**Assignment 2: K-Means & DBSCAN Clustering**

**Dinh Hoang Viet Phuong – 301123263**

**Import Libraries:**

Various libraries from sklearn and numpy are imported to help with fetching datasets, modeling, evaluation.

**Step 1: Retrieve and Display Olivetti Faces**

fetch\_olivetti\_faces retrieves the Olivetti Faces dataset.

data.images gives the images, data.data gives the flattened image data (feature vectors), and data.target gives the target labels (person IDs).

The first 20 faces are displayed using matplotlib.

A collage of faces of men and women

Description automatically generated

**Step 2: Splitting the Dataset**

The data is split into training, validation, and test sets. Stratified sampling is used to ensure that there are equal numbers of images per person in each set.

The data is initially split into a temporary set and a test set (80/20 ratio).

The temporary set is further split into training and validation sets (75/25 ratio). This ensures an overall 60/20/20 split.

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

**Step 3: Train a Classifier with Cross-Validation**

A linear SVM (SVC) is chosen as a classifier.

Stratified K-Fold cross-validation is used to ensure each fold has an equal proportion of each target class.

Cross-validation scores for the training set are computed and printed.

The classifier is trained on the entire training set and evaluated on the validation set.

Cross-validation scores: [0.89583333 0.9375 0.9375 0.97916667 0.89583333]  
Average cross-validation score: 0.9291666666666666  
Validation set score: 0.975

**Step 4: Dimensionality Reduction with K-Means**

The optimal number of clusters for K-Means is determined by the silhouette score. A higher silhouette score indicates better-defined clusters.

Silhouette scores for different numbers of clusters (from 2 to 50) are calculated.

These scores are plotted against the number of clusters to visually identify the optimal number.

The training data is then transformed using the optimal number of clusters.

A graph with blue dots and numbers

Description automatically generated

The optimal number of clusters is: 49

**Step 5: Train Classifier on Reduced Data**

A similar linear SVM classifier is trained on the reduced (transformed by KMeans) training data.

The process (from cross-validation to validation set evaluation) is analogous to step 3 but using the transformed data.

Cross-validation scores (reduced data): [0.8125 0.83333333 0.8125 0.83333333 0.8125 ]  
Average cross-validation score (reduced data): 0.8208333333333334  
Validation set score (reduced data): 0.85

**Step 6: Clustering with DBSCAN**

Data normalization is applied using StandardScaler to ensure features have a mean of 0 and a standard deviation of 1. This is essential for many distance-based algorithms, including DBSCAN.

DBSCAN clustering is applied with cosine similarity as the distance metric. The choice of cosine similarity is often justified for image data, as it measures the cosine of the angle between two vectors. This makes it less sensitive to the magnitude (brightness variations) and more focused on the patterns (like facial features).

The unique cluster labels (clusters found by DBSCAN) are printed.

Unique cluster labels: [-1 0 1 2 3 4 5]