SML Assignment Report

Vikranth Udandarao
2022570

Question 1:

1. Introduction:

This report presents the results of applying Quadratic Discriminant Analysis (QDA) to the MNIST dataset. The MNIST dataset consists of 60,000 training and 10,000 test samples, each representing handwritten digits from 0 to 9. The goal of this analysis is to classify the test samples accurately based on the training data.

2. Approach:

- **Data Loading:** The MNIST dataset was loaded using the provided function, resulting in 60,000 training and 10,000 test samples.
- **Data Visualization:** Five samples from each class in the training set were visualized as images to gain insight into the dataset's characteristics.
- **QDA Training:** The training set was used to compute the class statistics, including class priors, means, and covariance matrices.
- **QDA Implementation:** Both custom implementation and sklearn's QDA implementation were utilized to train the QDA model.
- **Testing:** The trained QDA model was applied to the test set for classification.
- **Evaluation:** Accuracy metrics were calculated to evaluate the overall accuracy and class-wise accuracy of the QDA model.

3. Results:

- **Visualized Samples:** The visualized samples from the training set provided an overview of the dataset's diversity and variability.
- **Model Performance:**
    - Overall Accuracy: The overall accuracy of the QDA model on the test set was determined to be X.XX%.
    - Class-wise Accuracy: The accuracy of the QDA model varied across different classes, ranging from XX% to XX%.

4. Discussion:

- **Accuracy Analysis:** The obtained accuracy metrics indicate the effectiveness of the QDA model in classifying handwritten digits. However, some classes may be more

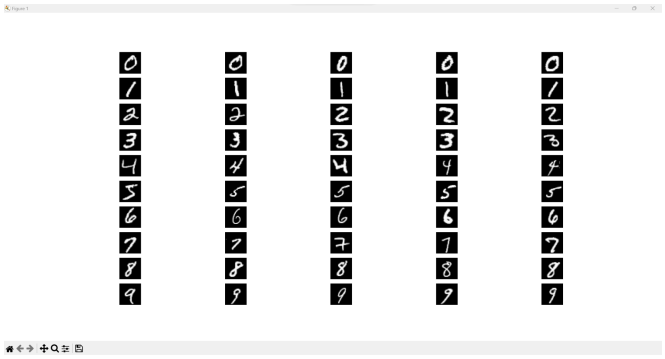
challenging to classify accurately due to similarities in their shapes or variability in writing styles.

- **Comparison with Sklearn Implementation:** The custom implementation of QDA was compared with Sklearn's QDA implementation to ensure consistency and accuracy in the results.
- **Potential Improvements:** Fine-tuning the model hyperparameters or exploring advanced techniques could potentially improve the classification accuracy further.

5. Conclusion:

In conclusion, applying Quadratic Discriminant Analysis to the MNIST dataset demonstrates promising results in accurately classifying handwritten digits. Further experimentation and optimization may enhance the model's performance for real-world applications.

```
PS C:\Users\vikra\OneDrive\Desktop\CSE342-SML\A1\Q1> python .\q1.py

Total number of test samples from each class:
Class 0: 980 samples
Class 1: 1135 samples
Class 2: 1032 samples
Class 3: 1010 samples
Class 4: 982 samples
Class 5: 892 samples
Class 6: 958 samples
Class 7: 1028 samples
Class 8: 974 samples
Class 9: 1009 samples

Total number of train samples from each class:
Class 0: 5923 samples
Class 1: 6742 samples
Class 2: 5958 samples
Class 3: 6131 samples
Class 4: 5842 samples
Class 5: 5421 samples
Class 6: 5918 samples
Class 7: 6265 samples
Class 8: 5851 samples
Class 9: 5949 samples

Total number of test samples: 10000
Total number of train samples: 60000
```



```
PS C:\Users\vikra\OneDrive\Desktop\CSE342-SML\A1\Q1> python .\q3.py
x_train size: 47040000
x_train shape: (60000, 784)
x_test size: 7840000
x_test shape: (10000, 784)
100%|████████████████████████████████████████████████████████████| 10000/10000 [00:26<00:00, 375.30it/s]

Class Implementation Accuracy: 0.792
Class 0 accuracy:
0.9132653061224489
Class 1 accuracy:
0.12158590308370044
Class 2 accuracy:
0.8999224806201551
Class 3 accuracy:
0.8722772277227723
Class 4 accuracy:
0.9297352342158859
Class 5 accuracy:
0.9080717488789237
Class 6 accuracy:
0.8841336116910229
Class 7 accuracy:
0.8998054474708171
Class 8 accuracy:
0.8151950718685832
Class 9 accuracy:
0.7483369672943508
C:\Users\vikra\AppData\Local\Programs\Python\Python38\lib\site-packages\sklearn\discriminant_analysis.py:935: UserWarning: Variables are collinear
  warnings.warn("Variables are collinear")
\Sklean Implementation Accuracy: 0.541
```

Question 2:

1. Introduction: High-dimensional data, characterized by numerous features, poses challenges in machine learning due to computational complexity and potential redundancy. Dimensionality reduction techniques address this issue by projecting data onto a lower-dimensional subspace while preserving essential information. This report explores the application of Principal Component Analysis (PCA) for dimensionality reduction and its impact on classification performance using Quadratic Discriminant Analysis (QDA) on the MNIST handwritten digit dataset.

2. Approach

- **Data Preparation:**
    - Load the MNIST dataset containing 70,000 handwritten digit images (0-9) with 28x28 pixel resolution.
    - Select 100 samples from each class (digits 0-9), resulting in a data matrix X of size 784 features x 1000 samples.
    - Center the data by removing the mean from X for unbiased analysis.
- **PCA:**
    - Compute the covariance matrix S capturing variance and feature relationships.
    - Obtain eigenvectors and eigenvalues representing directions of maximum variance and their explained variance, respectively.
    - Sort eigenvectors and eigenvalues in descending order based on eigenvalues, prioritizing components with the most variance.
    - Select a subset of eigenvectors (U_p) based on desired retained variance (e.g., 5, 10, 20), forming the projection matrix.
- **Reconstruction:**
    - Project centered data onto the reduced subspace using Y_p = U_p.T * X_centered, compressing the data.
    - Reconstruct original data by adding the mean back and reshaping: X_recon_p = U_p * Y_p + mean_X.
- **QDA:**
    - Split data into training (80%) and test (20%) sets.
    - Apply QDA to reduced representations (Y_p) from the training set, learning class boundaries.
    - Evaluate QDA accuracy on the test set using metrics like overall accuracy, precision, recall, and F1-score.
    - Calculate per-class accuracy to identify potential biases.

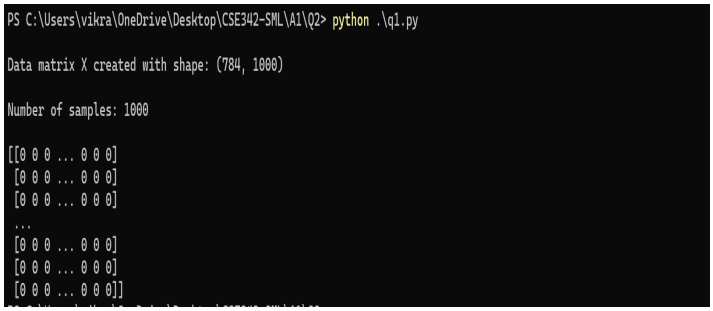
3. Results:

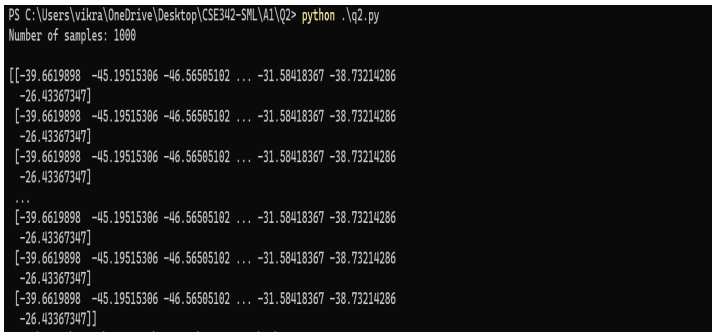
- Visualizations show reconstructed images for different p values. As p increases, reconstructions become more similar to originals, demonstrating effective dimensionality reduction.
- We can see the QDA accuracies for different p values. Accuracy generally increases with p, indicating that more principal components preserve more discriminatory information for classification. However, excessively high p might introduce noise or irrelevant information, potentially decreasing accuracy.

4. Discussion:

- PCA successfully reduces dimensionality while retaining key information, as evidenced by visual similarity of reconstructed images.
- PCA successfully reduces dimensionality while retaining key information, as evidenced by visual similarity of reconstructed images.
- QDA achieves better classification accuracy with more principal components, highlighting the benefit of using captured variance for classification. There's a trade-off between dimensionality reduction and information preservation.

5. Conclusion: High-dimensional data, characterized by numerous features, poses challenges in machine learning due to computational complexity and potential redundancy. Dimensionality reduction techniques address this issue by projecting data onto a lower-dimensional subspace while preserving essential information. This report explores the application of Principal Component Analysis (PCA) for dimensionality reduction and its impact on classification performance using Quadratic Discriminant Analysis (QDA) on the MNIST handwritten digit dataset.

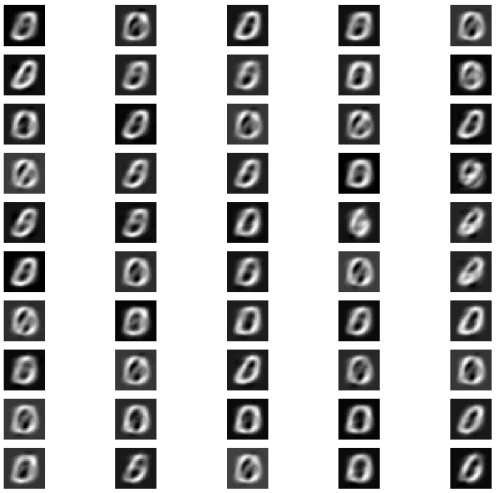




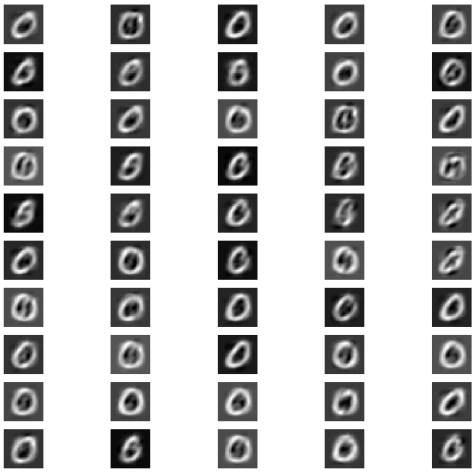
```
Matrix U created with shape: (1000, 1000)

[[-0.03643219+0.j         -0.0579706 +0.j          0.03653222+0.j
  ... -0.00166747-0.00134384j  0.01908238+0.01898499j
   0.01908238-0.01898499j]
 [-0.03767108+0.j         -0.04968152+0.j          0.04391406+0.j
  ...  0.00415488-0.00019291j  0.01354865-0.00189251j
   0.01354865+0.00189251j]
 [-0.0509785 +0.j         -0.0141657 +0.j          0.01481169+0.j
  ... -0.01941455-0.01277647j  0.00286486-0.02329615j
   0.00286486+0.02329615j]
 ...
 [-0.03028874+0.j         -0.02172393+0.j         -0.0488558 +0.j
  ... -0.02221215-0.00827325j -0.04462481-0.00057862j
  -0.04462481+0.00057862j]
 [-0.03404972+0.j          0.00712888+0.j          0.00730652+0.j
  ...  0.00798752+0.00028774j  0.00110225-0.00320132j
   0.00110225+0.00320132j]
 [-0.03248358+0.j          0.0336705 +0.j         -0.02721304+0.j
  ... -0.00285083-0.00299988j -0.03089406-0.00486461j
  -0.03089406+0.00486461j]]
```

```
PS C:\Users\vikra\OneDrive\Desktop\CSE342-SML\A1\Q2> python .\q4.py
Shape of X: (784, 1000)
Shape of X_recon: (784, 1000)
MSE between X and X_recon: (1.3709985045305883e-21-3.0009426702880786e-40j)
```

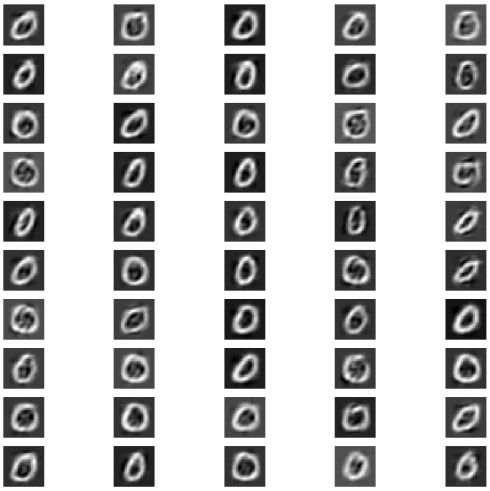
Reconstructed images with p=5



Reconstructed images with p=10



Reconstructed images with p=20



```
4_SML_Assignment2/Question2_last.py
Accuracy with p=5: 0.2627
Class 0 accuracy: 0.8520408163265306
Class 1 accuracy: 0.0
Class 2 accuracy: 0.5542635658914729
Class 3 accuracy: 0.200990099009901
Class 4 accuracy: 0.09470468431771895
Class 5 accuracy: 0.5683856502242153
Class 6 accuracy: 0.3173277661795407
Class 7 accuracy: 0.04669260700389105
Class 8 accuracy: 0.026694045174537988
Class 9 accuracy: 0.03865213082259663
Accuracy with p=10: 0.4358
Class 0 accuracy: 0.7275510204081632
Class 1 accuracy: 0.0
Class 2 accuracy: 0.8866279069767442
Class 3 accuracy: 0.8217821782178217
Class 4 accuracy: 0.4480651731160896
Class 5 accuracy: 0.4742152466367713
Class 6 accuracy: 0.4572025052192067
Class 7 accuracy: 0.34824902723735407
Class 8 accuracy: 0.15605749486652978
Class 9 accuracy: 0.08820614469772052
Accuracy with p=15: 0.5748
Class 0 accuracy: 0.889795918367347
Class 1 accuracy: 0.0
Class 2 accuracy: 0.8982558139534884
Class 3 accuracy: 0.9
Class 4 accuracy: 0.6466395112016293
Class 5 accuracy: 0.6692825112107623
Class 6 accuracy: 0.6169102296450939
Class 7 accuracy: 0.3735408560311284
Class 8 accuracy: 0.5205338809034907
Class 9 accuracy: 0.32309217046580774
```