**Assignment 3: Hierarchical Clustering**

Dinh Hoang Viet Phuong - 301123263

# Import all necessary libraries

A screen shot of a computer

Description automatically generated

**# 1. Retrieve and load the Olivetti faces dataset.**

A computer screen shot of a code

Description automatically generated

A collage of faces of men and women

Description automatically generated

**# 2. Split the training set, a validation set, and a test set using stratified sampling to ensure that there are the same number of images per person in each set.**

A computer code with white text

Description automatically generated

First, I split the original dataset into a temporary training set (X\_temp, y\_temp) and a test set (X\_test, y\_test). The test\_size=0.2 means that the test set will contain 20% of the original data.   
Then, I split the temporary training set into a final training set (X\_train, y\_train) and a validation set (X\_val, y\_val). The test\_size=0.25 means that the validation set will contain 25% of the temporary training set, which is 15% of the original data.  
As a result, the final splits will contain 60% of the data in the training set, 20% in the validation set, and 20% in the test set. The use of stratified sampling ensures that each set has the same distribution of face IDs as the original dataset.

Training set size: 240  
Validation set size: 80  
Test set size: 80

**# 3. Using k-fold cross validation, train a classifier to predict which person is represented in each picture, and evaluate it on the validation set.**

A computer screen shot of text

Description automatically generated

In the context of the Olivetti faces dataset, which consists of grayscale images of faces, SVM might work well because faces have consistent structures (eyes, nose, mouth) that a linear classifier might capture. The choice of SVM here was primarily for simplicity and demonstration.

Cross-validation scores: [0.95833333 0.9375 0.97916667 0.89583333 0.875]  
Mean accuracy: 0.9291666666666668  
Standard deviation: 0.038640077064565424

The 5-fold cross-validation scores on the training set range between 0.875 and 0.9792 with a mean accuracy of 0.9292. This indicates that the SVM classifier with a linear kernel performs very well on the training set. The validation set accuracy is even higher at 0.975. This suggests that the model generalizes well and is not overfitting to the training data.

**# 4. Using either Agglomerative Hierarchical Clustering (AHC) or Divisive Hierarchical Clustering (DHC) and using the centroid-based clustering rule, reduce the dimensionality of the set by using the following similarity measures:**

A screen shot of a computer program

Description automatically generated

Sample clusters with Euclidean distance:   
[ 4 29 9 35 18 0 27 34 30 4 2 22 15 19 17 13 3 2 21 29]

Sample clusters with Manhattan distance:   
[ 0 33 1 33 34 1 6 12 3 0 10 21 23 16 3 2 13 2 5 33]

Sample clusters with Cosine similarity:   
[ 2 21 6 10 1 6 17 1 28 2 4 32 6 12 28 6 24 6 11 26]

For each of the distance metrics (Euclidean, Manhattan, and Cosine), the sample clusters differ, highlighting the differing nature of these metrics in the context of hierarchical clustering.

**# 5. Use the silhouette score approach to choose the number of clusters for 4(a), 4(b), and 4(c)**

A screen shot of a computer program

Description automatically generated

A graph with red lines

Description automatically generated

The silhouette scores give a perspective on the distance between the resulting clusters. More distance indicates better clusterings. The plot showcases how the silhouette scores vary with the number of clusters. By analyzing this, one can determine an optimal number of clusters for each distance metric. Higher silhouette scores for a certain number of clusters signify better-defined clusters.

**# 6. Use the set from (4(a), 4(b), or 4(c)) to train a classifier as in (3) using k-fold cross validation.**

A computer screen with text

Description automatically generated

Cross-validation scores: [0.4 0.4125 0.4125 0.3875 0.4]  
Mean accuracy: 0.4025  
Standard deviation: 0.009354143466934839

Using the cluster assignments from Euclidean distance-based clustering as a feature for the SVM classifier, the performance has significantly decreased. The mean accuracy is now only 0.4025 with 5-fold cross-validation. This is a considerable drop from the previous ~0.93 accuracy. The standard deviation is also relatively low, indicating that the performance is consistently low across the folds.