Assignment-Clustering

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

Unsupervised learning is a type of machine learning where algorithms learn patterns from unlabeled data without explicit supervision or target variables. The system tries to discover hidden structures, relationships, or patterns in the data on its own.

2. How does K-Means clustering algorithm work?

K-Means works by:

1. Randomly initializing K cluster centroids
2. Assigning each data point to the nearest centroid
3. Recalculating centroids as the mean of all points in each cluster
4. Repeating steps 2-3 until centroids stabilize or max iterations reached

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

A dendrogram is a tree-like diagram that records the sequences of merges or splits in hierarchical clustering. The vertical axis represents distance or dissimilarity between clusters, while the horizontal axis shows the data points. It provides a visual representation of the clustering process at all levels.

4. What is the main difference between K-Means and hierarchical clustering?

The main difference is that K-Means requires pre-specifying the number of clusters (K) and produces flat clusters, while hierarchical clustering creates a tree of clusters (dendrogram) that can be cut at any level to obtain different numbers of clusters without recomputation.

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

DBSCAN advantages:

- Can find arbitrarily shaped clusters
- Doesn't require specifying number of clusters
- Can identify noise/outliers
- Works well with clusters of varying densities
- Not sensitive to initialization like K-Means

6. When would you use Silhouette Score in clustering?

Silhouette Score is used to:

- Evaluate clustering quality when true labels are unknown
- Compare different clustering results
- Determine optimal number of clusters
- Assess how well each point fits its assigned cluster

7. What are the limitations of Hierarchical Clustering?

Limitations include:

- Computationally expensive ($O(n^3)$ for most methods)
- Sensitive to noise and outliers
- Once a decision is made to combine clusters, it cannot be undone
- Difficult to handle large datasets
- Memory intensive for big data

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

Feature scaling is important because:

- K-Means uses distance measures (typically Euclidean)
- Features on larger scales dominate the distance calculation
- Without scaling, features with larger ranges get more weight
- Can lead to poor clustering results

9. How does DBSCAN identify noise points?

DBSCAN identifies noise points as:

- Points that don't fall within the ε-neighborhood of any core point
- Points that aren't density-reachable from any core point
- Points that don't have enough neighbors (min_samples) within ε distance

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

Inertia is the sum of squared distances of samples to their closest cluster center. It's the objective function that K-Means tries to minimize. Lower inertia indicates better clustering where points are closer to their centroids.

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

The elbow method is a technique to determine the optimal number of clusters (K) by:

1. Running K-Means for different K values
2. Plotting inertia against K
3. Looking for the "elbow" point where inertia starts decreasing linearly
4. Selecting K at this elbow point as optimal

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

In DBSCAN, density is defined by two parameters:

- ε (eps): Radius of the neighborhood around a point
- min_samples: Minimum number of points required within ε to form a dense region
  A point is considered a core point if it has at least min_samples points within ε distance.

13. Can hierarchical clustering be used on categorical data?

Yes, but with appropriate distance measures:

- Need to use similarity measures for categorical data (e.g., Hamming distance, Jaccard similarity)
- Gower's distance can handle mixed data types
- Standard Euclidean distance won't work for pure categorical data

14. What does a negative Silhouette Score indicate?

A negative Silhouette Score indicates that:

- Many points might be assigned to wrong clusters
- On average, points are closer to points in other clusters than their own
- The clustering configuration may be worse than random assignment

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

Linkage criteria determine how to measure distance between clusters when merging them. Common types:

- Single linkage: Minimum distance between any two points
- Complete linkage: Maximum distance between any two points
- Average linkage: Average distance between all point pairs
- Ward's method: Minimizes variance when merging clusters

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

K-Means performs poorly because:

- It assumes clusters are spherical and equally sized
- Centroids get pulled toward larger clusters
- Dense clusters may get split while sparse clusters may get merged
- Uses Euclidean distance which favors equal-sized clusters

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

Core parameters:

- ε (eps): Radius of neighborhood - larger values form larger clusters
- min_samples: Minimum points to form dense region - higher values make more noise points
  They influence:
- Number of clusters formed
- What's considered noise
- Cluster density requirements
- Ability to find nested clusters

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

K-Means++ improves initialization by:

- Choosing first centroid randomly
- Selecting subsequent centroids with probability proportional to distance² from nearest existing centroid
- This spreads out initial centroids
- Leads to faster convergence and better final results

19. What is agglomerative clustering?

Agglomerative clustering is a bottom-up hierarchical clustering approach where:

- Each point starts as its own cluster
- Closest pairs of clusters are merged iteratively
- Continues until all points are in one cluster
- Forms a dendrogram showing merge sequence

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

Silhouette Score is better because:

- Considers both intra-cluster and inter-cluster distances
- Not biased toward spherical clusters like inertia
- Values are normalized (-1 to 1) allowing comparison across datasets
- Can detect when clusters are poorly defined
- Works better for non-convex clusters