2DLDA and Its Applications
on the MNIST handwritten digits Classification Problems
Xixi Lu and Terry Situ
San Jose State University

About This Study
In this study, we are going to investigate how the algorithms of (2D) matrix-based linear discriminant analysis (LDA) perform on the classification problems of the MNIST handwritten digits dataset, and to compare its performance to the traditional (1D) vector-based dimension reduction method: Principal Component Analysis (PCA).

Linear Discriminant Analysis (LDA) is most commonly used as dimensionality reduction technique in the pre-processing step for pattern-classification and machine learning applications. The goal is to project a dataset onto a lower-dimensional space with good class-separability in order to avoid overfitting (“curse of dimensionality”) and also reduce computational costs. The main ideas behind 2DLDA are that they are based on 2D matrices as opposed to the traditional LDA, which are based on 1D vector. [1]

With all the benefits come with 2DLDA, you may ask: “how does it perform with different classification methods (e.g. k nearest neighbors algorithm and LDA classification)?” or “is it better than 1D-PCA?”

We investigate these questions and analyze extensively them based on a large database of handwritten digits. The database contains 60,000 training images and 10,000 testing images. In addition, each image contains 28 x 28 pixels. [3]

The study is conducted in two parts:

1) We use 2DLDA to reduced the dimension of the MNIST handwritten digit dataset and applied different classification methods on the reduced dataset.

2) We compare the performance between 2DLDA and PCA (in term of test errors) by using different classifiers.

We compared the accuracy of those classifiers in term of test error rates.

How Does 2DLDA Work?

2DLDA transforms $r \times c$ images to smaller $r \times c'$ images. Let $X \in \mathbb{R}^{r \times c}$ be a given image. The transformation is defined by two matrices with orthonormal column, $L \in \mathbb{R}^{r \times r}$ and $R \in \mathbb{R}^{c \times c}$:

$$Y = X' LR$$

Like FDA, 2DLDA finds the best transformations $L, R$ by preserving the most discriminatory information in the projection space:

$$\max_{L, R} \quad \frac{\text{between-class scatter}}{\text{within-class scatter}}$$

And the between-class scatter and within-class scatter are defined as follow:

Within-class scatter:

$$\hat{S}^w = \frac{1}{n} \sum_{i=1}^{n} (X - M_i) (X - M_i)^T$$

Between-class scatter:

$$\hat{S}^b = \frac{1}{n} \sum_{i=1}^{n} n_i (M_i - \mu) (M_i - \mu)^T$$

Where $L \in \mathbb{R}^{r \times r}$ and $R \in \mathbb{R}^{c \times c}$ are tall matrices with orthonormal columns.

The iterative procedure:

1. Initialize $R = \left[ \begin{array}{c} \tilde{L} \\ \tilde{R} \end{array} \right] \in \mathbb{R}^{c \times c}$

2. Iterative until convergence:

   • $L =$ top $r$ eigenvecors of $(\hat{S}^b)^{-1} \hat{S}^w$

   • $\tilde{R} =$ top $r$ eigenvecors of $(\hat{S}^b)^{-1} \hat{S}^w$

$$\hat{S}^w = \sum_{i=1}^{n} n_i (M_i - \mu)(M_i - \mu)^T$$

3. Return final versions of $L$ and $R$.

Examples of Projected Displays by 2DLDA

The image on the left displays the first six images were projected here. The projected dimension is 15 x 15 for each image. Can our human recognize these images? (They are 5, 0, 4, 1, 9 & 2).

Investigations & Results: Case I 2DLDA + Local kMeans

For 2DLDA + kNN / Local Kmeans:

- The choice of distance functions will have a huge impact on the optimal combination of projected dimension and number of $k$;
- This is possible for lower projected dimension to yield higher accuracy rate;
- The higher the projected dimension will not always produce significant lower test error;
- 2DLDA vs. LDA;
- Both of LDA and QDA will produce lower min. test error rates PCA; LDA & Naïve Bayes does not yield good classifying results;
- Limitations:
  - Due to the resource constrains, we DID NOT implement exhaustive search to obtain the global optimal combination(s) of projected dimension and classifier for either 2DLDA or PCA;
  - Even though we had obtained the global optimal combination here, that may ONLY apply to MNIST dataset.

Conclusions & Limitations

References


Algorithm For 2DLDA

We need to obtain the optima $L$ and $R$ by solving the following equation:

$$\min_{L,R} \sum_{i=1}^{n} \|L(M_i - \mu)\|_F^2$$

Where $L \in \mathbb{R}^{r \times r}$ and $R \in \mathbb{R}^{c \times c}$ are tall matrices with orthonormal columns.

The iterative procedure:

1. Initialize $R = \left[ \begin{array}{c} \tilde{L} \\ \tilde{R} \end{array} \right] \in \mathbb{R}^{c \times c}$

2. Iterative until convergence:

   • $L =$ top $r$ eigenvecors of $(\hat{S}^b)^{-1} \hat{S}^w$

   • $\tilde{R} =$ top $r$ eigenvecors of $(\hat{S}^b)^{-1} \hat{S}^w$

$$\hat{S}^w = \sum_{i=1}^{n} n_i (M_i - \mu)(M_i - \mu)^T$$

3. Return final versions of $L$ and $R$.

Investigations & Results: Case II 2DLDA + kNN

For 2DLDA + kNN / Local Kmeans:

- The lowest test error (0.0272) reached when the dimension is 14 x 14 and $k = 10$.

- The lowest test error (0.0272) reached when the projected dimension is 14 x 14 and $k = 10$.

Investigations & Results: Case III 2DLDA vs. PCA

Investigations & Results: Case II 2DLDA + kNN

Conclusions & Limitations

- For 2DLDA + kNN / Local Kmeans:
  - The choice of distance functions will have a huge impact on the optimal combination of projected dimension and number of $k$;
  - This is possible for lower projected dimension to yield higher accuracy rate;
  - The higher the projected dimension will not always produce significant lower test error;

- 2DLDA + QDA:
  - Both of LDA and QDA will produce lower min. test error rates PCA; LDA & Naïve Bayes does not yield good classifying results;

- Limitations:
  - Due to the resource constrains, we DID NOT implement exhaustive search to obtain the global optimal combination(s) of projected dimension and classifier for either 2DLDA or PCA;
  - Even though we had obtained the global optimal combination here, that may ONLY apply to MNIST dataset.

Conclusions & Limitations

References


