Assignment – Clustering

Q What is unsupervised learning in the context of machine learning.

→ Unsupervised learning is a type of machine learning where the algorithm is trained on unlabeled data.

- model tries to find hidden patterns, structure & grouping.

Q.2 How does k-means clustering algorithm work –

i) choose the No. of cluster k.

ii) Initialize k cluster centroid (randomly or using k-means++).

iii) Assign each data point to the nearest centroid

iv) update centroid as the mean of point in each cluster.

v) Repeat (3) & (4) until convergence (centroid stop moving significantly

Q.3 Explain the concept of dendrogram in hierarchical clustering

A dendrogram is a tree-like diagram that shows how cluster are merged or split at each step in hierarchical clustering.

- The vertical axis represents the distance dissimilarity b/w cluster. By cutting the

the cluster

Q4 what is the main difference b/w K-means & Hierarchical clustering.

K-means :- Requires No. of cluster K in advance, partitions data into fixed cluster.

Hierarchical clustering:

Builds a hierarchy of cluster (merging or splitting) without K - upfront

Q5 what are the advantages of DBSCAN over K-means.

- Can find cluster of arbitrary shape.
- Does not require specifying the No. of cluster in advance.
- Indentifies and labels noise/outliers.
- Work well with varying cluster densities.

Q.6 when would you use Silhouette Score in clustering?

Silhoutte Score measures how well-separated and compact the clusters are.

- Use it to evaluate cluster quality.
- Useful for choosing the optimal No. of cluster.

clustering

very expensive computationally for large datasets $O(n^2)$
Sensitive to noise and outliers
once a merge / split done, it cannot be undone.

Q8 Why is feature scaling important in clustering algorithms like k-means
k-means uses Euclidean distance. If features are not scaled, variable with larges ranges dominate the distance calculation leading to biased clusters

Q9 How does DBSCAN identify noise points.
A point is labeled as noise if it is:
→ Not within the neighborhood (ε) of any core point.
→ Not reachable from any clusters

Q10 Define inertia in the context of k-means.
Inertia is the sum of squared distance of each data point to its assigned cluster centroid.
- Lower inertia means tighter & more compact cluster.

Q.11 What is the elbow method in k-means clustering.

It is a method to determine the optimal No. of clusters k. Plot Intertia Vs k and look for a point (the elbow) where adding more clusters does not significantly reduce Inertia.

Q.12 Describe the concept of 'density' in DBSCAN

- ε (Epsilon): Radius of neighborhood around a point.
- MinPts: Minimum No. of points within ε to a dense region. Dense Regions form clusters, are noise.

Q.13 Can hierarchical clustering used on categorical data?

- Yes but not directly with standard distance metrics. You need to use special similarity measure (like Hamming distance or Gower distance) suitable categorical data

Q.14 What does a negative Silhouette score indicate?

A negative score means the point is closer to another cluster than its assigned cluster.

→ Indicates wrong clustering.

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

- Single linkage :- Min. distance b/w points of two clusters.
- complete linkage: Maximum distance.
- Average linkage: Average distance.
- Ward's method: minimize variance b/w cluster

Q.16 why might k-means clustering perform poorly on data with varying cluster size or densities

- Clusters are spherical are equally sized
- All clusters have similar density.
- It struggle with cluster of different size or densities

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

- ε (Epsilon): Radius of neighborhood. Larger ε – longer clusters
- MinPts: Min. No. of neighbors to form a dense region

Q.18 How does K-means++ improve upon standard K-means initialization

Instead of random initialization, K-means++ spreads intial centroid for apart, which:

- Reduces chances of poor clustering.
- Leads to faster convergence.
- Produces more stable results

Q.19 What is agglomerative clustering.

A bottom-up hierarchical clustering method where each data point starts as its own cluster, and clusters are merged step by step until only one cluster remains

Q.20 What make Silhouette Score a better metric than just inertia for model evaluation

- Inertia: Only measures compactness (lower is better) but ignores separation.

- Silhouette Score: considers both compactness and separation, giving a better overall evaluation of cluster quality.