**1.Explain the concept of hierarchical clustering and discuss its advantages and disadvantages.**

**2.What are the key differences between agglomerative and divisive hierarchical clustering?**

Agglomerative Clustering is a type of hierarchical clustering algorithm that merges the most similar pairs of data points or clusters, building a hierarchy of clusters until all the data points belong to a single cluster.\

Agglomerative clustering is generally more computationally expensive, especially for large datasets as this approach requires the calculation of all pairwise distances between data points.

It tends to produce more interpretable results

Applications are: Image segmentation, Customer segmentation, Social network analysis,

Divisive Clustering is the technique that starts with all data points in a single cluster and recursively splits the clusters into smaller sub-clusters based on their dissimilarity.

Comparatively less expensive as divisive clustering only requires the calculation of distances between sub-clusters,

It is more difficult to interpret.

Applications are :Market segmentation, Anomaly detection, Natural language processing, etc.

**3.Describe the dendrogram in hierarchical clustering. How is it useful in determining the number of clusters?**

A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters.

**4.Discuss the difference between single-linkage, complete-linkage, and average-linkage clustering methods in hierarchical clustering.**

Linkage criteria determine the distance between clusters and impact the structure of the dendrogram in hierarchical clustering. Common linkage criteria include single-linkage, complete-linkage, average-linkage, and ward's method.

Single linkage: computes the minimum distance between clusters before merging them. Complete linkage: computes the maximum distance between clusters before merging them. Average linkage: computes the average distance between clusters before merging them.

**5.Explain the concept of density-based clustering and how it differs from centroid-based clustering.**

Density-Based Clustering refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in a data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density.

Compared to centroid-based clustering like k-means, density-based clustering works by identifying “dense” clusters of points, allowing it to learn clusters of arbitrary shape and identify outliers in the data.

**6.What are the main parameters in a DBSCAN algorithm, and how do they affect the clustering results?**

**7.Compare and contrast DBSCAN with K-means clustering. In what scenarios is each algorithm more suitable?**

DBSCAN is particularly useful when dealing with datasets that have irregular shapes and different densities.

K-Means, is a centroid-based algorithm that partitions data into k clusters based on the mean distance between points and their assigned centroid

K-means is sensitive to noise and outliers since it uses centroids. A few outlying points can significantly shift the position of centroids, leading to suboptimal clusters. DBSCAN inherently identifies and separates noise from clusters.

**8.Explain the core points, border points, and noise points in the context of DBSCAN clustering.**

A point is a core point if it has more than a specified number of points (MinPts) within epsilon, indicating that it is in a dense region and likely part of a cluster. These are points that are at the interior of a cluster. A border point has fewer than MinPts within epsilon, but is in the neighbourhood of a core point. A noise point is any point that is not a core point nor a border point.

**9.Provide a step-by-step explanation of the DBSCAN clustering algorithm, including how it identifies core points and expands clusters.**

1. Classify the points.

2. Discard noise.

3. Assign cluster to a core point.

4. Color all the density connected points of a core point.

5. Color boundary points according to the nearest core point.

To define core points it must satisfy one condition that the number of neighbors must be greater than or equal to our threshold min\_samples or z.

DBSCAN iteratively expands the cluster, by going through each individual point within the cluster, and counting the number of other data points nearby.

**10.What are the limitations of DBSCAN clustering? How can these limitations be addressed?**

 DBSCAN cannot cluster data-sets with large differences in densities well, since then the minPts-eps combination cannot be chosen appropriately for all clusters. Choosing a meaningful eps value can be difficult if the data isn't well understood.

**11.Explain the concept of reachability in DBSCAN clustering and its significance in identifying clusters.**

Reachability : A point to be reachable from another if it lies within a particular distance (eps) from it, which indicates how densely reachable a cluster is.

It states if a data point can be accessed from another data point directly or indirectly,

Based on *reachability* and *connectivity* we can define the cluster and noise points.

Objects which are directly density-reachable from at least one core object are known as cluster  *and not reachable are known as Noise points*.

**12.Discuss the challenges of determining the optimal number of clusters in hierarchical clustering and DBSCAN.**

One of the fundamental challenges in clustering is determining the appropriate number of clusters in the data. Choosing the incorrect number of clusters can lead to poor clustering results. There are various techniques, such as the elbow method, silhouette analysis, are used to help estimate the optimal number of clusters.