INTRODUCTIO TO DATA SCIENCE

2nd Homework Assignment

Due on: April 29, 2024

This homework explores Principal Component Analysis (PCA) for facial recognition and generation using CelebFaces Attributes (CelebA) Dataset.

Tasks:

1. Data Preparation

To prepare the data, you will first load the facial images from the dataset. You will then ensure all images have a consistent size to simplify processing. Next, you will convert the images from color to grayscale, as facial recognition algorithms often focus on luminance information. Finally, you will transform each image into a one-dimensional vector. This vector will represent the intensity of each pixel in the image, creating a numerical representation suitable for further analysis.

2. PCA Analysis

To analyze the facial data, you will first arrange all the preprocessed face image vectors into a single matrix. Each row in this matrix will represent a unique face. To focus on the variations between faces, you will center the data by subtracting the average face vector (obtained by averaging all the image vectors) from each row. Then, you can calculate the covariance matrix, which captures how these centered faces vary in relation to each other. Finally, you will leverage scikit-learn's PCA function. This function will identify the key directions (eigenvectors) of greatest variance within the data, which are called principal components.

3. Face Reconstruction

To capture the key variations in the faces, you will project the centered data matrix onto a lower-dimensional space spanned by the most informative principal components. Use your face image for face reconstruction. The number of components used ('*n*') will determine the balance between detail and accuracy in the reconstruction. You will then utilize the projection coefficients and these chosen components to recreate the original

faces. By visualizing both the original and reconstructed versions, you can analyze how well the essential facial features are preserved as we vary '*n*' (e.g., *n* is 20, 30, 40, 50). This will help you understand the impact of dimensionality reduction on the reconstruction quality.

4. Project Your Face

Find the celebrity in the dataset whose facial representation in PCA space is closest to your face.

5. Random Face Generation

Utilize the existing PCA components captured from the facial data and generate random face vectors by manipulating the principal components within a controlled range. Project the manipulated vectors back into the original data space for visualization.

6. Evaluation

Assess the reconstruction accuracy using metrics like mean squared error between original and reconstructed faces. Evaluate the quality and realism of the generated faces.

Submission:

Submit your Python code or Jupyter Notebook in a file named your-student-id.py (e.g. 490606-hw2.py or 490606-hw2.ipynb) through Itslearning. Please upload the Python/Notebook file only.

IMPORTANT

- Academic dishonesty, including but not limited to cheating, plagiarism, and collaboration, is unacceptable and subject to disciplinary action. Any student found guilty will have a grade of F. Assignments are due in class on the due date. Late assignments will generally not be accepted. Any exception must be approved. Approved late assignments are subject to a grade penalty.