Compare Clustering Methods.

Introduction

This report analyzes the application of K-means and Hierarchical clustering algorithms, on the Iris and MNIST datasets. The performance of these models was evaluated based on sum of squared errors (SSE) / inertia, silhouette score, adjusted rand index (ARI) and execution time.

Methodology

Dataset:

- For this assignment, we are using the Iris dataset and the MNIST dataset, which we load from sklearn.datasets.
- Iris dataset shape is (150, 4).
- The MNIST dataset shape is (70000, 784). We extract a subset of the dataset by using the train-test-split method with the stratify option (stratify=yes) and then split MNIST subset to X_train and X_test for the calculations. The MNIST subset shape is(2100, 784) .
- The datasets were standardized using StandardScaler.
- Iris dataset has 3 distinct labels and each of the 3 labels has 50 instances, indicating a balanced dataset.
- MNIST dataset has 10 distinct labels prior to and after subset while the counts of instances for each digit/distinct label are different.

Evaluation Metrics:

The clustering techniques were implemented in Python using the Scikit-learn and SciPy libraries and evaluated based on the following metrics.

- Sum of Squared Errors (SSE) / Inertia
- Silhouette score
- Adjusted Rand Index (ARI)
- Execution time

Results and Analysis:

Elbow Method & Silhouette Score Analysis.

For Iris dataset, the plot shows potential elbow around K=3 in the SSE plot, while the Silhouette Score peaked at K=2.

For the MNIST dataset, the SSE plot shows a gradual decrease without a clear elbow and the Silhouette Score peaked at K=2 with a very low score, indicating poorly separated clusters.

Terminal Output:

- The Optimal K selected from the Elbow Method for mnist: 2.
- The Optimal K selected from the Elbow Method for iris: 2.

The table below shows the terminal output of the evaluation metrics.

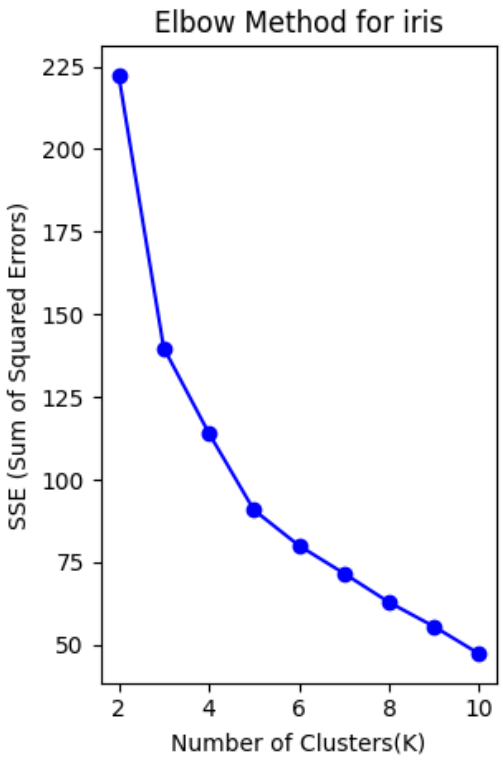
For the Iris dataset, all algorithms were very fast. K-Means showed slightly better clustering for Iris, achieving higher scores for both silhouette and adjusted rand index compared to both scikit-learn and SciPy hierarchical clustering.

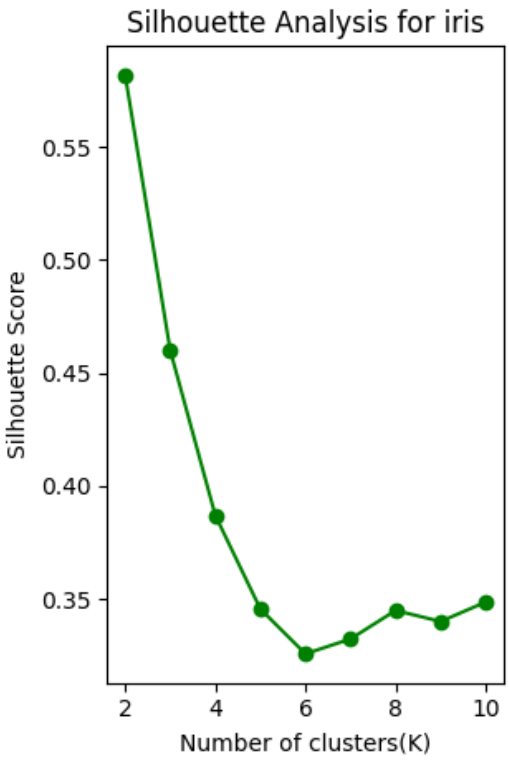
For the MNIST dataset, all algorithms had comparable execution time and struggled with cluster separation. Both scikit-learn and SciPy hierarchical clustering showed a better alignment with the true digit labels (higher adjusted rand index) compared to K-Means.

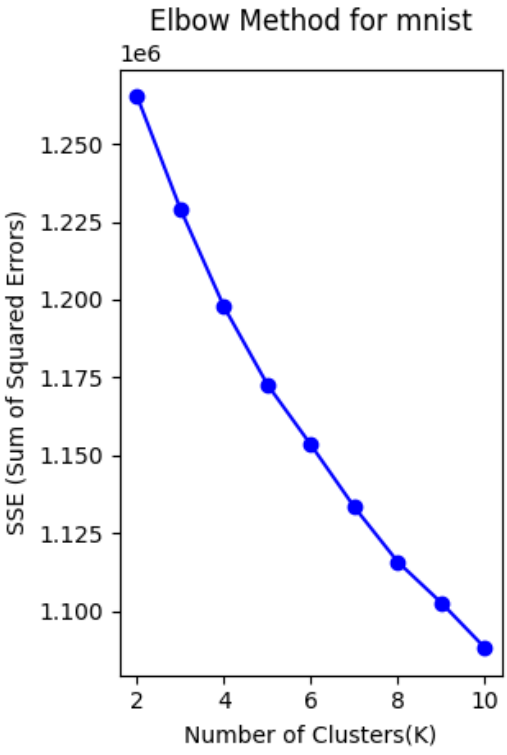
IRIS DATASET				
ALGORITHM	Sum of Squared Errors (SSE) / Inertia	Silhouette Coefficient	Adjusted Rand Index	Execution Time
Scikit-learn Hierarchical		**0.447**	**0.615**	**0.002**
SciPy Hierarchical		**0.447**	**0.615**	**0.000**
K-Means	**139.820**	**0.460**	**0.620**	**0.099**
MNIST DATASET				
ALGORITHM	Sum of Squared Errors (SSE) / Inertia	Silhouette Coefficient	Adjusted Rand Index (ARI)	Execution Time (seconds)
Scikit-learn Hierarchical		**-0.027**	**0.430**	**1.494**
SciPy Hierarchical		**-0.027**	**0.430**	**1.453**
K-Means	**1089400.261**	**0.009**	**0.355**	**1.427**

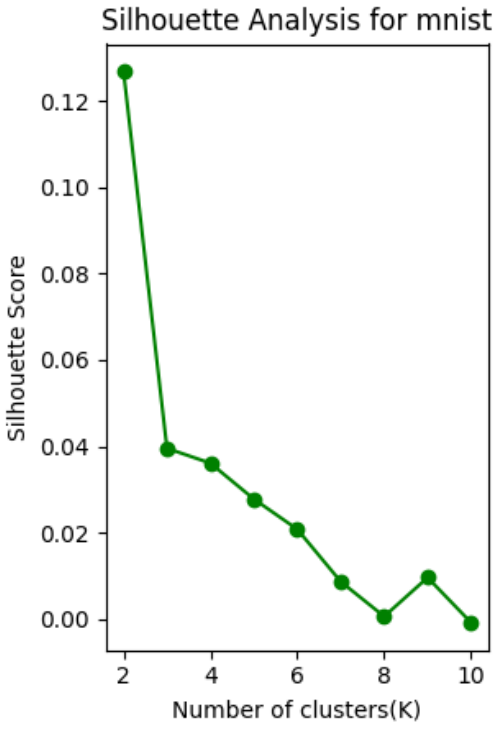
Plots

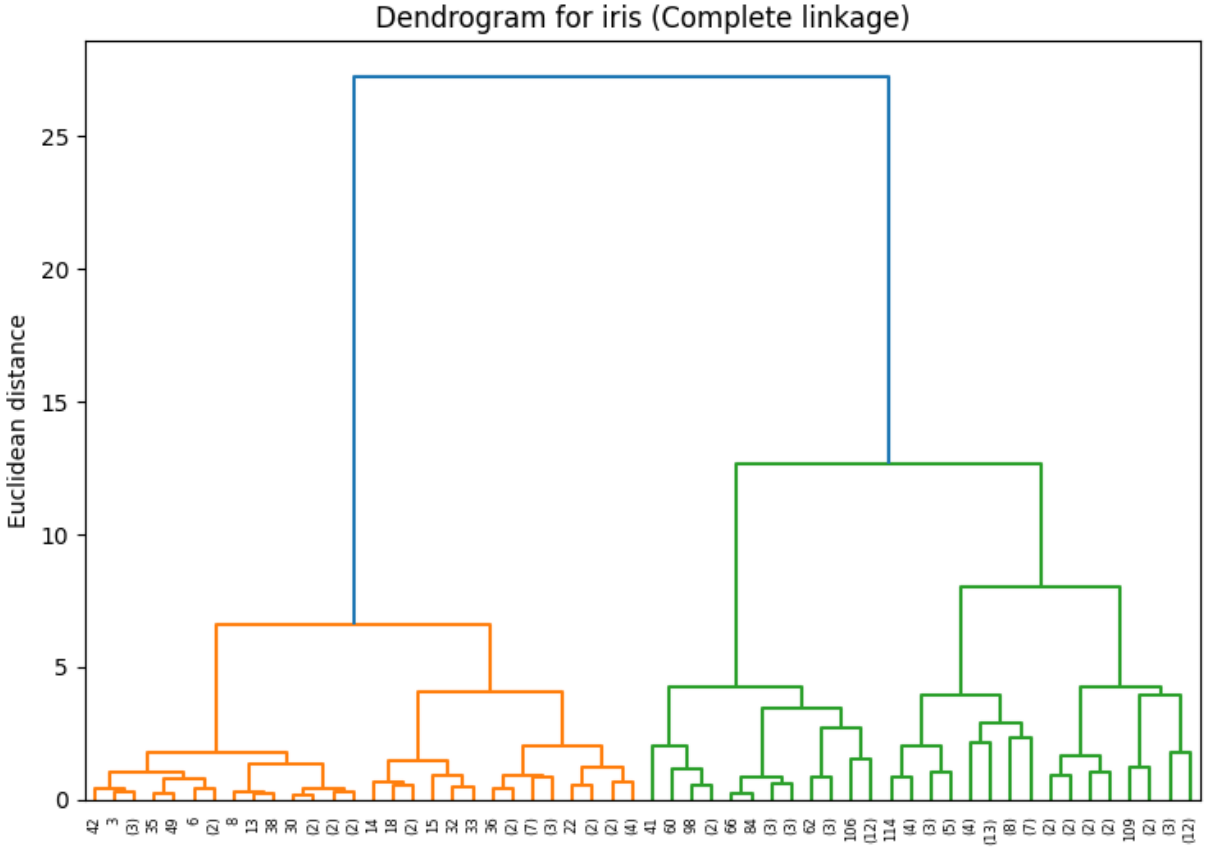
Below are the plots generated.

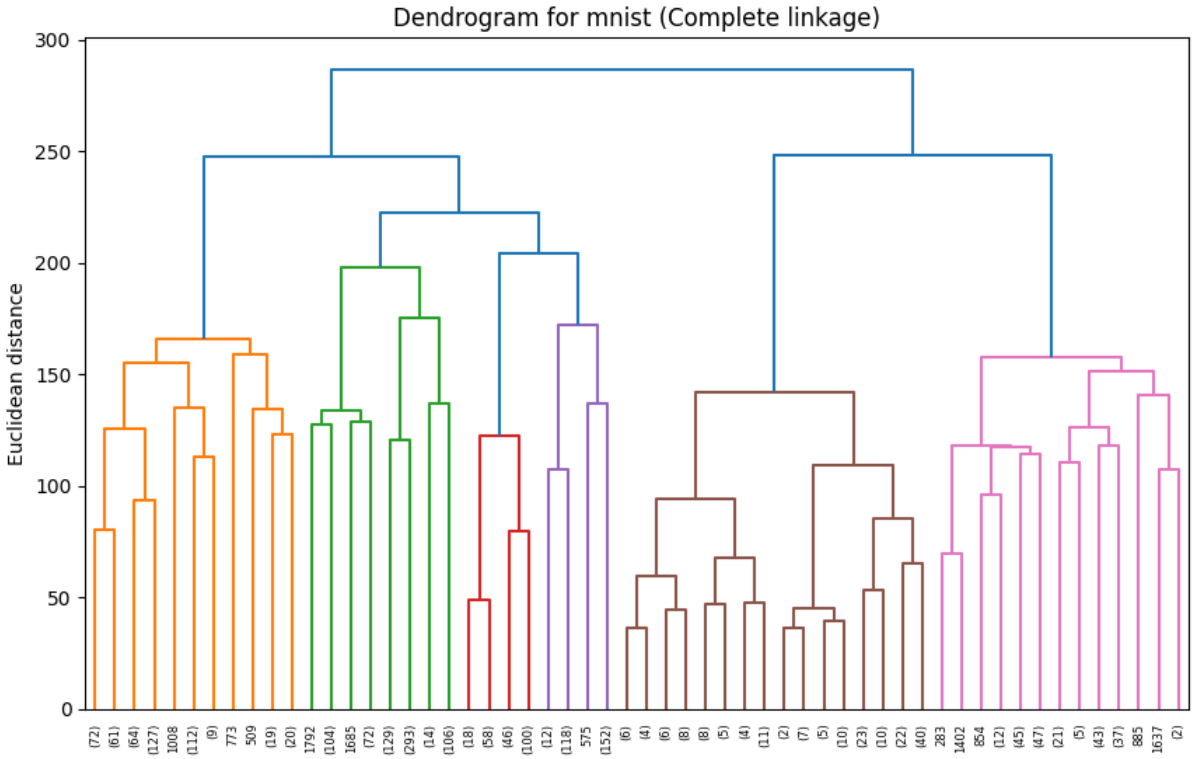

Elbow Method for iris
SSE (Sum of Squared Errors)
225
200
175
150
125
100
75
50
2
4
6
8
10
Number of Clusters(K)

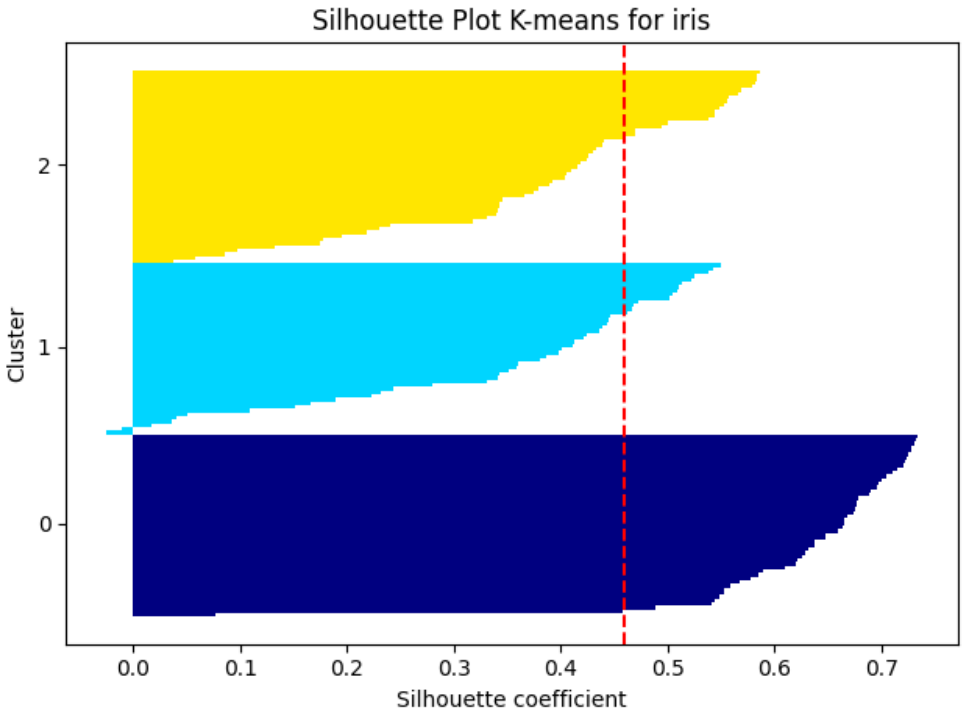

Silhouette Analysis for iris
Silhouette Score
0.55
0.50
0.45
0.40
0.35
2
4
6
8
10
Number of clusters(K)

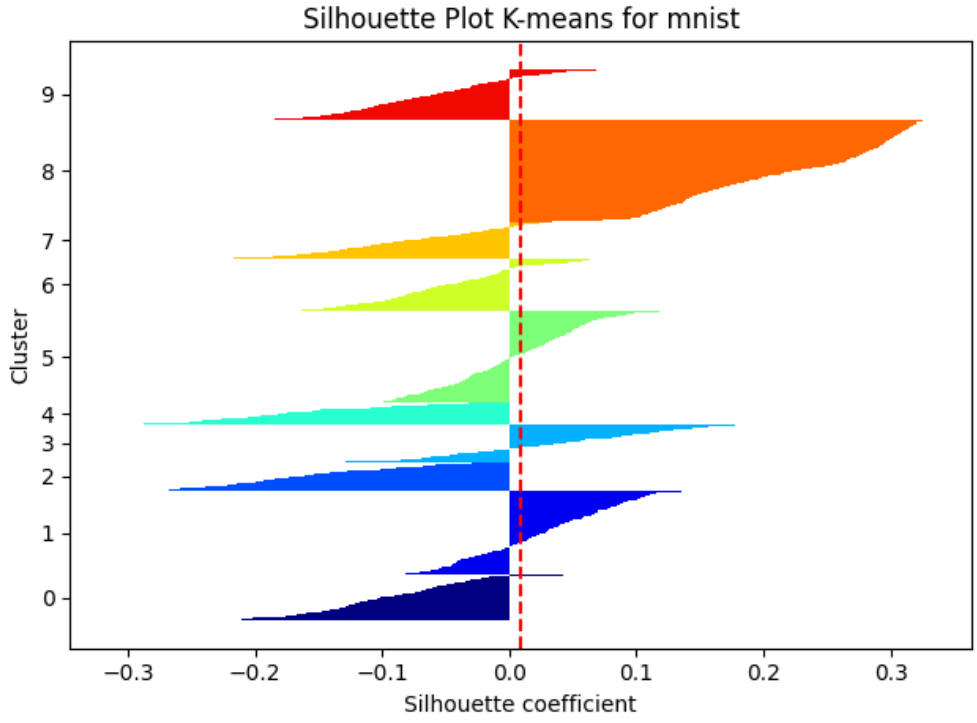

Elbow Method for mnist
1e6
SSE (Sum of Squared Errors)
1.250
1.225
1.200
1.175
1.150
1.125
1.100
2
4
6
8
10
Number of Clusters(K)

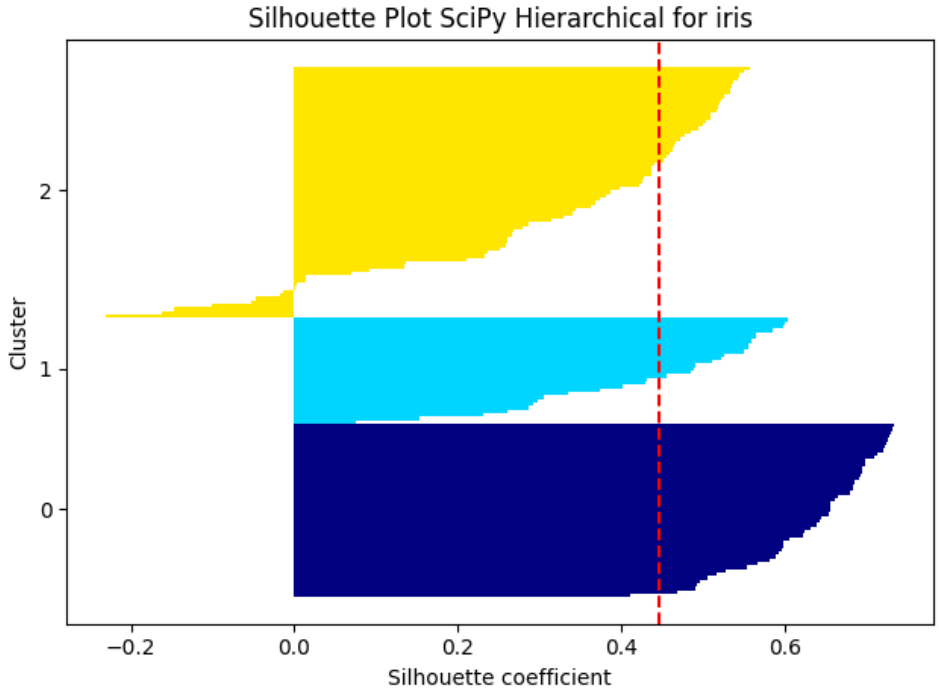

Silhouette Analysis for mnist
Silhouette Score
0.12
0.10
0.08
0.06
0.04
0.02
0.00
2
4
6
8
10
Number of clusters(K)

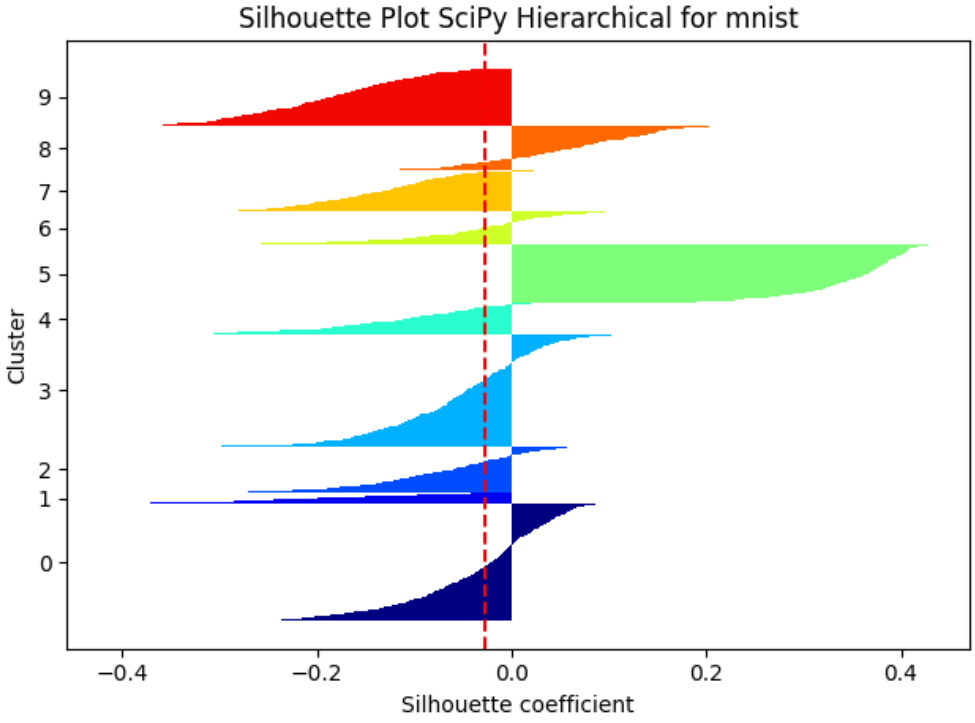

Dendrogram for iris (Complete linkage)
Euclidean distance
0
5
10
15
20
25

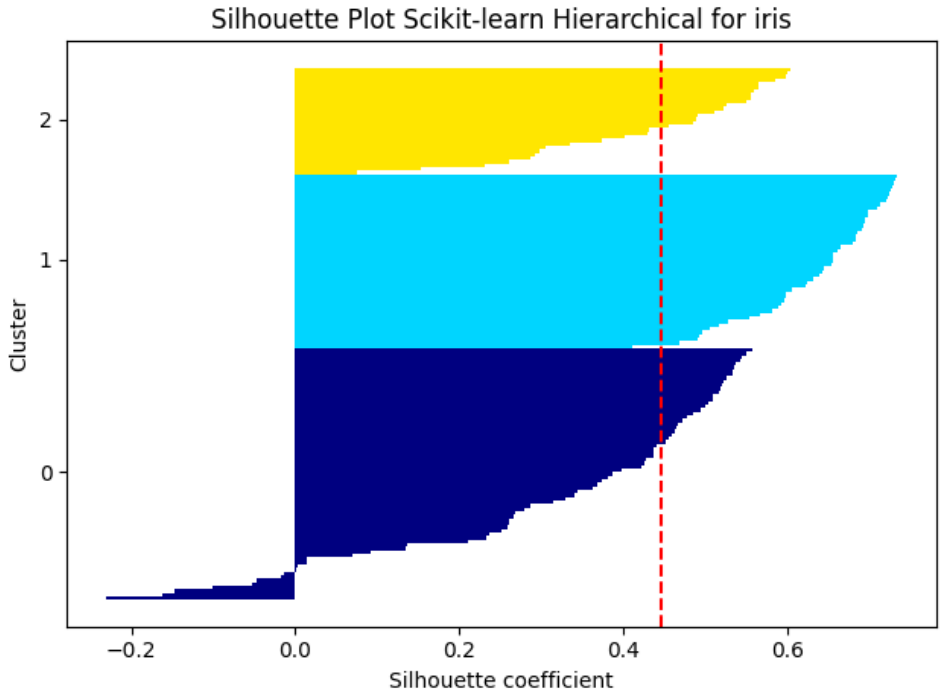

Dendrogram for mnist (Complete linkage)
Euclidean distance
0
50
100
150
200
250
300

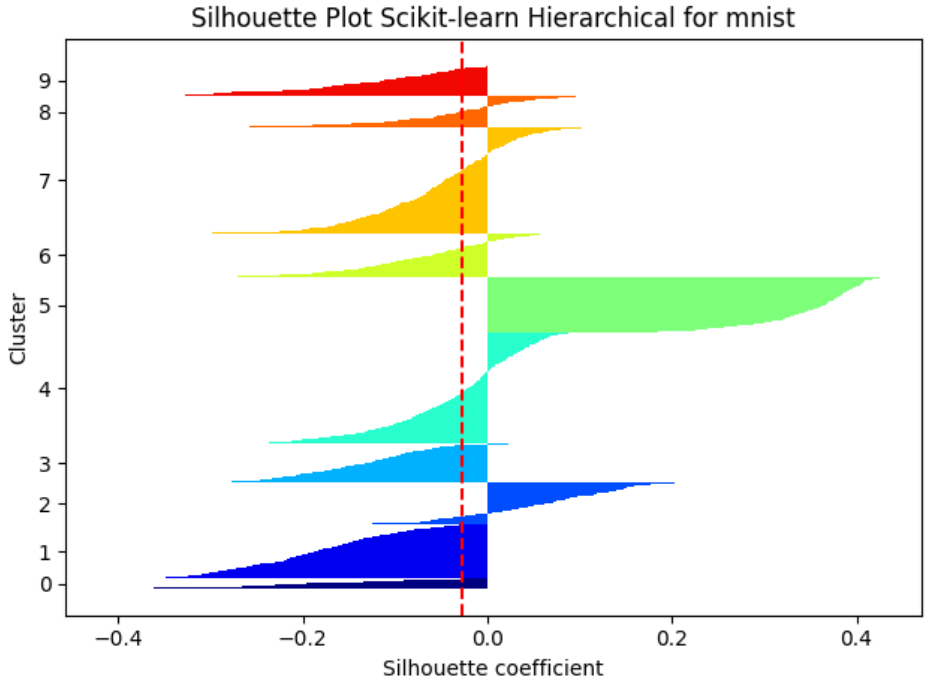

Silhouette Plot K-means for iris
Cluster
2
1
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Silhouette coefficient


Silhouette Plot K-means for mnist
Cluster
9
8
7
6
5
4
3
2
1
0
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
Silhouette coefficient


Silhouette Plot SciPy Hierarchical for iris
Cluster
2
1
0
−0.2
0.0
0.2
0.4
0.6
Silhouette coefficient


Silhouette Plot SciPy Hierarchical for mnist
Cluster
9
8
7
6
5
4
3
2
1
0
−0.4
−0.2
0.0
0.2
0.4
Silhouette coefficient


Silhouette Plot Scikit-learn Hierarchical for iris
Cluster
2
1
0
−0.2
0.0
0.2
0.4
0.6
Silhouette coefficient


Silhouette Plot Scikit-learn Hierarchical for mnist
Cluster
9
8
7
6
5
4
3
2
1
0
−0.4
−0.2
0.0
0.2
0.4
Silhouette coefficient