MACHINE LEARNING ANSWERKEY-Worksheetset1

1. b)4
2. d)1,2 and 4
3. d) formulating the clustering problem
4. a) Euclidean distance
5. b) Divisive clustering
6. d) All answers are correct
7. a) Divide the data points into groups
8. b) Unsupervised learning
9. d) All of the above
10. a) K-means clustering algorithm
11. d) All of the above
12. a) Labeled data
13. How is cluster analysis calculated?

The objective of this algorithm is to partition a data set S consisting of n-tuples of real numbers into k **clusters** $C_1, \ldots, C_k$ in an efficient way. For each cluster C_j, one element c_j is chosen from that cluster called a **centroid**.

Definition 1: The basic **k-means clustering algorithm** is defined as follows:

- Step 1: Choose the number of clusters k
- Step 2: Make an initial selection of k centroids
- Step 3: Assign each data element to its nearest centroid (in this way k clusters are formed one for each centroid, where each cluster consists of all the data elements assigned to that centroid)
- Step 4: For each cluster make a new selection of its centroid
- Step 5: Go back to step 3, repeating the process until the centroids don't change (or some other convergence criterion is met)

There are various choices available for each step in the process.

An alternative version of the algorithm is as follows:

- Step 1: Choose the number of clusters k
- Step 2: Make an initial assignment of the data elements to the k clusters
- Step 3: For each cluster select its centroid

- Step 4: Based on centroids make a new assignment of data elements to the *k* clusters
- Step 5: Go back to step 3, repeating the process until the centroids don't change (or some other convergence criterion is met)

14. How is cluster quality measured?

Cluster quality can be measured using various metrics, such as:

Within-cluster sum of squares (WCSS) or within-cluster variance, which measures the similarity of data points within a cluster.

Silhouette Score, which measures how similar an object is to its own cluster compared to other clusters.

Davies-Bouldin index, which measures the average similarity between each cluster and its most similar cluster.

Calinski-Harabasz index, which compares the ratio of between-cluster variance to within-cluster variance.

Rand Index, which compares the similarity of the cluster labels to the true labels of the data points.

Mutual Information based scores, such as Normalized Mutual Information (NMI) and Adjusted Mutual Information (AMI)

15. What is cluster analysis and its types?

Cluster analysis is a multivariate data mining technique whose goal is to groups objects (eg., products, respondents, or other entities) based on a set of user selected characteristics or attributes.

A)Centroid-based Clustering: Algorithms such as k-means and k-medoids are examples of centroid-based clustering. They work by defining clusters around a central point, also known as a centroid.

B)Hierarchical Clustering: Algorithms such as Agglomerative and Divisive are examples of hierarchical clustering. They create a hierarchy of clusters, where each cluster is a subset of the previous one.

C)Density-based Clustering: Algorithms such as DBSCAN are examples of density-based clustering. They work by defining clusters as high-density regions of the data.

D)Distribution-based Clustering: Algorithms such as Gaussian Mixture Models (GMM) are examples of distribution-based clustering. They model the clusters as probability distributions and assign data points to clusters based on the likelihood of them being generated by the distribution.