SML Assignment 2

Sakshat Sachdeva

February 2025

1 Introduction

The following is the report of all the parts in the SML assignment. The data was extracted from the following link: MNIST Data Set.

2 Data Processing

This section describes the data preprocessing steps performed on the MNIST dataset. The dataset contains images of handwritten digits from 0 to 9. For this assignment, we focused on the subset of data containing only the digits 0, 1, and 2.

2.1 Loading and Filtering

The MNIST dataset was loaded using a custom data loader class, `MnistDataloader`, which reads the image and label files in their original format. We then filtered the training data to include only images of digits 0, 1, and 2. A random selection process was used to ensure a balanced data set of 100 samples per class, resulting in a total of 300 training images.

2.2 Image Conversion and Normalization

Each 28×28 image was converted into a feature vector of length 784 by stacking the columns of the image. These feature vectors were then normalized to have pixel values in the range of 0 to 1 by dividing each pixel value by 255.0.

2.3 Data Storage

The resulting preprocessed data was store in`dataSet.npz` using the NumPy `savez` function. This file can then be easily loaded for subsequent steps in the assignment.

2.4 Code Snippet

The code can be found in the Github repository:

3 Report and Analysis

3.1 Performac of LDA/QDA after PCA/FDA

We had to report the performance of Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) on the data after applying dimensionality reduction using PCA or FDA. In particular, the results were recorded for the following three conditions[1]:

- **PCA with Variance of 95%:**
  The dimension chosen for PCA is 80.
  LDA after PCA Test Accuracy: 97.67%
  QDA after PCA Test Accuracy: 99.00%
  LDA after PCA Train Accuracy: 99.00%
  QDA after PCA Train Accuracy: 100.00%

- **PCA with Variance of 90%:**
  The dimension chosen for PCA is 49.
  LDA after PCA Test Accuracy: 96.67%
  QDA after PCA Test Accuracy: 98.33%
  LDA after PCA Train Accuracy: 97.67%
  QDA after PCA Train Accuracy: 100.00%

- **FDA for 3 Classes:**
  (Using two discriminant directions for 3 classes)
  LDA after FDA Test Accuracy: 94.00%
  QDA after FDA Test Accuracy: 95.33%
  LDA after FDA Train Accuracy: 96.00%
  QDA after FDA Train Accuracy: 96.67%

Both the training and test sets used in these experiments are of dimension (300×784).

3.2 Analysis of PCA on Classification Performance

PCA was applied to the data to reduce its dimensionality before classification, and the resulting classification accuracies were evaluated for various numbers of principal components. The results obtained were as follows:

- **3 dimensions:** 93%
- **4 dimensions:** 92.67%
- **10 dimensions:** 96%
- **25 dimensions:** 96.67%
- **49 dimensions:** 96.67%

[1] The results are after one iteration. May vary with the flattening changing everytime.

- **80 dimensions:** 97.67%

These results indicate that as the number of principal components increases, the classification accuracy generally improves.
With very low dimensions (e.g., 3 or 4), a significant amount of discriminative information is lost, resulting in lower accuracies. As more components are retained, more of the original variance is captured, leading to an improvement in classification performance. However later on, the improvements starts to reduce (around 50 dimensions). This ensures our hypothesis that the amount of information retained in following Eigen matrices of U is lesser than the ones prior to it (It is sorted descending)

3.3 Visualisation of FDA and PCA

The graph below is for PCA for 95%:

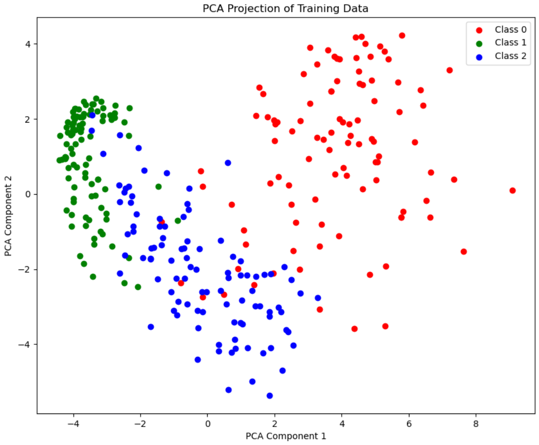

Figure 1: PCA Projection of Training Data (95% Variance).

This shows us that the data is well scattered and can be separated. Some parts of class 0 lie in class 2 but are far from class 1.

The graph in Figure 2 is for PCA with a variance of 90%.

The graph is very similar to the 95% variance one. This similarity arises because most of the essential information is already captured, and the eigenvalues corresponding to the remaining components do not contribute significantly to the overall variance.

The graph in Figure 3 is for FDA.

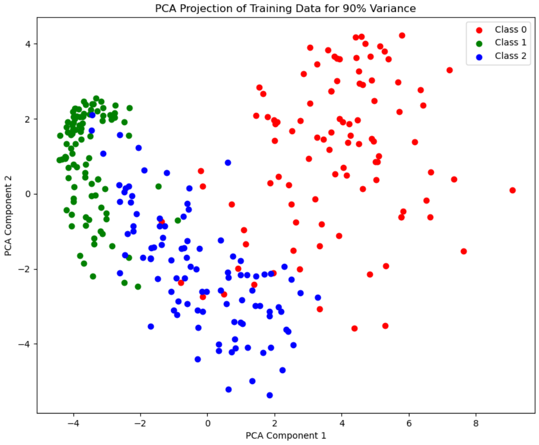

Figure 2: PCA Projection of Training Data (90% Variance).

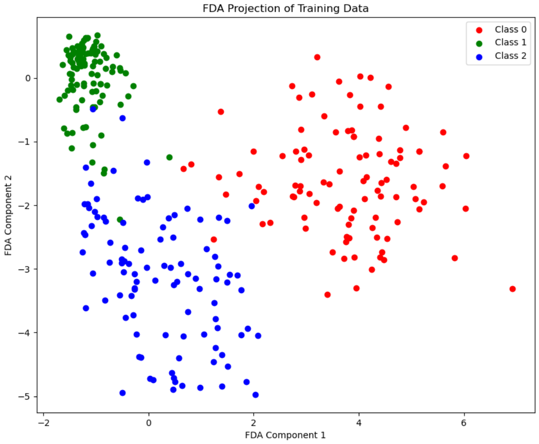

Figure 3: FDA Projection of Training Data.

The FDA graph shows that the data are much more distinctly separated. Unlike PCA, which focuses on capturing the overall variance in the data, FDA maximizes the ratio of between-class scatter to within-class scatter. This results in a projection where the classes are more clearly differentaited with near no overlap. In this projection, the distinct clusters indicate that the FDA transformation has effectively enhanced the separability of the classes.

4 Link for Repository

Link for Github: Classifier for Numbers 0-1-2 SML