M1 Project

Instructions

1. You are required to submit a report in PDF format that covers all the tasks provided in this project document.
2. The report should include sections and subsections for better clarity and evaluation of the methods used.
3. Ensure that the code files (.py or .ipynb) are submitted along with the report. Submissions without code will not be evaluated.
4. Plagiarism will be penalized. Please provide original work.
5. Include any visualizations or tables that support your analysis and findings.
6. Write your name on the first page of the report.

Abstract

This project focuses on the mathematical foundation for AI/ML using the MNIST dataset. We perform eigen decomposition, apply PCA for dimensionality reduction, and assess reconstruction error using PSNR.

1 Introduction

The MNIST dataset is a comprehensive collection of handwritten digits ranging from 0 to 9. It includes 60,000 grayscale images for training and 10,000 images for testing. Each image is 28 by 28 pixels, with pixel values ranging from 0 to 255. In this project, we focus on analyzing a subset of the test dataset and apply Principal Component Analysis (PCA) to reduce dimensionality.

2 Tasks

Task 1: Eigen Decomposition

Compute the eigen decomposition of the sample covariance matrix. Use the eigenvalues to calculate the percentage of variance explained. Plot the cumulative sum of these percentages versus the number of components.

Task 2: PCA for Dimensionality Reduction

Apply PCA via eigen decomposition to reduce the dimensionality of the images for each $p \in \{50, 250, 500\}$.

Task 3: Data Reconstruction

Using the reduced data from Task 2, reconstruct the original images. Use the property of orthonormal matrices for reconstruction.

Task 4: Error Comparison (PSNR)

Compare the error between the original and reconstructed images for 5 randomly selected images using Peak Signal-to-Noise Ratio (PSNR).

3 Methodology

3.1 Dataset

The MNIST dataset consists of 60,000 training and 10,000 testing images. In this project, we use the first 2000 samples from the test set. The images are scaled to the range [0, 1] by dividing each pixel value by 255.

3.2 Data Preprocessing

Each image is flattened into a vector of size 784 (28x28). A matrix $X \in \mathbb{R}^{2000\times784}$ is created, where each row represents a flattened image.

3.3 Eigen Decomposition

We compute the sample covariance matrix of X and perform eigen decomposition. The eigenvalues represent the variance explained by each principal component.

3.4 PCA

We reduce the dimensionality of the dataset using PCA for $p \in \{50, 250, 500\}$ and reconstruct the data using the principal components.

4 Results and Discussion

4.1 Variance Explained

The cumulative variance explained by the principal components shows that most of the variance can be captured with fewer components.

4.2 Data Reconstruction

The reconstructed images from different values of p show varying levels of detail. Higher values of p result in better reconstruction.

4.3 Error Comparison

The PSNR values indicate the reconstruction quality. Higher PSNR values suggest less reconstruction error.

5 Conclusion

In this project, we applied PCA to the MNIST dataset and successfully reduced the dimensionality while retaining most of the variance. The reconstructed images showed good quality with minimal error, as indicated by the PSNR values.

References

1. MNIST Dataset: `http://yann.lecun.com/exdb/mnist/`
2. Python Peak Signal-to-Noise Ratio (PSNR): `https://www.geeksforgeeks.org/python-peak-signal-to-noise-ratio-psnr/`