DIGIT CLASSIFICATION AND IMAGE COMPRESSION USING PCA

PROJECT PRESENTATION

Mohamed Achak

OVERVIEW

- INTRODUCTION
- PCA FOR IMAGE COMPRESSION
- LOGISTIC REGRESSION FOR CLASSIFICATION
- METHODOLOGY & IMPLEMENTATION
- MODEL EVALUATION (REPORT + CONFUSION MATRIX)
- METHMATHEMATICAL FOUNDATIONS
- VISUAL RESULTS: RECONSTRUCTION & PREDICTIONS
- CONCLUSION

INTRODUCTION

This project applies PCA and Logistic Regression to classify handwritten digits.
PCA reduces the image data from 64 to 30 key features by projecting it onto the directions of highest variance.
Then, logistic regression learns to predict digit labels using the compressed data.
This combination improves efficiency while maintaining high accuracy, supported by clear mathematical steps like centering, covariance, eigenvectors, and projection.

PCA FOR IMAGE COMPRESSION

Principal Component Analysis (PCA) is a linear technique for dimensionality reduction.
It transforms the data to a new coordinate system where the axes (principal components) capture the directions of greatest variance.
By projecting high-dimensional data onto fewer components, PCA compresses the dataset while preserving its essential structure.
This improves both computational performance and model interpretability, especially for image and pattern recognition tasks.

LOGISTIC REGRESSION

Logistic Regression is a simple and effective way to classify data into categories.
It draws a line (or boundary) in the data space that helps decide which group a new point belongs to.
In our case, after PCA compresses the digit images, logistic regression learns from examples by adjusting weights .
And then it turns them into probabilities to pick the most likely predection.

METHODOLOGY & IMPLEMENTATION

PCA IMPLEMENTATION

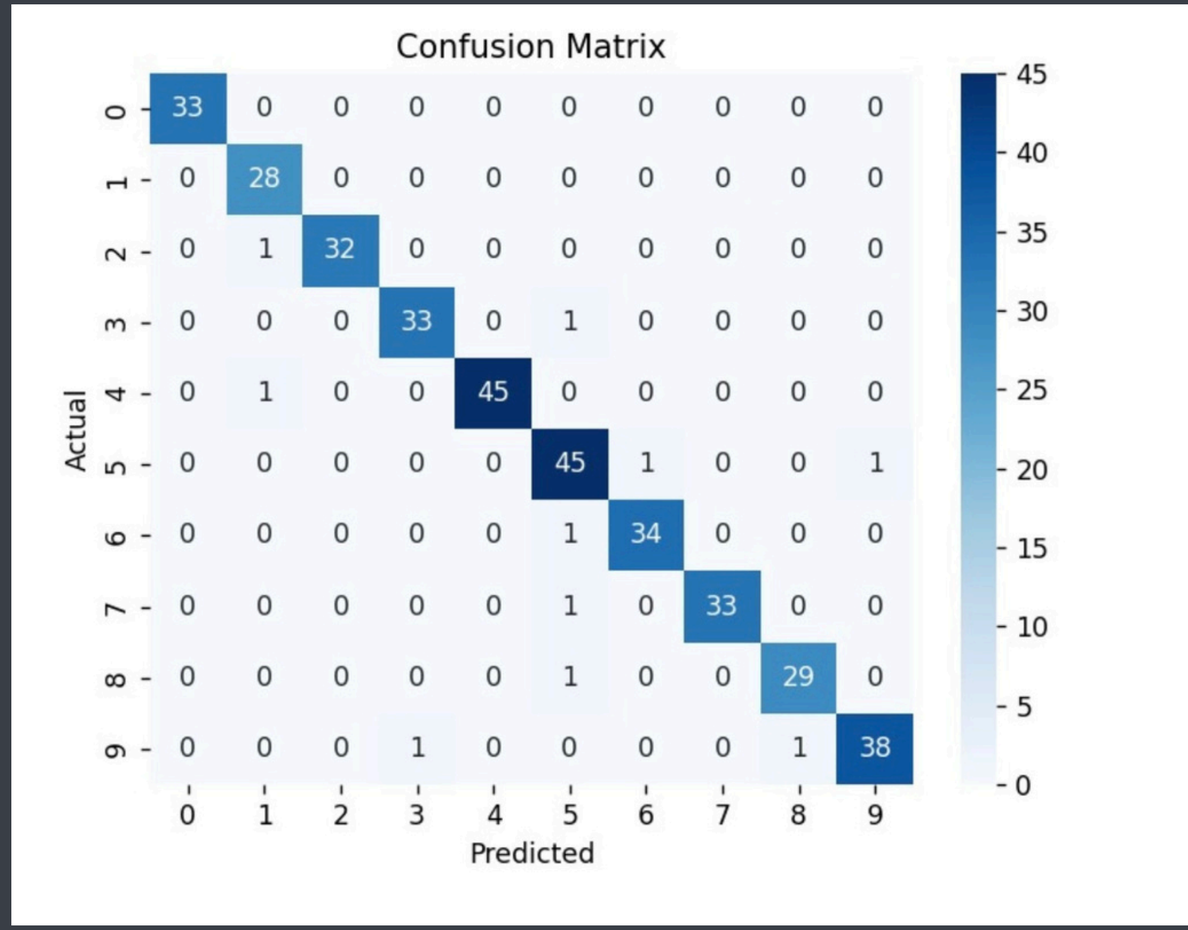
- Applied: PCA(n_components=30)
- fit_transform(X): compresses the data
- inverse_transform(X_pca): reconstructs it to visualize loss
- Cut data from 64 pixels to 30, reducing noise and speeding up training

TRAINING AND CLASSIFICATION

- Split data into 80% training / 20% testing
- Trained Logistic Regression on compressed features.
- Predicted labels using model.predict(X_test)

EVALUATION

- Classification Report gave us:
  - Accuracy, Precision
- Confusion Matrix (heatmap):
  - Showed which digits are confused (e.g., 4 vs. 9)



MATHEMATIC AREA

MEAN CENTERING

$$X_{centered} = X - \mu$$

EIGENVECTORS AND EIGENVALUES

$$\Sigma \vec{v} = \lambda \vec{v}$$

COVARIANCE MATRIX

$$\mathrm{Cov}(X) = \frac{1}{n-1} X_{\mathrm{centered}}^{\top} X_{\mathrm{centered}}$$

RESULT

- Compared original and reconstructed images
- Also predicted on a test image:
- “Predicted: 6 / Actual: 6” → shows PCA retained essential features

Original Image 64 vecs

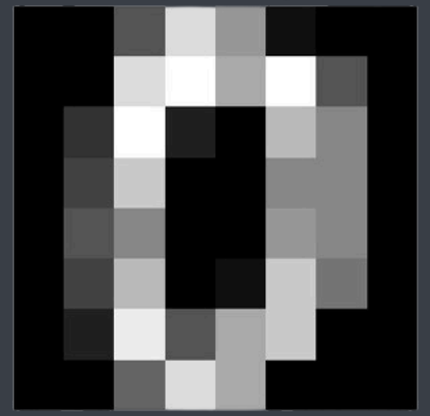


Image after PCA 30 Comp.

Predicted: 6, Actual: 6

CONCLUSION

By combining Principal Component Analysis and Logistic Regression, we were able to efficiently classify handwritten digit images.
PCA helped us reduce the number of features by keeping only the most important patterns in the data, making it easier and faster to process.
Then, logistic regression learned to recognize digit categories based on those compressed features, using a linear decision-making process.
This approach showed how powerful mathematical tools like vector projection, eigen decomposition, and probability-based classification can work together to solve real-world problems.

THANK YOU