Assignment – Clustering: Theoretical Questions

1. **What is unsupervised learning in the context of machine learning?**
   Unsupervised learning is a type of machine learning where the model identifies patterns or groupings in data without labeled outputs. Clustering and dimensionality reduction are common examples.
2. **How does K-Means clustering algorithm work?**
   K-Means partitions data into *k* clusters by randomly initializing centroids, assigning points to the nearest centroid, and updating centroids iteratively until convergence.
3. **Explain the concept of a dendrogram in hierarchical clustering.**
   A dendrogram is a tree-like diagram that shows the order and distances at which clusters are merged or split in hierarchical clustering, helping to decide the number of clusters.
4. **What is the main difference between K-Means and Hierarchical Clustering?**
   K-Means is a partitional algorithm requiring a predefined number of clusters, while hierarchical clustering builds a tree structure without needing to predefine the cluster count.
5. **What are the advantages of DBSCAN over K-Means?**
   DBSCAN can find arbitrarily shaped clusters and detect noise points. It doesn't require specifying the number of clusters and performs better on non-spherical datasets.
6. **When would you use Silhouette Score in clustering?**
   Use it to evaluate the quality of clusters. It measures how similar a point is to its own cluster vs. others. A higher score indicates better-defined clusters.
7. **What are the limitations of Hierarchical Clustering?**
   It's not scalable to large datasets, sensitive to noise, and once merged or split, clusters cannot be undone. It also lacks flexibility in reassigning points.
8. **Why is feature scaling important in clustering algorithms like K-Means?**
   K-Means relies on distance calculations. Without feature scaling, variables with larger ranges dominate the distance measure, skewing clustering results.
9. **How does DBSCAN identify noise points?**
   DBSCAN labels points as noise if they don't have enough neighbors within a given radius (eps). These points are not assigned to any cluster.

Assignment – Clustering: Theoretical Questions

10. **Define inertia in the context of K-Means.**
    Inertia is the sum of squared distances between data points and their assigned cluster centroids. Lower inertia indicates more compact clusters.
11. **What is the elbow method in K-Means clustering?**
    The elbow method plots inertia vs. number of clusters. The “elbow point” where inertia stops decreasing sharply suggests the optimal number of clusters.
12. **Describe the concept of "density" in DBSCAN.**
    Density refers to the number of data points within a specified radius (eps). Clusters form in dense regions where the number of points exceeds a threshold (minPts).
13. **Can hierarchical clustering be used on categorical data?**
    Yes, by using appropriate distance metrics like Hamming or Gower distance. However, it’s more common with continuous data unless pre-processed carefully.
14. **What does a negative Silhouette Score indicate?**
    A negative Silhouette Score means the point may be assigned to the wrong cluster, as it’s closer to a neighboring cluster than its own.
15. **Explain the term "linkage criteria" in hierarchical clustering.**
    Linkage criteria define how distances between clusters are calculated, such as single (minimum), complete (maximum), or average linkage.
16. **Why might K-Means clustering perform poorly on data with varying cluster sizes or densities?**
    K-Means assumes equal-sized, spherical clusters. It struggles with clusters that differ in shape, size, or density, leading to inaccurate groupings.
17. **What are the core parameters in DBSCAN, and how do they influence clustering?**
    The key parameters are eps (neighborhood radius) and minPts (minimum points to form a dense region). They determine cluster formation and noise sensitivity.
18. **How does K-Means++ improve upon standard K-Means initialization?**
    K-Means++ spreads out initial centroids more strategically, reducing the chances of poor clustering and improving convergence speed and accuracy.

Assignment – Clustering: Theoretical Questions

19. **What is agglomerative clustering?**

    Agglomerative clustering is a bottom-up hierarchical approach where each point starts in its own cluster, and pairs of clusters are merged iteratively.

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

    Unlike inertia, which only measures compactness, Silhouette Score also considers how well-separated clusters are, providing a more complete evaluation.