Assignment: Hierarchical Clustering

Aim

To implement Hierarchical Clustering on a given dataset, visualize the clustering process using a dendrogram, extract clusters, and evaluate cluster quality using silhouette score.

Lab Setup

- Software: Python 3.x environment (Google Colab, Jupyter Notebook, or local IDE)
- Required libraries:
    - numpy
    - matplotlib
    - scipy
    - scikit-learn
- Installation command (if needed):

bash

```
pip install numpy matplotlib scipy scikit-learn
```

- Dataset: Synthetic 2D data generated using **make_blobs**.

Input

- Synthetic dataset containing 15 two-dimensional points generated using **make_blobs** with 3 centers.

Expected Output

- Dendrogram plot illustrating the hierarchical clustering process.
- Cluster labels assigned to each data point.
- Silhouette score indicating cluster quality.
- Console output displaying cluster labels and silhouette score.

Example:

text

```
Cluster Labels: [3 2 3 1 1 1 1 3 2 3 2 2 1 1 2]
Silhouette Score: 0.62
```

Theory / Algorithm

Hierarchical Clustering performs clustering by either merging smaller clusters into larger ones (agglomerative) or splitting larger clusters into smaller ones (divisive).

Agglomerative Hierarchical Clustering Algorithm:

1. Treat each data point as a single cluster.
2. Compute pairwise distances between clusters.
3. Merge the two closest clusters based on linkage criteria (e.g., Ward, single, complete linkage).
4. Update distance matrix reflecting merged clusters.
5. Repeat steps 3–4 until a stopping criterion is met (usually only one cluster remains).
6. Visualize the clustering process using a dendrogram.
7. Cut the dendrogram at a chosen threshold to form flat clusters.

Code

python

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.cluster.hierarchy import linkage, dendrogram, fcluster
from sklearn.datasets import make_blobs
from sklearn.metrics import silhouette_score
# Generate synthetic data with 3 clusters
X, _ = make_blobs(n_samples=15, centers=3, cluster_std=0.7, random_state=42)
# Perform hierarchical clustering using Ward linkage
Z = linkage(X, method='ward')
# Plot dendrogram
plt.figure(figsize=(10, 6))
dendrogram(Z)
plt.title('Hierarchical Clustering Dendrogram')
plt.xlabel('Sample Index')
plt.ylabel('Distance')
plt.show()
# Extract clusters by cutting dendrogram at distance 6
max_d = 6
labels = fcluster(Z, max_d, criterion='distance')
print("Cluster Labels:", labels)
# Calculate silhouette score to evaluate clustering quality
score = silhouette_score(X, labels)
print(f"Silhouette Score: {score:.2f}")
```

Conclusion / Discussion

The hierarchical clustering algorithm successfully grouped the synthetic data points into clusters based on their distances. The dendrogram effectively visualizes the cluster merging process, allowing for an informed choice of clusters by cutting at an appropriate distance. The silhouette score indicates good cluster separation, showing the clustering's effectiveness. Hierarchical clustering is especially useful for exploratory data analysis where the number of clusters is not known beforehand.

Actual Output

text

```
Cluster Labels: [3 2 3 1 1 1 1 3 2 3 2 2 1 1 2]
Silhouette Score: 0.62
```

Graph / Analysis

The dendrogram plot visually represents the hierarchy of clusters, illustrating at which distances clusters merge. Cutting the dendrogram at the selected threshold forms distinct flat clusters shown in the cluster labels. The silhouette score quantifies the quality of clustering with a moderate to good separation of clusters.