to another cluster (the most similar and closest one) to form one single cluster. This process is repeated until all subjects are in one cluster. This particular method is known as **Agglomerative method**. Agglomerative clustering starts with single objects and starts grouping them into clusters.

The divisive method is another kind of Hierarchical method in which clustering starts with the complete data set and then starts dividing into partitions.

2. Centroid-based Clustering

In this type of clustering, clusters are represented by a central entity, which may or may not be a part of the given data set. K-Means method of clustering is used in this method, where k are the cluster centres and objects are assigned to the nearest cluster centres.

3. Distribution Based clustering

It is a type of clustering model closely related to statistics based on the modals of distribution. Objects that belong to the same distribution are put into a single cluster. This type of clustering can capture some complex properties of objects like correlation and dependence between attributes.

4. Density Based Clustering

In this type of clustering, clusters are defined by the areas of density that are higher than the remaining of the data set. Objects in sparse areas are usually required to separate clusters. The objects in these sparse points are usually noise and border points in the graph. The most popular method in this type of clustering is DBSCAN.