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

13. Cluster Analysis calculation for k means:

- First, an initial partition with k clusters (given number of clusters) is created.
- Then, starting with the first object in the first cluster, Euclidean distances of all objects to all cluster foci are calculated.
- If an object is detected whose distance to the centre of gravity of the own cluster is greater than the distance to the centre of gravity (centroid) of another cluster, this object is shifted to the other cluster.
- Finally, the centroids of the two changed clusters are calculated again, since the compositions have changed here.
- These steps are repeated until each object is located in a cluster with the smallest distance to its centroid (centre of the cluster) (optimal solution)

Cluster Analysis calculation for BDSCAN:

- DBSCAN algorithm requires two parameters:

  eps: It defines the neighbourhood around a data point i.e. if the distance between two points is lower or equal to ‘eps’ then they are considered neighbours. If the eps value is chosen too small then a large part of the data will be considered as outliers. If it is chosen very large then the clusters will merge and the majority of the data points will be in the same clusters. One way to find the eps value is based on the k-distance graph.

MinPts: Minimum number of neighbors (data points) within eps radius.

In this algorithm, we have 3 types of data points.

1. Core Point: A point is a core point if it has more than MinPts points within eps.

2. Border Point: A point which has fewer than MinPts within eps but it is in the neighborhood of a core point.

3. Noise or outlier: A point which is not a core point or border point.

DBSCAN algorithm can be abstracted in the following steps:

1. Find all the neighbor points within eps and identify the core points or visited with more than MinPts neighbors.
2. For each core point if it is not already assigned to a cluster, create a new cluster.
3. Find recursively all its density connected points and assign them to the same cluster as the core point.
4. point a and b are said to be density connected if there exists a point c which has a sufficient number of points in its neighbors and both points a and b are within the eps distance. This is a chaining process. So, if b is neighbor of c, c is neighbor of d, d is neighbor of e, which in turn is neighbor of a implies that b is neighbor of a.
5. Iterate through the remaining unvisited points in the dataset. Those points that do not belong to any cluster are noise.

14. Most widely used method to evaluate the quality of a cluster is the Silhouette coefficient. The way the Silhouette coefficient works is it measures how similar is an object of one cluster compared to another cluster. Its values lie in the range of [-1:1] where a high positive integer indicates that all objects within the cluster are well matched and on another end if we get a negative value it indicates either the object is

not placed in a proper cluster or we have too many or too fewer clusters so data was not able to be clustered properly. Silhouette can be calculated using both Euclidean distance and Manhattan distance. To calculate the silhouette coefficient, cluster cohesion (a) and cluster separation (b) must be calculated. Cluster cohesion refers to the average distance between an instance (sample) and all other data points within the same cluster while cluster separation refers to the average distance between an instance (sample) and all other data points in the nearest cluster. The silhouette coefficient is essentially the difference between cluster separation and cohesion divided by the maximum of the two.

15. When we try to group a set of objects that have similar kind of characteristics and attributes these groups are called clusters. The process is called clustering. It is a very difficult task to get to know the properties of every individual object instead, it would be easy to group those similar objects and have a common structure of properties that the group follows. Cluster analysis is a multivariate data mining technique whose goal is to group objects based on a set of user-selected characteristics or attributes.

    Types of Cluster Analysis:

    1. Hierarchical Cluster Analysis

        1. Agglomerative method.

            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 the Agglomerative method.

        2. Divisive method.
            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 models 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.