1. What is Unsupervised Learning?

Unsupervised learning is a type of machine learning where the model learns patterns or structures in data without using labeled outcomes. The goal is to discover hidden structures, groupings, or distributions in the data. Examples include clustering algorithms (e.g., K-Means, DBSCAN) and dimensionality reduction methods (e.g., PCA).

---

2. How Does K-Means Clustering Work?

K-Means is an iterative clustering algorithm:

1. Choose the number of clusters ((k)).
2. Initialize (k) cluster centroids randomly.
3. Assign each data point to the nearest centroid based on distance (commonly Euclidean).
4. Recalculate the centroids as the mean of points assigned to each cluster.
5. Repeat steps 3 and 4 until centroids no longer change significantly or a maximum number of iterations is reached.

---

3. What is a Dendrogram in Hierarchical Clustering?

A dendrogram is a tree-like diagram that visually represents the merging process in hierarchical clustering. Each leaf node corresponds to a data point, and branches show how clusters are combined at different levels of similarity.

---

4. Main Difference Between K-Means and Hierarchical Clustering

- **K-Means**: Divides the data into (k) clusters directly. It's fast but requires predefining (k).
- **Hierarchical Clustering**: Builds a nested hierarchy of clusters using a linkage criterion and does not require specifying (k) upfront.

---

5. Advantages of DBSCAN Over K-Means

- DBSCAN handles clusters of varying shapes and densities well.
- Identifies noise points effectively.
- Does not require specifying the number of clusters (k) beforehand.

---

6. When to Use Silhouette Score?

Use the Silhouette Score to evaluate clustering quality. It measures how similar a sample is to its assigned cluster compared to other clusters. Higher scores indicate better-defined clusters.

---

7. Limitations of Hierarchical Clustering

- Computationally expensive for large datasets.
- Sensitive to noise and outliers.
- Difficult to undo decisions once clusters are merged or split.

---

8. Why is Feature Scaling Important in Clustering?

Clustering algorithms like K-Means are distance-based, meaning features with larger scales dominate. Scaling (e.g., StandardScaler or MinMaxScaler) ensures all features contribute equally.

---

9. How Does DBSCAN Identify Noise Points?

DBSCAN labels points as noise if they:

1. Have fewer than `min_samples` neighbors within the specified `eps` distance.
2. Cannot be assigned to any cluster.

---

10. Define Inertia in K-Means

Inertia is the sum of squared distances between data points and their assigned cluster centroids. Lower inertia indicates tighter clusters.

---

11. What is the Elbow Method in K-Means?

The elbow method helps determine the optimal number of clusters ((k)) by plotting (k) against inertia. The "elbow" point, where the curve starts flattening, indicates a good (k).

---

12. Concept of "Density" in DBSCAN

Density in DBSCAN is defined by:

- (\text{eps}): Maximum distance between two points to be considered neighbors.
- (\text{min_samples}): Minimum number of points required to form a dense region (cluster).

---

13. Can Hierarchical Clustering Be Used on Categorical Data?

Yes, by calculating a dissimilarity matrix (e.g., Hamming distance for categorical variables). Specialized methods like **Gower Distance** can also handle mixed data types.

---

14. What Does a Negative Silhouette Score Indicate?

A negative score indicates that a sample is closer to points in other clusters than its own, suggesting poor clustering or overlap.

---

15. What is "Linkage Criteria" in Hierarchical Clustering?

Linkage criteria determine how distances between clusters are computed:

- **Single**: Shortest distance between points.
- **Complete**: Farthest distance between points.
- **Average**: Mean distance between points.
- **Ward**: Minimizes variance within clusters.

---

16. Why Might K-Means Perform Poorly on Varying Cluster Sizes/Densities?

K-Means assumes clusters are spherical and equally sized, so it struggles with:

- Non-spherical shapes.
- Clusters with unequal sizes or densities.

---

17. Core Parameters in DBSCAN

- **(\text{eps})**: Maximum distance for two points to be neighbors.
- **(\text{min_samples})**: Minimum points needed to form a cluster. Larger `eps` forms bigger clusters; smaller `min_samples` increases sensitivity.

---

18. How Does K-Means++ Improve Initialization?

K-Means++ selects initial centroids more strategically, reducing the likelihood of poor clustering by:

1. Picking one random point.
2. Choosing subsequent points far from existing centroids.

---

19. What is Agglomerative Clustering?

Agglomerative clustering is a **bottom-up approach** in hierarchical clustering. Each data point starts as its own cluster, and clusters are iteratively merged based on a linkage criterion.

---

20. Why is Silhouette Score Better than Inertia for Evaluation?

Inertia only measures compactness (distance within clusters), while Silhouette Score evaluates both:

- Compactness (tightness within a cluster).
- Separation (distance between clusters).
