**COMP257 SECTION 003**

**Assignment 3: Hierarchical Clustering**

**Github:** <https://github.com/venussirada/COMP257/blob/main/Assignment3/SiradaThoungvitayasutee_COMP257_Assignment3.ipynb>

**Output**

1. Olivetti faces dataset

A collage of different faces

Description automatically generated

1. Dendrogram of Clustering Result

A graph of different colored lines

Description automatically generated

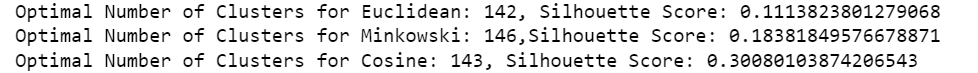
A diagram of a graph

Description automatically generated with medium confidence

A graph of different colored lines

Description automatically generated

1. Optimal Clusters



1. Silhouette Score for Euclidean Distance

A graph with blue lines

Description automatically generated

1. Silhouette Score for Minkowski Distance

A graph with blue lines

Description automatically generated

1. Silhouette Score for Cosine Similarity

A graph with blue lines

Description automatically generated

1. Discrepancies between Euclidean distance, Minkowski distance and Cosine similarity

Answer All three approaches – Euclidean Distance, Minkowski Distance, and Cosine Similarity – identify the optimal number of clusters in the range of 140 clusters. However, there is a difference clustering quality based on the silhouette scores:

* Cosine Similarity produces the best silhouette score of 0.3008, indicating that it has the most well-defined clusters among the three models.
* Minkowski Distance performs better than Euclidean Distance, with a silhouette score of 0.1838.
* Euclidean Distance, with a silhouette score of 0.1114, has the least score among the three measures.

1. Reduced Dataset Size

A screenshot of a computer

Description automatically generated

1. Cross Validation Score

A white background with numbers

Description automatically generated

1. Validation Accuracy

A close up of words

Description automatically generated

**Code Explanation**

Firstly, I began by loading Olivetti faces dataset and used Stratified Split to ensure that training, validation and testing dataset have the same number of images for each label.

Then I trained the Logistic Regression classifier and received the validation accuracy of 95%. After that, I used Agglomerative Clustering, along with centroid-based rules, to reduce the dimensionality of the dataset. I experimented it with Euclidean Distance, Milkowski Distance and Cosine Similarity.

Then I computed Silhouette Score for each approach to retrieve the optimal number of clustering. The result shows that Cosine Similarity produces the best Silhouette Score of 0.3.

Lastly, I used the optimal number of clusters to train the classifier for each approach and received the validation accuracy score (refer to image in output 9 and 10 above).

**Lesson Learned**

* AgglomerativeClustering in sci-kit learn does not support ‘centroid’ linkage. As a result, scipy’s linkage function is used for centroid-based clustering.
* When using Minkowski Distance with scipy’s pdist function, it is important to set p = 1 to ensure the use of Manhattan distance. Otherwise, the function will use the default Euclidean distance.

**References:**

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.dendrogram.html>

<https://scikit-learn.org/dev/modules/generated/sklearn.cluster.AgglomerativeClustering.html>

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html>

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.pdist.html#scipy.spatial.distance.pdist>

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.fcluster.html>