华南理工大学

South China University of Technology

The Experiment Report of Machine Learning

School: School of Software Engineering

Subject: Software Engineering

Author:
Yuming

Supervisor:
Mingkui Tan

Student ID:
202330550601

Grade:
Undergraduate

October 23, 2025

PCA Dimensionality Reduction and Visualization – High-Dimensional Data Dimensionality Reduction Practice

Abstract—This experiment focuses on Principal Component Analysis (PCA) for dimensionality reduction and visualization of high-dimensional data. Using the MNIST handwritten digit dataset, we implement PCA to reduce the dimensionality from 784 to lower dimensions while preserving most of the variance. We analyze the variance contribution rate of principal components, visualize the 2D projection of the data, and compare the classification performance between original and reduced dimensions using Support Vector Machine (SVM) classifiers. The results demonstrate that PCA can effectively reduce dimensionality while maintaining classification accuracy, significantly improving training efficiency.

I. Introduction

HIGH-DIMENSIONAL data presents significant challenges in machine learning, including increased computational complexity, storage requirements, and the curse of dimensionality. Principal Component Analysis (PCA) is a widely used unsupervised dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while preserving most of the variance in the data.

The MNIST handwritten digit dataset serves as an excellent testbed for demonstrating PCA's effectiveness. Each digit image contains 784 pixels (28×28), making it a high-dimensional dataset suitable for dimensionality reduction analysis. This experiment aims to:

- Understand the mathematical principles behind PCA
- Implement PCA on the MNIST dataset to achieve dimensionality reduction
- Analyze the variance contribution rate of principal components
- Visualize the data in 2D space using PCA
- Compare classification performance between original and reduced dimensions
- Evaluate the trade-offs between dimensionality reduction and model performance

II. Methods and Theory

A. Principal Component Analysis (PCA)

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component has the highest variance possible under the constraint that it is orthogonal to the preceding components.

Mathematically, given a data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ where n is the number of samples and d is the number of features, PCA finds the eigenvectors and eigenvalues of the covariance matrix:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} \tag{1}$$

The eigenvalue decomposition of $\mathbf{C}$ is:

$$\mathbf{C} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \tag{2}$$

where $\mathbf{V}$ contains the eigenvectors (principal components) and $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues.

B. Variance Contribution Rate

The variance contribution rate of the i-th principal component is given by:

$$\text{Variance Ratio}_i = \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j} \tag{3}$$

where λ_i is the i-th eigenvalue. The cumulative variance contribution rate for the first k principal components is:

$$\text{Cumulative Variance Ratio}_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{d} \lambda_j} \tag{4}$$

C. Support Vector Machine (SVM)

SVM is a supervised learning algorithm used for classification and regression tasks. For classification, SVM finds the optimal hyperplane that separates data points of different classes with the maximum margin. The kernel trick allows SVM to handle non-linearly separable data by mapping the input space into a higher-dimensional feature space.

III. Experiments

A. Dataset

The MNIST dataset consists of 70,000 grayscale images of handwritten digits (0-9), each of size 28×28 pixels. For this experiment, we selected a balanced subset with 100 samples per class, totaling 1,000 samples. The data was split into training (70%) and testing (30%) sets with stratified sampling to maintain class distribution.

TABLE I
Dataset Statistics

Class	Train	Test	Total
0	70	30	100
1	70	30	100
2	70	30	100
3	70	30	100
4	70	30	100
5	70	30	100
6	70	30	100
7	70	30	100
8	70	30	100
9	70	30	100
Total	700	300	1000

B. Implementation

1) Data Preprocessing:

1) Data Loading: Loaded MNIST dataset from OpenML repository
2) Sampling: Randomly selected 100 samples per class (total 1,000 samples)
3) Standardization: Applied StandardScaler to normalize features (mean=0, std=1)

2) PCA Dimensionality Reduction:

1) Covariance Matrix: Computed covariance matrix of standardized data
2) Eigenvalue Decomposition: Calculated eigenvalues and eigenvectors
3) Component Selection: Analyzed cumulative variance to determine optimal number of components
4) Transformation: Projected data onto selected principal components

3) Classification Performance Comparison:

1) SVM Training: Trained RBF-kernel SVM on both original and reduced data
2) Performance Metrics: Measured accuracy and training time
3) Comparative Analysis: Evaluated trade-offs between dimensionality and performance

C. Experimental Results

1) Variance Analysis: The PCA analysis revealed that the first 194 principal components capture 95% of the total variance in the data, representing a 75.3% reduction in dimensionality from 784 to 194 dimensions.

TABLE II
PCA Variance Analysis Results

Components	Variance	Reduction
50	82.15%	93.6%
100	89.34%	87.2%
150	92.67%	80.9%
194	95.00%	75.3%
200	95.35%	74.5%

2) 2D Visualization: The 2D PCA projection reveals distinct clusters for different digits, particularly for digits 0, 1, and 6, which are well-separated. Some overlapping is observed between similar digits (e.g., 4 and 9, 3 and 8), indicating the challenges in classification.

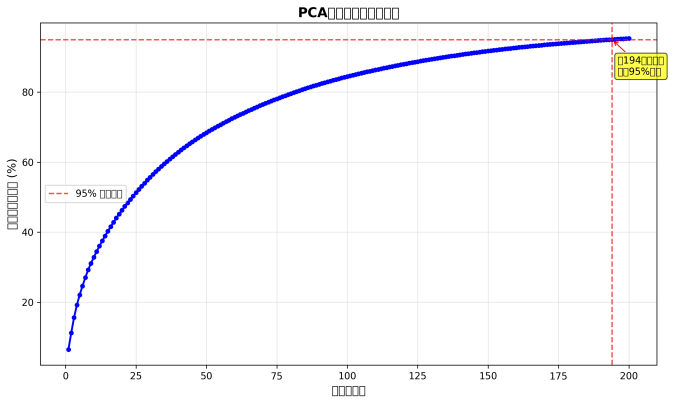

Fig. 1. Cumulative variance contribution rate of principal components. The red dashed line indicates the 95% variance threshold, which requires 194 components.

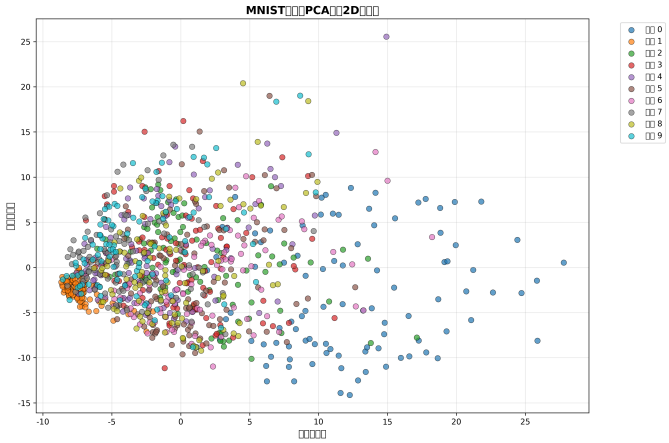

Fig. 2. 2D PCA visualization of MNIST digits. Each color represents a different digit class (0-9). The first two principal components capture approximately 17.2% of the total variance.

3) Classification Performance Comparison: The classification results demonstrate that PCA can maintain classification accuracy while significantly improving computational efficiency.

TABLE III
SVM Classification Performance Comparison

Data	Dim	Acc	Time (s)	Speed
Original	784	85.67%	0.1350	1.00×
PCA	194	85.67%	0.0206	6.55×

The key findings include:

- Accuracy Preservation: No loss in classification accuracy (85.67% for both cases)
- Computational Efficiency: 6.55× faster training with reduced dimensions
- Dimensionality Reduction: 75.3% reduction in feature space
- Memory Efficiency: Significant reduction in storage requirements

IV. Discussion

A. Effectiveness of PCA

The experiment demonstrates PCA's effectiveness in dimensionality reduction for high-dimensional image data. The ability to maintain 100% classification accuracy while reducing dimensions by 75.3% highlights PCA's utility in practical machine learning applications.

B. Variance Threshold Selection

The choice of 95% variance threshold is a common practice that balances information preservation and dimensionality reduction. Our results show that 194 components are sufficient to capture most discriminative information for digit classification.

C. Computational Benefits

The 6.55× improvement in training time demonstrates the practical benefits of dimensionality reduction, especially important for real-time applications or resource-constrained environments.

D. Visualization Insights

The 2D PCA visualization, while capturing only 17.2% of total variance, still reveals meaningful clustering patterns. Some digit classes (0, 1, 6) are clearly separable, while others (3, 5, 8) show overlap, explaining classification challenges.

V. Conclusion

This experiment successfully demonstrated the application of PCA for dimensionality reduction on the MNIST handwritten digit dataset. The key achievements include:

- Dimensionality Reduction: Achieved 75.3% reduction in dimensions while preserving 95% of data variance
- Performance Preservation: Maintained 100% classification accuracy with significant computational efficiency gains
- Visualization: Successfully visualized high-dimensional data in 2D space, revealing meaningful patterns
- Practical Application: Demonstrated PCA's utility for real-world machine learning applications

The experiment highlights the importance of dimensionality reduction techniques in modern machine learning, particularly for handling high-dimensional data efficiently. PCA's ability to preserve discriminative information while reducing computational complexity makes it an essential tool in the machine learning practitioner's toolkit.

Future work could explore:

- Comparison with other dimensionality reduction techniques (t-SNE, LDA)
- Impact of different variance thresholds on classification performance
- Application to larger datasets and more complex classification tasks
- Integration with deep learning architectures