Clustering

Q1. What is unsupervised learning in the context of machine learning?

Unsupervised learning is a type of machine learning where models learn patterns from unlabeled data. The goal is to discover hidden structures such as clusters or relationships without predefined output labels.

Q2. How does the K-Means clustering algorithm work?

K-Means works by selecting k initial centroids, assigning each data point to the nearest centroid, and then updating the centroids as the mean of assigned points. This process repeats until convergence.

Q3. Explain the concept of a dendrogram in hierarchical clustering.

A dendrogram is a tree-like diagram that shows how data points or clusters are merged step by step in hierarchical clustering, along with the distance at which each merge occurs.

Q4. What is the main difference between K-Means and Hierarchical Clustering?

K-Means requires the number of clusters to be specified in advance, while hierarchical clustering does not and instead builds a hierarchy of clusters.

Q5. What are the advantages of DBSCAN over K-Means?

DBSCAN can find clusters of arbitrary shape, does not require specifying the number of clusters, and can identify noise points, unlike K-Means.

Q6. When would you use Silhouette Score in clustering?

Silhouette Score is used to evaluate clustering quality by measuring how well each data point fits within its cluster compared to other clusters.

Q7. What are the limitations of Hierarchical Clustering?

Hierarchical clustering is computationally expensive, sensitive to noise and outliers, and not suitable for very large datasets.

Q8. Why is feature scaling important in clustering algorithms like K-Means?

Feature scaling is important because K-Means uses distance calculations. Without scaling, features with larger values can dominate the clustering process.

Q9. How does DBSCAN identify noise points?

DBSCAN labels points as noise if they do not belong to a dense region, meaning they have fewer than the minimum required neighboring points within a given radius.

Q10. Define inertia in the context of K-Means.

Inertia is the sum of squared distances between each data point and its nearest cluster centroid. It measures how compact the clusters are.

Q11. What is the elbow method in K-Means clustering?

The elbow method helps determine the optimal number of clusters by plotting inertia against different values of k and identifying a point where the rate of decrease sharply changes.

Q12. Describe the concept of "density" in DBSCAN.

Density refers to the number of data points within a specified neighborhood. Dense regions form clusters, while sparse regions are considered noise.

Q13. Can hierarchical clustering be used on categorical data?

Yes, hierarchical clustering can be applied to categorical data if an appropriate distance or similarity measure is used.

Q14. What does a negative Silhouette Score indicate?

A negative Silhouette Score indicates that a data point may be assigned to the wrong cluster and is closer to another cluster.

Q15. Explain the term "linkage criteria" in hierarchical clustering.

Linkage criteria define how the distance between clusters is calculated, such as single, complete, average, or ward linkage.

Q16. Why might K-Means clustering perform poorly on data with varying cluster sizes or densities?

K-Means assumes clusters are spherical and equally sized, so it struggles when clusters differ in size, shape, or density.

Q17. What are the core parameters in DBSCAN, and how do they influence clustering?

The core parameters are eps (neighborhood radius) and min_samples (minimum points to form a dense region). They control cluster formation and noise detection.

Q18. How does K-Means++ improve upon standard K-Means initialization?

K-Means++ selects initial centroids more carefully by spreading them out, leading to faster convergence and better clustering results.

Q19. What is agglomerative clustering?

Agglomerative clustering is a bottom-up hierarchical method where each data point starts as its own cluster and clusters are merged step by step.

Q20. What makes Silhouette Score a better metric than just inertia for model evaluation?

Silhouette Score considers both cluster cohesion and separation, whereas inertia only measures compactness, making Silhouette Score more informative.