Clustering & Unsupervised Learning - Assignment Answers

1. Unsupervised Learning in Machine Learning:

Unsupervised learning involves finding patterns or structures in data without using labeled outcomes. It's used for clustering, dimensionality reduction, and anomaly detection.

2. K-Means Clustering Algorithm:

- Step 1: Choose the number of clusters (K) and randomly initialize K centroids.
- Step 2: Assign each data point to the nearest centroid (using distance metric).
- Step 3: Recompute centroids as the mean of assigned points.
- Repeat until assignments stop changing.

3. Dendrogram in Hierarchical Clustering:

A dendrogram is a tree diagram that shows the order and distances of merges during hierarchical clustering. Cutting the tree at a certain height gives cluster groups.

4. K-Means vs. Hierarchical Clustering:

- K-Means: Requires pre-specified K, works on large datasets, faster.
- Hierarchical: No need to pre-specify clusters, better for small datasets, more interpretable.

5. Advantages of DBSCAN Over K-Means:

- Does not require specifying number of clusters.
- Can find clusters of arbitrary shapes.
- Handles noise (outliers).

6. Silhouette Score in Clustering:

Measures how well each data point fits into its cluster (value ranges from -1 to 1). High scores indicate well-separated clusters.

7. Limitations of Hierarchical Clustering:

- Computationally expensive (slow on large datasets).
- Sensitive to noise and outliers.
- Once split/merged, cannot be undone.

8. Why Feature Scaling is Important in Clustering?

Distance-based clustering (K-Means, Hierarchical) is sensitive to feature scales. Scaling ensures no feature dominates due to its magnitude.

9. How DBSCAN Identifies Noise:

Points not assigned to any cluster (due to insufficient nearby points) are marked as noise.

10. Inertia in K-Means:

It's the sum of squared distances from each point to its cluster centroid. Lower inertia means tighter clusters.

11. Elbow Method in K-Means:

Plots inertia vs. number of clusters (K) to find the 'elbow' point where inertia decrease slows down, indicating optimal K.

12. Density in DBSCAN:

Clusters are formed based on point density, where high-density areas form clusters, and sparse areas are noise.

13. Hierarchical Clustering for Categorical Data:

Yes, using appropriate distance metrics (like Hamming distance).

14. Negative Silhouette Score Meaning:

Indicates that points may be assigned to the wrong cluster.

15. Linkage Criteria in Hierarchical Clustering:

Defines how distances between clusters are measured: Single (nearest), Complete (farthest), Average, Ward (minimizes variance).

16. K-Means on Varying Cluster Sizes/Densities:

K-Means assumes spherical clusters of equal size and density. It struggles when clusters differ significantly.

17. Core Parameters of DBSCAN:

- eps (eps): Maximum distance between two points to be neighbors.
- min_samples: Minimum points required to form a dense region (cluster).

18. K-Means++ Initialization:

Improves K-Means by carefully choosing initial centroids to spread them out, reducing convergence issues.

19. Agglomerative Clustering:

A bottom-up approach where each data point starts as its own cluster, and clusters are merged iteratively based on distance.

20. Silhouette Score vs. Inertia:

Silhouette score considers both intra-cluster and inter-cluster distances, giving a more reliable evaluation of clustering quality than inertia alone.