CSE342: Statistical Machine Learning

Assignment 2

Name: Aditya Aggarwal **Roll Number:** 2022028

Problem 1

ASSUMPTIONS

1. The data is provided in the MNIST dataset format, where images are represented as matrices of pixel values.
2. The dataset contains labelled images of handwritten digits from 0 to 9.
3. Each label (digit) has approximately the same number of samples.
4. The determinant of the covariance matrix for each label is approximately same and tending to zero.

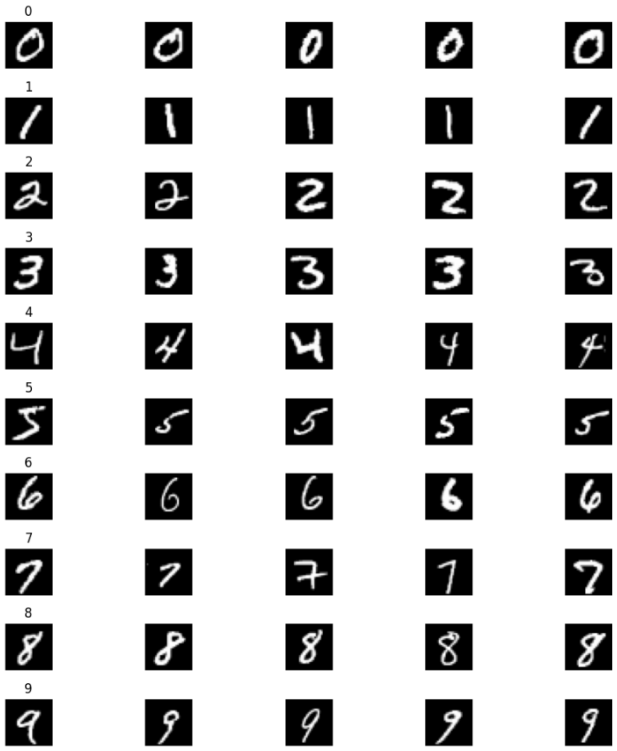
APPROACH

1. **Loading the Data:** The MNIST dataset is loaded from the provided URL using the *urllib.request.urlretrieve()* function. The dataset is then unpacked and loaded into NumPy arrays. The train set contains 60000 images and the test set contains 10000 images. All images are 784 by 784 pixels and labelled.
2. **Exploratory Data Analysis:** Basic exploratory analysis is performed to understand the structure and size of the dataset. This includes checking the number of unique labels, the size of the training set, and the number of pixels in each image.
3. **Data Preprocessing:** The images are vectorized into a 2D array to prepare them for further processing. The mean and inverse covariance matrix for each label is computed using the training data.
4. **QDA Classifier Implementation:** The Quadratic Discriminant Analysis (QDA) classifier is implemented with the *QDA()* and *max_QDA()* functions. The QDA function computes the QDA score for each label using the mean and covariance of the label. It also uses priori probabilities. It utilizes *np.pinv* to calculate the pseudo-inverse of the covariance matrix. Since the determinant of the matrix is zero, taking the pseudo-inverse is calculated by adding a negligibly small value to the diagonal elements and using the resultant determinant in the inverse.
5. **Evaluation:** The accuracy of the QDA classifier is evaluated on the test set. For each test sample, the label which produces the maximum QDA score is assigned as the predicted label. The predicted label is compared to the true label, and the accuracy is computed as the ratio of correctly classified samples to the total number of test samples.

RESULTS

Overall Accuracy: The QDA classifier achieved an overall accuracy of **85.72%** on the test set.

Per-Label Accuracy: The accuracy for each digit label varies, with some classes achieving higher accuracy than others. These accuracies provide insights into the classifier's performance on individual digits.



DATA VISUALISATION

```
Label 0: 93.47%
Label 1: 67.40%
Label 2: 93.60%
Label 3: 87.52%
Label 4: 90.94%
Label 5: 79.93%
Label 6: 89.04%
Label 7: 86.38%
Label 8: 88.81%
Label 9: 82.16%
```

PER-LABEL ACCURACY

CONCLUSION

The QDA classifier shows promising results for MNIST digit recognition, achieving an overall accuracy of 85.72%. However, there is variation in accuracy across different digit classes, suggesting that the classifier may struggle with certain digits more than others.

```
correct classifications: 1  total samples tested: 1
correct classifications: 2  total samples tested: 2
correct classifications: 3  total samples tested: 3
correct classifications: 4  total samples tested: 4
correct classifications: 5  total samples tested: 5
correct classifications: 6  total samples tested: 6
correct classifications: 7  total samples tested: 7
correct classifications: 8  total samples tested: 8
correct classifications: 9  total samples tested: 9
correct classifications: 10  total samples tested: 10
correct classifications: 11  total samples tested: 11
correct classifications: 12  total samples tested: 12
correct classifications: 13  total samples tested: 13
correct classifications: 14  total samples tested: 14
correct classifications: 15  total samples tested: 15
correct classifications: 15  total samples tested: 16
correct classifications: 16  total samples tested: 17
correct classifications: 17  total samples tested: 18
correct classifications: 17  total samples tested: 19
correct classifications: 18  total samples tested: 20
correct classifications: 19  total samples tested: 21
correct classifications: 20  total samples tested: 22
correct classifications: 21  total samples tested: 23
correct classifications: 22  total samples tested: 24
correct classifications: 23  total samples tested: 25
...
correct classifications: 8570  total samples tested: 9998
correct classifications: 8571  total samples tested: 9999
correct classifications: 8572  total samples tested: 10000
Accuracy: 85.72%
```

OVERALL ACCURACY

Problem 2

ASSUMPTIONS

1. The dataset is loaded and pre-processed, ensuring that each digit label has 100 samples for training.
2. The frequency of images is approximately the same across each label.
3. Principal Component Analysis (PCA) is used to extract eigenfaces from the training data.
4. Quadratic Discriminant Analysis (QDA) is employed for classification based on the extracted features.

APPROACH

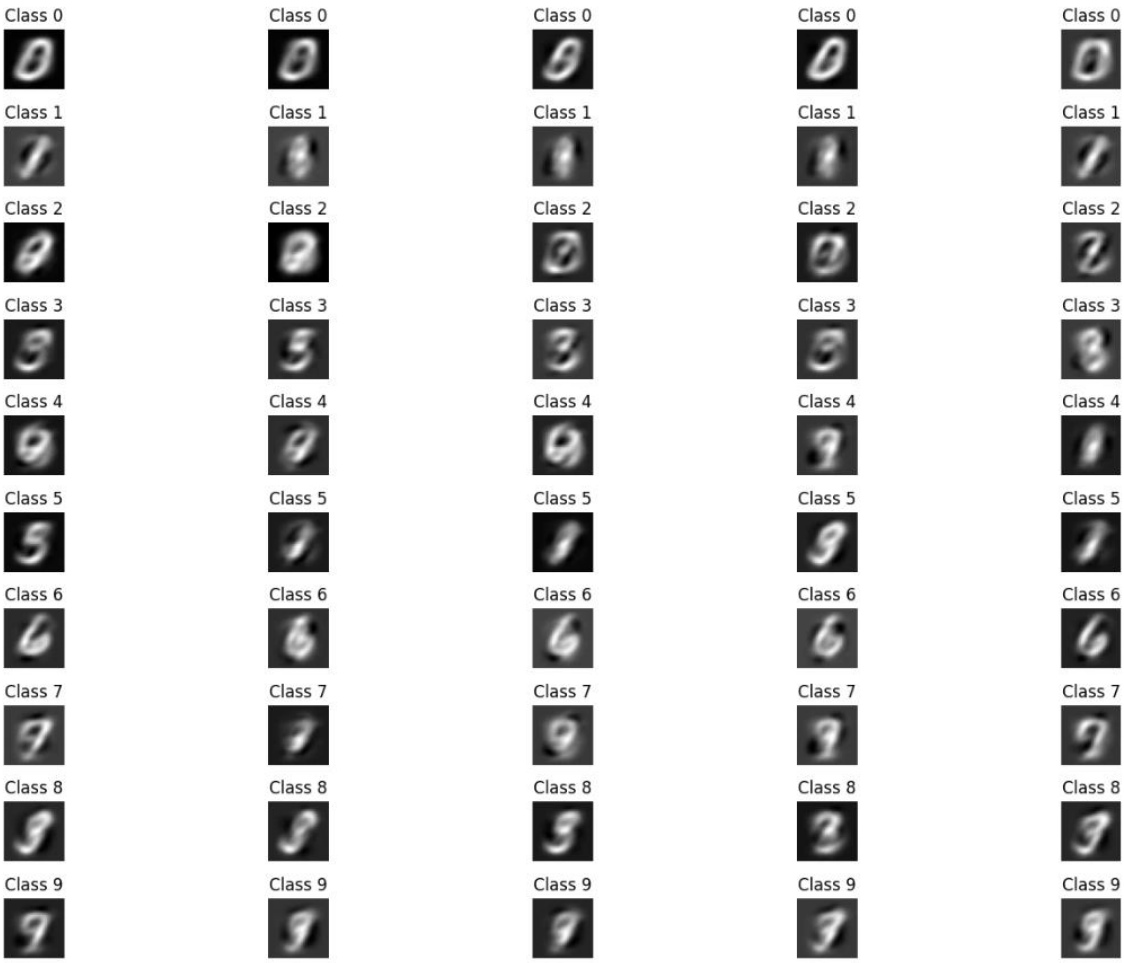
1. **Data Loading and Preprocessing:** The MNIST dataset is loaded, and 100 samples from each digit label are extracted for training. The images are reshaped and centralized by subtracting the mean image from the dataset.
2. **Eigenfaces Extraction:** Principal Component Analysis (PCA) is performed on the centralized training data to extract eigenfaces. The covariance matrix is computed, and its eigenvectors and eigenvalues are obtained. The eigenvectors corresponding to the largest eigenvalues represent the principal components or eigenfaces.
3. **Dimensionality Reduction:** The training data is projected onto the subspace spanned by the top 'p' higher priority eigenvectors (features) to obtain the feature vectors.
4. **Feature Selection:** It is noticed that the value of p directly influences quality of images plotted. As the value of p is increased from 5 to 784, the images are sharpened.
5. **QDA Classification:** Quadratic Discriminant Analysis (QDA) is applied to the projected feature vectors for classification. For each digit class, the mean and inverse covariance matrix are computed based on the projected training data. During testing, the QDA scores for each label are computed, and the label with the highest score is selected as the predicted label.
6. **Evaluation:** The classification accuracy is evaluated on the training set to assess the performance of the QDA classifier.

CONCLUSION

The Principal Component Analysis combined with Quadratic Discriminant Analysis on the MNIST digit classification dataset shows that by extracting eigenfaces and using QDA for classification, the model effectively learns discriminative features and achieves reasonable accuracy on the training set. Performing PCA and selecting prominent features before QDA increases its accuracy as opposed to simple QDA. The accuracy of the QDA classifier as well as the quality of images increases as the number of top features extracted (p) increases.

RESULTS

1. *p = 5*



DATA VISUALISATION FOR *p = 5*

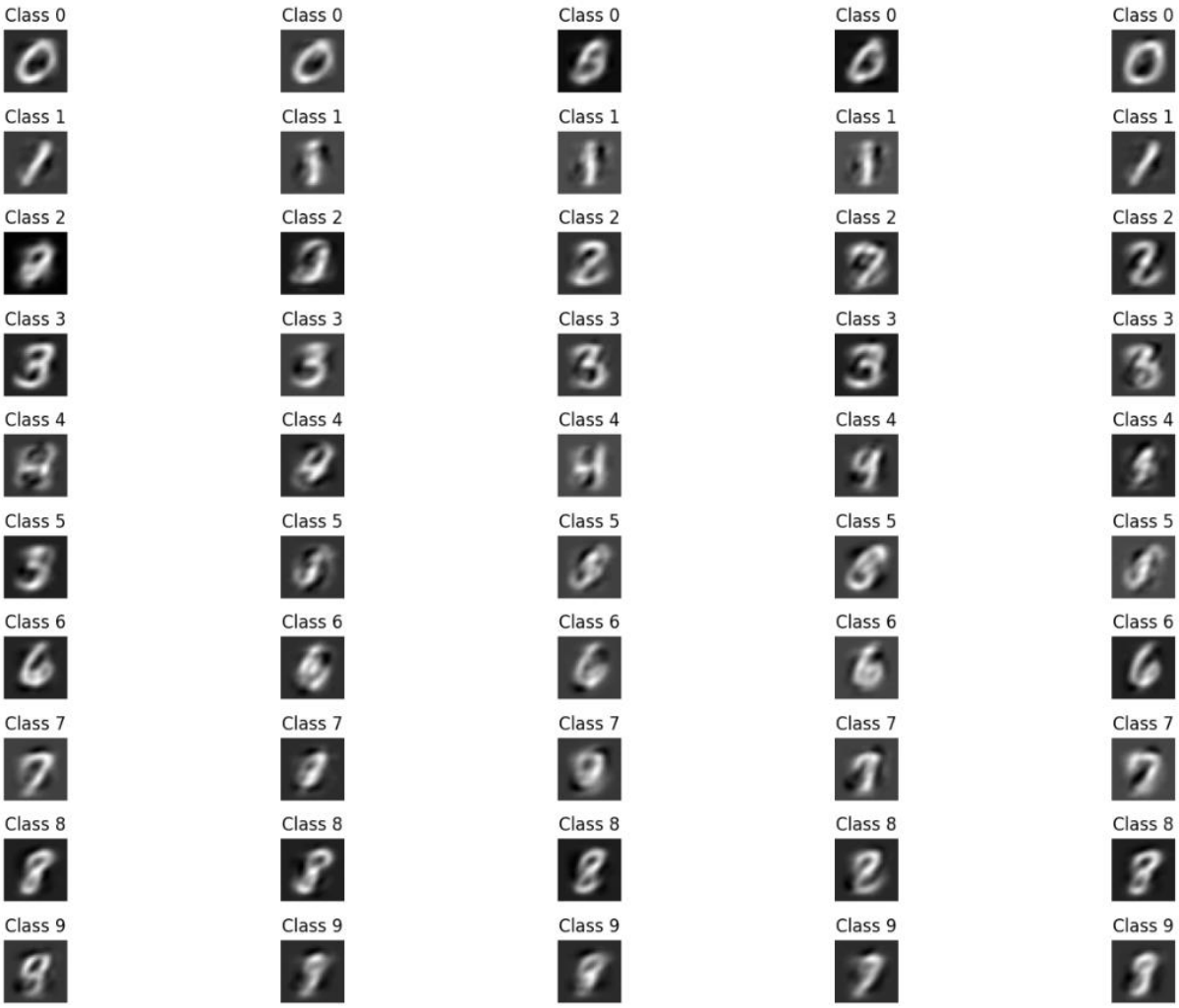
```
correct classifications: 7240  total samples tested: 9991
correct classifications: 7240  total samples tested: 9992
correct classifications: 7240  total samples tested: 9993
correct classifications: 7241  total samples tested: 9994
correct classifications: 7242  total samples tested: 9995
correct classifications: 7243  total samples tested: 9996
correct classifications: 7244  total samples tested: 9997
correct classifications: 7244  total samples tested: 9998
correct classifications: 7245  total samples tested: 9999
correct classifications: 7246  total samples tested: 10000
Accuracy: 72.46%
```

OVERALL ACCURACY FOR *p = 5*

```
Label 0: 89.18%
Label 1: 91.89%
Label 2: 82.27%
Label 3: 78.02%
Label 4: 59.16%
Label 5: 63.45%
Label 6: 86.01%
Label 7: 67.12%
Label 8: 40.45%
Label 9: 63.13%
```

PER- LABEL ACCURACY

2. *p = 10*



DATA VISUALISATION FOR *p = 10*

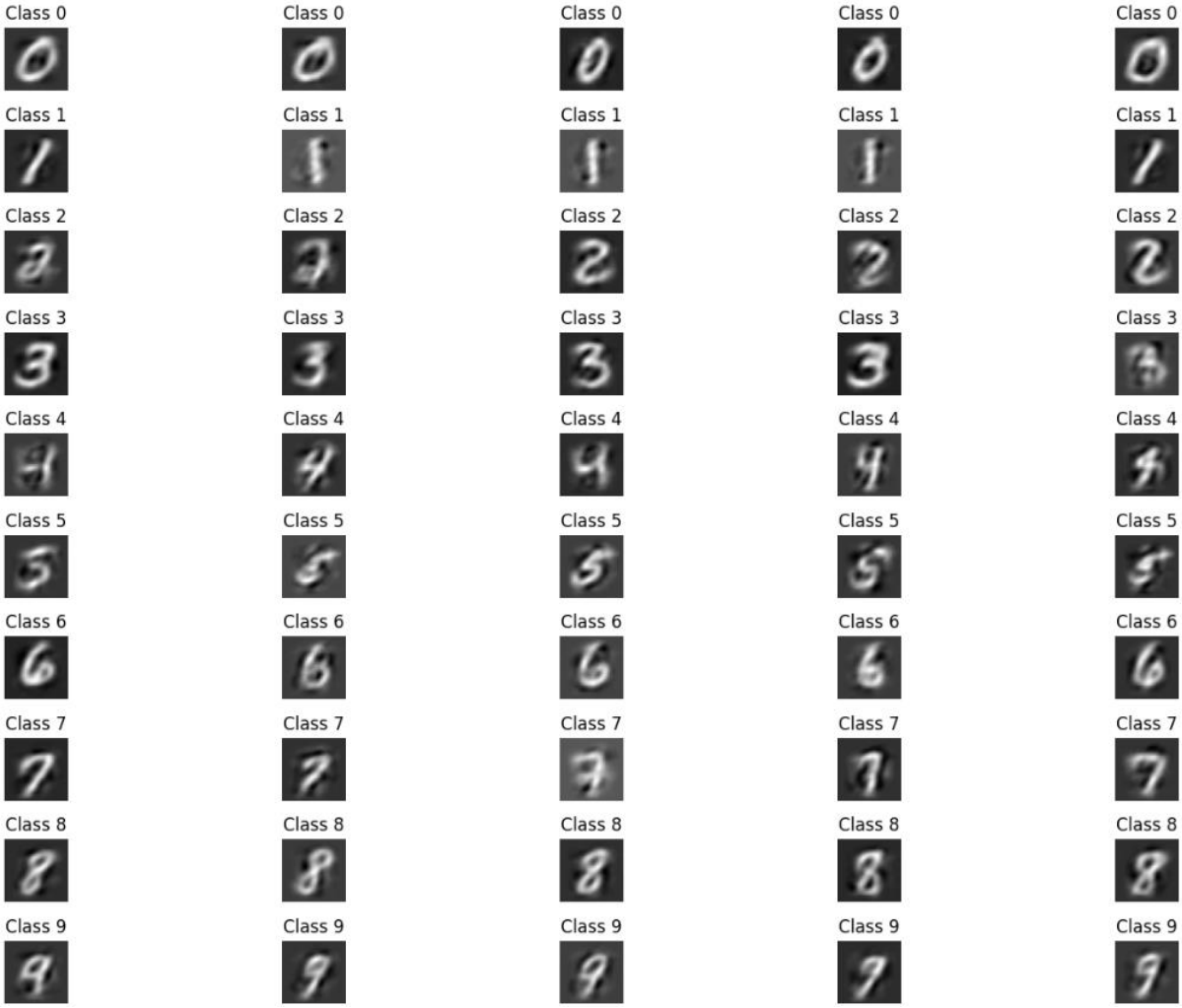
```
correct classifications: 8812  total samples tested: 9991
correct classifications: 8813  total samples tested: 9992
correct classifications: 8813  total samples tested: 9993
correct classifications: 8814  total samples tested: 9994
correct classifications: 8815  total samples tested: 9995
correct classifications: 8816  total samples tested: 9996
correct classifications: 8817  total samples tested: 9997
correct classifications: 8817  total samples tested: 9998
correct classifications: 8818  total samples tested: 9999
correct classifications: 8819  total samples tested: 10000
Accuracy: 88.19%
```

OVERALL ACCURACY FOR *p = 10*

```
Label 0: 97.14%
Label 1: 91.45%
Label 2: 93.22%
Label 3: 92.08%
Label 4: 86.25%
Label 5: 85.54%
Label 6: 89.56%
Label 7: 82.00%
Label 8: 81.93%
Label 9: 82.06%
```

PER- LABEL ACCURACY

3. p = 20



DATA VISUALISATION FOR *p = 20*

```
correct classifications: 9258  total samples tested: 9991
correct classifications: 9259  total samples tested: 9992
correct classifications: 9260  total samples tested: 9993
correct classifications: 9261  total samples tested: 9994
correct classifications: 9262  total samples tested: 9995
correct classifications: 9263  total samples tested: 9996
correct classifications: 9264  total samples tested: 9997
correct classifications: 9265  total samples tested: 9998
correct classifications: 9266  total samples tested: 9999
correct classifications: 9267  total samples tested: 10000
Accuracy: 92.67%
```

OVERALL ACCURACY FOR *p = 20*

```
Label 0: 98.37%
Label 1: 86.17%
Label 2: 96.32%
Label 3: 94.85%
Label 4: 95.52%
Label 5: 93.50%
Label 6: 91.86%
Label 7: 86.19%
Label 8: 93.63%
Label 9: 91.48%
```

PER- LABEL ACCURACY