MNIST Classification using MLE, PCA, FDA, and Discriminant Analysis

Deadline: 14th Friday 11:59pm

Problem Statement

You are given MNIST (https://www.kaggle.com/datasets/hojjatk/mnist-dataset), which consists of handwritten digit images (0-9). Your task is to build a classification pipeline using only the digits 0, 1, and 2. Randomly select 100 samples from train set of each class leading to a total of 300 samples. Call this as train set. Randomly sample 100 samples from test set for each class and denote it as test set. This exercise incorporates the following concepts:

1. **Maximum Likelihood Estimation (MLE)**: Estimate the mean and covariance matrix of each digit class assuming a Gaussian distribution.
2. **Principal Component Analysis (PCA)**: Reduce the dimensionality of the MNIST dataset while retaining 95% variance.
3. **Fisher's Discriminant Analysis (FDA)**: Find the optimal low-dimensional projection that maximizes class separation.
4. **Discriminant Analysis for Classification**: Use Linear Discriminant Analysis (LDA) or Quadratic Discriminant Analysis (QDA) to classify test samples.

Task Description

Implement the following steps:

1. **Data Preprocessing**
   - Load the MNIST dataset and filter only classes $0, 1$, and 2.
   - Convert images into feature vectors by stacking columns and normalize them to be in the range 0 to 1.
2. **Compute MLE Estimates** [1]

- Estimate the mean $\boldsymbol{\mu}_c$ and covariance matrix $\boldsymbol{\Sigma}_c$ for each digit class using MLE. Use only train set.
- Assume the data follows a multivariate Gaussian distribution:

$$P(\mathbf{x}|y=c) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_c|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1}(\mathbf{x}-\boldsymbol{\mu}_c)\right).$$

3. **Dimensionality Reduction using PCA** [2]

   - Apply PCA to reduce MNIST features to a lower-dimensional representation. Use only train set.
     For PCA:
     a. You need to obtain data matrix $X \in R^{784\times 300}$. Then obtain its mean μ.
     b. Remove the mean form X to obtain X_c.
     c. Obtain covariance $S = X_c X_c^\top/(300-1)$.
     d. Obtain eigenvectors and eigenvalues. Sort the eigenvectors in descending order of eigenvalues.
     e. Obtain U_p and perform $Y = U_p^\top X_c$.
     f. Later on for applying LDA, you will use Y where the question asks to apply PCA first.
     g. For a new test point x_{test}, do $y_{test} = U_p^\top x_{test}$. Plug y_{test} in the trained LDA/QDA and check for the discriminant which gives maximum value.
   - Retain 95% variance.

4. **Class Projection using Fisher's Discriminant Analysis (FDA) [will be covered on monday]** [2]

   - Use FDA to find the optimal projection direction for classification.
   - Compute the between-class scatter matrix $\mathbf{S}_B$ and within-class scatter matrix $\mathbf{S}_W$:

$$\mathbf{S}_B = \sum_c N_c(\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T$$

$$\mathbf{S}_W = \sum_c \sum_{\mathbf{x}_i \in C_c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^T.$$

   - Compute the optimal projection matrix $\mathbf{W}$ that maximizes class separability:

$$\mathbf{W} = \arg\max_{\mathbf{W}} \det(\mathbf{W}^T\mathbf{S}_B\mathbf{W})/\det(\mathbf{W}^T\mathbf{S}_W\mathbf{W}).$$

Note: You will be solving a generalized eigenvalue problem. We will see how to obtain between and within class scatter matrices. In the code, this needs to be explicitly computed.

5. **Evaluate and Compare Performance**
    - Apply FDA on the test set. Compute the classification accuracy of both LDA and QDA on the test set. Report the accuracy for both train and test set separately. [1]
    - Apply PCA and then apply LDA. Report the accuracy. [1]
    - Analyse the accuracy by changing the variance to 90%.
    - Analyse the accuracy by using only first two principal components. Report the accuracy for both train and test set separately. [1]

Your final report should include:

- Accuracy of LDA and QDA classifiers.
- Analysis of how PCA affects classification performance.
- A visualization of the transformed feature space. For both FDA and PCA, you will obtain 2D data, plot them by using different colors and markers. See if the samples from 3 classes look separable.

Accuracy Definition

Accuracy is a commonly used metric in classification tasks, measuring the proportion of correct predictions over the total number of predictions.

Mathematical Definition

Accuracy is defined as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{1}$$

If we define:

- N as the total number of test samples,
- $\hat{y}_i$ as the predicted label for the i-th sample,
- y_i as the true label for the i-th sample,

then the accuracy can be expressed as:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \nVdash(\hat{y}_i = y_i) \tag{2}$$

where $\nVdash(\hat{y}_i = y_i)$ is an indicator function that returns 1 if the prediction is correct and 0 otherwise.

Special Cases

1. **Perfect Accuracy:** If all predictions are correct:

$$\text{Accuracy} = 1.0 \quad (100\% \text{ correct}) \tag{3}$$

2. **Worst Case Accuracy:** If no predictions are correct:

$$\text{Accuracy} = 0.0 \quad (0\% \text{ correct}) \tag{4}$$

Considerations

- Accuracy is useful when classes are balanced.
- For imbalanced datasets, accuracy can be misleading. Alternative metrics like Precision, Recall, and F1-score should be considered.