Week 10 - PCA

DS3010 - Introduction to Machine Learning

Instructions

1. Provide commented, indented code. Variables should have meaningful names.
2. Submit one .ipynb file containing all answers. The name should be [student name][roll_number] assignment[number].ipynb
3. Read the questions carefully before answering. If a question asks to follow a particular approach or to use a specific data structure, then it must be followed.
4. Write questions in separate text blocks in Jupyter Notebook before the code blocks containing answers.
5. All plots should have appropriate axis labels, titles, and legends.
6. Late submissions will not be accepted.
7. Do not use any library functions unless it is mentioned in the questions.
8. Download the data from the following link.

Tasks for the Lab

1. Load the dataset and visualize (2 Points)

    a. Load the given ‘emnist-letters-train.csv’ file using Pandas.

    b. Extract the data of any single label of your choice. Use this data for the questions given below.(e.g Choose a label 1 for 'a')

    c. Split the data into features and targets.Convert the features to images

    d. Visualize the extracted features using the function given to you.

2. Get the eigen-values and eigen-vectors (4 Points)

    a. Find the mean and demean or centralize the data (subtracting mean from images).

b. Reshape the demeaned data to 2D (num_samples, num_features).Use the demeaned data to get the covariance matrix.

c. Find the eigen-values and eigen-vectors using the covariance matrix. Sort

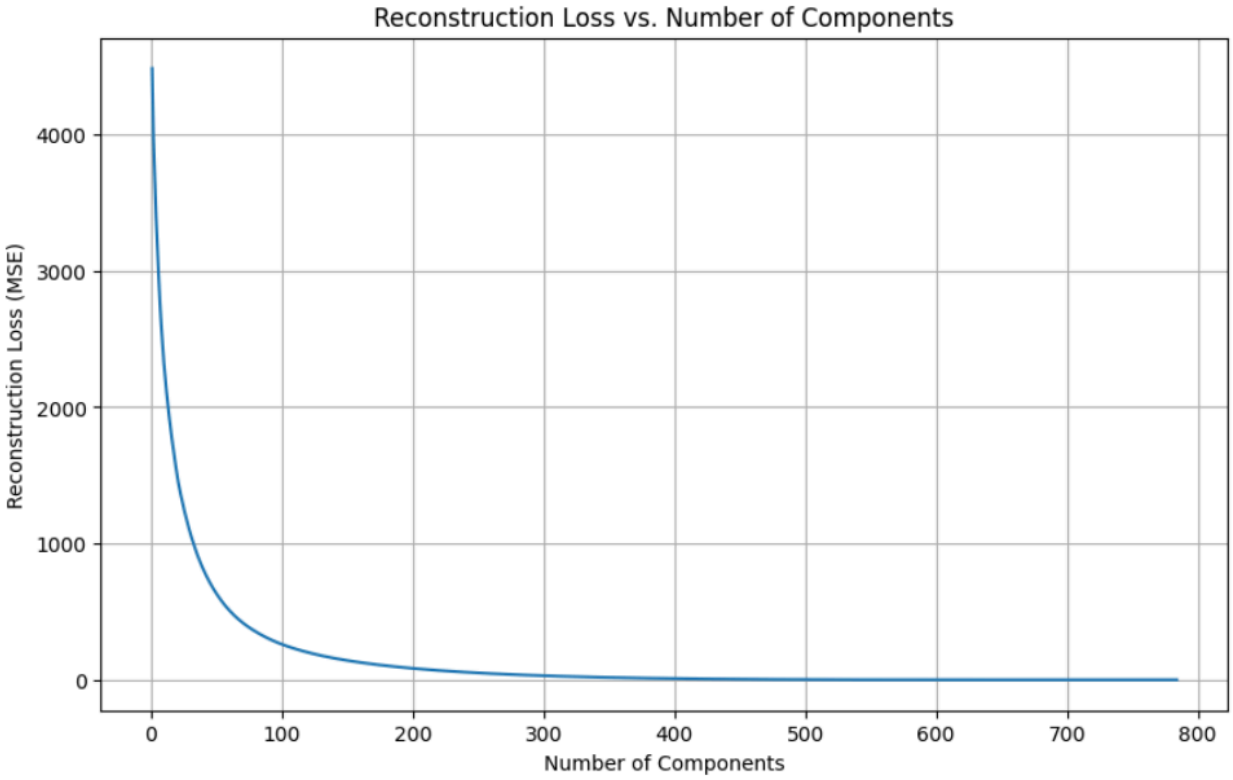
them in descending order.

d. Plot the eigen-values corresponding to dimension.

3. Find the optimal number of components (6 Point)

a. Project the data into lower dimensions and calculate the reconstruction loss using the flattened demeaned data.

b. Plot the reconstruction loss corresponding to the number of components.



4. Perform PCA with optimal number of components (3 Point)

a. Select the optimal number of components from the plot and perform PCA to get the reduced feature vectors.

b. Get the reduced feature vectors.

c. Perform reconstruction to get back the original features. Reshape the reconstructed data back to the original image shape.

5. Visualize and write your observations (5 Point)

a. Plot the images using reduced features.

b. Plot the images after reconstructing the features.

c. Write your observation about the reconstructed images and original images.