Q: What is unsupervised learning in the context of machine learning?

A: Unsupervised learning is a type of machine learning where the model is not provided with labeled data. Instead, it identifies hidden patterns or structures in the input data. Clustering and dimensionality reduction are common examples.

Q: How does the K-Means clustering algorithm work?

A: K-Means clustering follows these steps:

1. Choose K initial centroids randomly or using K-Means++.

2. Assign each data point to the nearest centroid (forming clusters).

3. Recalculate centroids as the mean of points in each cluster.

4. Repeat steps 2-3 until centroids don't change significantly or a max iteration is reached.

Q: Explain the concept of a dendrogram in hierarchical clustering.

A: A dendrogram is a tree-like diagram showing the hierarchical relationships among data points. It represents how clusters are merged or split step-by-step and helps in deciding the number of clusters by "cutting" the tree at a certain height.

Q: What is the main difference between K-Means and Hierarchical Clustering?

A: K-Means is a partitioning method needing the number of clusters in advance; it creates flat, non-overlapping clusters. Hierarchical clustering builds a nested hierarchy of clusters and doesn't require specifying the number of clusters initially.

Q: What are the advantages of DBSCAN over K-Means?

A: - No need to predefine K

- Handles arbitrary shapes

- Detects noise/outliers

- Better for clusters with varying densities

Q: When would you use Silhouette Score in clustering?

A: Use Silhouette Score to evaluate clustering quality. It measures how well a point fits within its cluster versus others. Values range from -1 to 1.

Q: What are the limitations of Hierarchical Clustering?

A: - Computationally expensive ($O(n^2)$)

- Sensitive to noise/outliers

- Irreversible

- Harder to scale to large datasets

Q: Why is feature scaling important in clustering algorithms like K-Means?

A: Because K-Means relies on Euclidean distance, features with larger scales can dominate. Feature scaling ensures each feature contributes equally to distance calculations.

Q: How does DBSCAN identify noise points?

A: A point is labeled as noise if it has fewer than min_samples points in its -neighborhood.

Q: Define inertia in the context of K-Means.

A: Inertia is the sum of squared distances from each point to its assigned centroid. It indicates cluster compactness - lower inertia means tighter clusters.

Q: What is the elbow method in K-Means clustering?

A: The elbow method helps determine the optimal number of clusters (K) by plotting K vs. inertia. The 'elbow point' suggests the best K.

Q: Describe the concept of 'density' in DBSCAN.

A: In DBSCAN, density refers to the number of points within a given radius. High-density regions form clusters.

Q: Can hierarchical clustering be used on categorical data?

A: Yes, but not directly. Use Gower distance, Hamming distance, or convert categories appropriately.

Q: What does a negative Silhouette Score indicate?

A: It means a point is likely assigned to the wrong cluster, as it's closer to another cluster than to its own.

Q: Explain the term 'linkage criteria' in hierarchical clustering.

A: Linkage criteria determine how the distance between clusters is calculated: single, complete, average, or Ward's linkage.

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

A: Because K-Means assumes equal-sized, spherical clusters and is sensitive to density and shape variations.

Q: What are the core parameters in DBSCAN, and how do they influence clustering?

A: - eps (): radius to search for neighbors

- min_samples: minimum number of points to form a dense region

Too small eps many noise points; too large merges clusters

Q: How does K-Means++ improve upon standard K-Means initialization?

A: K-Means++ spreads out initial centroids smartly, reducing poor local minima and improving convergence.

Q: What is agglomerative clustering?

A: A bottom-up hierarchical approach where each point starts as its own cluster and merges closest pairs iteratively.

Q: What makes Silhouette Score a better metric than just inertia for model evaluation?

A: Inertia only measures compactness. Silhouette Score also considers separation, making it more balanced and interpretable.