MACHINE LEARNING – WORKSHEET-1

ANSWERS

1. B
2. C
3. D
4. A
5. B
6. D
7. A
8. B
9. A
10. A
11. D
12. A

13.How is cluster analysis calculated?

Ans:- **Clustering** is a process of segregating populations into a number of groups that are supposed to have similarity. Clustering analysis is a concept which is used to group data points which are supposed to exhibit similar patterns.It is used on a unsupervised data,that means that data which do not have labels.

Clustering helps to find natural groups in the feature space of input data.

There are following ways to perform clustering analysis:-

- Set an optimal number of cluster centroids k. These will be the total number of clusters formed in your dataset.
- For each value of calculate the total sum of square. This is done by taking the mean vector of points in that cluster until convergence and perform this for every cluster.
- The optimal number of clusters k could be calculated by ELBOW Method. The point where the bend is present in the plot is generally indicated as number of clusters.

  Other methods to calculate number of clusters can be Average silhouette method and the Gap statistic method.

14. How is cluster quality measured?

Ans:-There are few method for measuring the quality of cluster. These are grouped into two.

1)Extrinsic Methods:-In extrinsic method ,we take some algorithms and try to evaluate them in solving another problem. These includes:-

- **Clustering Homogeneity**: A clustering result satisfies homogeneity if all of its clusters contains only datapoints which are members of a single class.
- **Cluster Completeness:** A clustering result satisfies completeness if all the datapoints that a members of a given class are elements of the same cluster
- **Small cluster preservation:** splitting a small category into pieces is more harmful than spitting a large category in to pieces.

2)Intrinsic Methods:-In Intrinsic method, we try to see if clustering is useful in and of itself. These includes:-

- **Silhouette method**- it refer to interpretation and validation of consistency with in clusters of data. The silhouette value is a measure of how similar an object is to its own cluster(cohesion) compared to other clusters(separation).

The silhouette ranges from -1 to +1, which indicates how well is object is matched.

- **The Dunn Index(DI)-** its is a metric for judging a clustering algorithm. A higher DI implies better clustering which means clusters are compact and well-separated from other clusters.

15. What is cluster analysis and its types?

Ans:- Clustering analysis is a concept which is used to group data points which are supposed to exhibit similar patterns.It is used on a unsupervised data,that means that data which do not have labels.

Clustering helps to find natural groups in the feature space of input data.

TYPES OF CLUSTER ANALYIS ARE :-

I. **CENTRAL CLUSTERING:-** In central cluster ,we choose the number of clusters that we want to classify. The initial cluster centroid are selected randomly and after that it is assigned to the closest cluster centroid.
   k-means is the most widely-used centroid-based clustering algorithm. Centroid-based algorithms are efficient but sensitive to initial conditions and outliers.

II. **DENSITY CLUSTERING:-** Density clustering connects areas of high example density into clusters. This allows for arbitrary-shaped distributions as long as dense areas can be connected. These algorithms have difficulty with data of varying densities and high dimensions. Further, by design, these algorithms do not assign outliers to clusters. EX-DBSCAN Clustering

III. **DISTRIBUTION CLUSTERING:**-its identifies the probability that a point belongs to a cluster.Around eacg possible centroid the algorithm defines rhe density distributions for each cluster, quantifying the probability of belonging based on distribution .

This clustering approach assumes data is composed of distributions, such as Gaussian distributions.

IV. **HIERACHICAL CLUSTERING**:-It is a tree- based representation of the objects, which is also called as dendrogram. Observations can be subdivided into groups by cutting the dendrogram at a desired level.