Clustering

Assignment Questions

PW SKILLS

Theoretical Questions:

1. What is unsupervised learning in the context of machine learning?
2. How does K-Means clustering algorithm work?
3. Explain the concept of a dendrogram in hierarchical clustering.
4. What is the main difference between K-Means and Hierarchical Clustering?
5. What are the advantages of DBSCAN over K-Means?
6. When would you use Silhouette Score in clustering?
7. What are the limitations of Hierarchical Clustering?
8. Why is feature scaling important in clustering algorithms like K-Means?
9. How does DBSCAN identify noise points?
10. Define inertia in the context of K-Means.
11. What is the elbow method in K-Means clustering?
12. Describe the concept of "density" in DBSCAN.
13. Can hierarchical clustering be used on categorical data?
14. What does a negative Silhouette Score indicate?
15. Explain the term "linkage criteria" in hierarchical clustering.
16. Why might K-Means clustering perform poorly on data with varying cluster sizes or densities?
17. What are the core parameters in DBSCAN, and how do they influence clustering?
18. How does K-Means++ improve upon standard K-Means initialization?
19. What is agglomerative clustering?
20. What makes Silhouette Score a better metric than just inertia for model evaluation?

Practical Questions:

21. Generate synthetic data with 4 centers using make_blobs and apply K-Means clustering. Visualize using a scatter plot.
22. Load the Iris dataset and use Agglomerative Clustering to group the data into 3 clusters. Display the first 10 predicted labels.
23. Generate synthetic data using make_moons and apply DBSCAN. Highlight outliers in the plot.
24. Load the Wine dataset and apply K-Means clustering after standardizing the features. Print the size of each cluster.
25. Use make_circles to generate synthetic data and cluster it using DBSCAN. Plot the result.
26. Load the Breast Cancer dataset, apply MinMaxScaler, and use K-Means with 2 clusters. Output the cluster centroids.
27. Generate synthetic data using make_blobs with varying cluster standard deviations and cluster with DBSCAN.
28. Load the Digits dataset, reduce it to 2D using PCA, and visualize clusters from K-Means.
29. Create synthetic data using make_blobs and evaluate silhouette scores for k = 2 to 5. Display as a bar chart.
30. Load the Iris dataset and use hierarchical clustering to group data. Plot a dendrogram with average linkage.
31. Generate synthetic data with overlapping clusters using make_blobs, then apply K-Means and visualize with decision boundaries.
32. Load the Digits dataset and apply DBSCAN after reducing dimensions with t-SNE. Visualize the results.
33. Generate synthetic data using make_blobs and apply Agglomerative Clustering with complete linkage. Plot the result.
34. Load the Breast Cancer dataset and compare inertia values for K = 2 to 6 using K-Means. Show results in a line plot.
35. Generate synthetic concentric circles using make_circles and cluster using Agglomerative Clustering with single linkage.
36. Use the Wine dataset, apply DBSCAN after scaling the data, and count the number of clusters (excluding noise)
37. Generate synthetic data with make_blobs and apply KMeans. Then plot the cluster centers on top of the data points.
38. Load the Iris dataset, cluster with DBSCAN, and print how many samples were identified as noise.
39. Generate synthetic non-linearly separable data using make_moons, apply K-Means, and visualize the clustering result.
40. Load the Digits dataset, apply PCA to reduce to 3 components, then use KMeans and visualize with a 3D scatter plot.

41. Generate synthetic blobs with 5 centers and apply KMeans. Then use silhouette_score to evaluate the clustering.
42. Load the Breast Cancer dataset, reduce dimensionality using PCA, and apply Agglomerative Clustering. Visualize in 2D.
43. Generate noisy circular data using make_circles and visualize clustering results from KMeans and DBSCAN side-by-side.
44. Load the Iris dataset and plot the Silhouette Coefficient for each sample after KMeans clustering.
45. Generate synthetic data using make_blobs and apply Agglomerative Clustering with 'average' linkage. Visualize clusters.
46. Load the Wine dataset, apply KMeans, and visualize the cluster assignments in a seaborn pairplot (first 4 features).
47. Generate noisy blobs using make_blobs and use DBSCAN to identify both clusters and noise points. Print the count.
48. Load the Digits dataset, reduce dimensions using t-SNE, then apply Agglomerative Clustering and plot the clusters.