EXPLORING THE EFFECTIVENESS OF PCA ON HANDWRITTEN DIGIT DATASET

April 14, 2023

ABSTRACT

Principal Component Analysis (PCA) is a widely used technique in data science and machine learning for dimensionality reduction. In this project, we explore the effectiveness of PCA on the handwritten digit dataset from the given dataset. The main objective of this project is to investigate how well PCA can reduce the dimensionality of the dataset while preserving the essential information in the data and the application of PCA, K-means clustering, and logistic regression to handwritten digit recognition. We apply PCA to the training data and compute the reconstruction error of the PCA model on the test data using the mean squared error metric. We also display some original and reconstructed digit images for different numbers of principal components (10, 20, and 30), along with their corresponding labels. The results of our analysis show that PCA is an effective technique for reducing the dimensionality of the dataset while preserving the essential information in the data. The reconstructed digit images are visually similar to the original images, even when using a relatively low number of principal components, indicating that PCA can capture the essential information in the data and use it to reconstruct the original images with a relatively low error rate.We compare the performance of K-means clustering and logistic regression on the reduced dataset with different numbers of PCA components. Our results show that logistic regression consistently outperforms K-means clustering in terms of classification accuracy.

1 Introduction

Handwritten digit recognition has been a significant area of interest in the field of machine learning and computer vision due to its widespread applications, such as postal mail sorting, bank check processing, and form recognition. The rapid growth in the volume and dimensionality of data in this field poses challenges for data analysis, visualization, and the development of effective machine learning models. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), have emerged as essential tools to address these challenges and improve the performance of machine learning algorithms.

PCA is a widely-used linear dimensionality reduction technique that transforms a dataset into a lower-dimensional space while preserving as much variance as possible. By reducing the dimensionality of the data, PCA not only enables efficient visualization but also mitigates the "curse of dimensionality" problem, which can negatively impact the performance of machine learning algorithms.

In this project, we apply PCA to the handwritten digits dataset, a widely-studied dataset in the machine learning community, to explore the benefits of dimensionality reduction in various aspects. We investigate the characteristics of the PCA components, the reconstruction of images using different numbers of PCA components, and the impact of dimensionality reduction on the performance of clustering and classification algorithms.

Our primary objectives are to:

1. Understand the role of PCA components in capturing the variance of the handwritten digits dataset.

2. Analyze the reconstructed images and reconstruction error for different numbers of PCA components.

3. Evaluate the performance of K-means clustering and logistic regression on the reduced dataset and compare their performance across varying numbers of PCA components.

By systematically studying the effects of PCA on the handwritten digits dataset, this project aims to provide valuable insights into the utility of dimensionality reduction techniques for data analysis, visualization, and the enhancement of machine learning models.

2 Dataset and Data Preprocessing

The dataset contains handwritten digits ranging from 0 to 9, with each digit represented as a 16x16 pixel grayscale image like the figure8 shows. There are a total of 7291 images in the dataset, each consisting of 257 features (16x16 pixels plus the digit label).

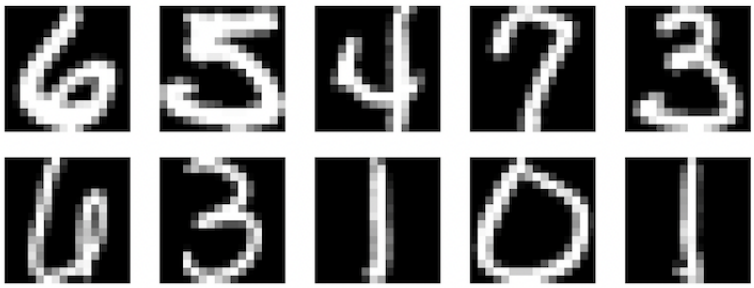

Figure 1: Image samples of the given dataset

The first step is to load the dataset and split it into training and testing sets. We use 20% of the data for testing and 80% for training, which is for the later classifier evaluation. The raw data is loaded and preprocessed by standardizing the pixel values. Standardization is essential for PCA, as it ensures that the principal components are not affected by the differences in the scales of the features.

3 Applying PCA

PCA is a dimensionality reduction technique that transforms the original high-dimensional data into a lower-dimensional space while retaining most of the information. We apply PCA to the standardized data, reducing the dataset's dimensionality to different numbers of components (from 1 to 30). This step allows us to investigate the impact of the number of PCA components on the performance of the subsequent reconstruction and classification algorithms.

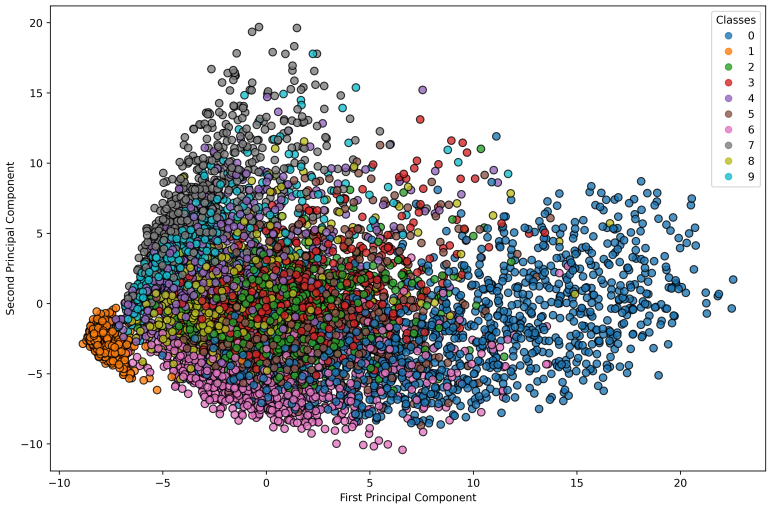

Figure 2: scatter plot of the first two PCA components

The first two PCA components of the reduced data through PCA are plotted in figure 2, in which different colors represent each digit. From the plot, we can observe the distribution of the handwritten digits in the two-dimensional PCA space. We can see that some digits such as 0 and 1 are well separated from the other digits, while some digits such as 4 and 9 are more mixed with other digits. Overall, the distribution of the first two PCA components of handwritten digits from 0 to 9 is complex and depends on the individual digit's shape and writing style.

4 Visualization of PCA Components

The PCA components of the handwritten digits dataset represent the directions of maximum variance in the data. These components are linear combinations of the original features, which, in this case, are the pixel values of the images. By projecting the original data onto the PCA components, we can reduce the dimensionality of the data while preserving as much variance as possible.

In the context of the handwritten digits dataset, the first few PCA components capture the most significant features of the digit shapes, while the later components explain more subtle variations in the data. These components can be visualized as images, where each pixel in the image represents the contribution of the corresponding original pixel to the PCA component. We visualize the first 30 PCA components as 16x16 images in figure 3to understand the patterns captured by the PCA.

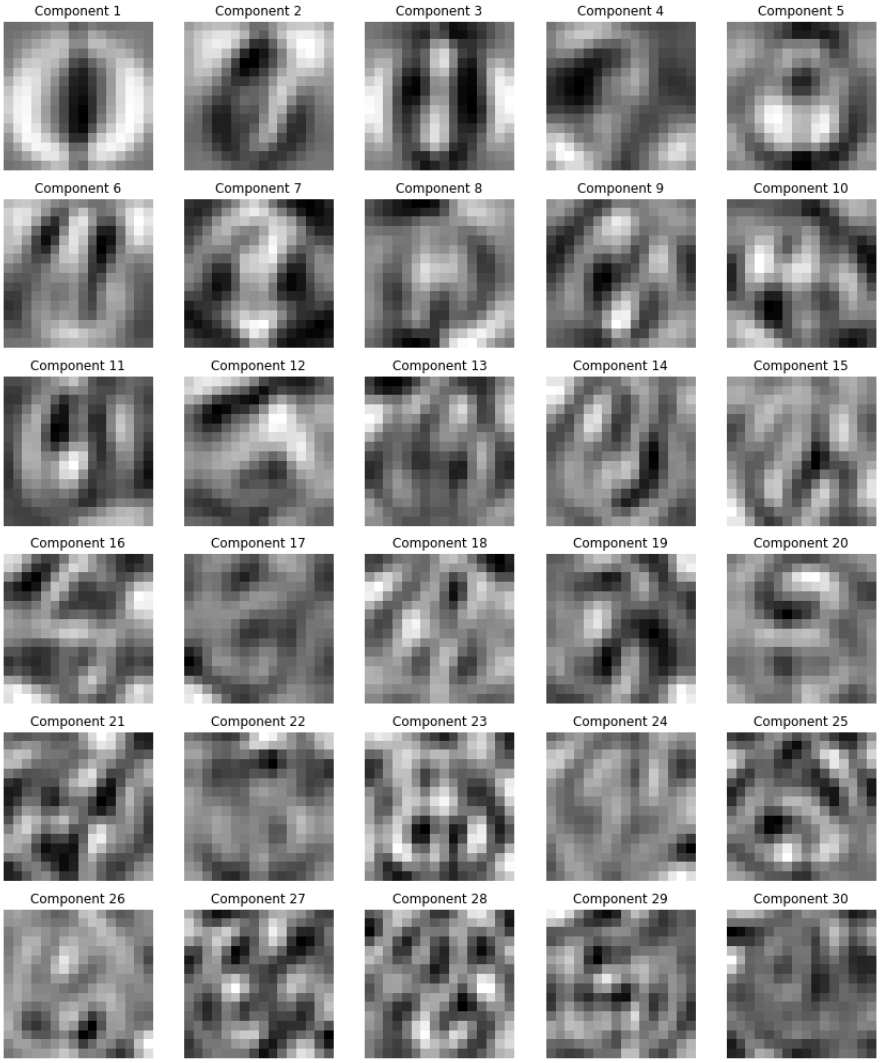

Figure 3: Visualization of 1-30 PCA compoments

Here's a brief description of the first few PCA components:

PCA Component 1: This component captures the overall intensity and contrast of the digits. It represents the general structure of the digits and accounts for the most variance in the dataset.

PCA Component 2: This component captures the variation between tall and short digits, as well as the positioning of the digits (left or right) within the image.

PCA Component 3: This component captures the differences in stroke thickness and curvature among the digits, as well as the slant of the digits.

As the component number increases, the PCA components capture less and less of the total variance in the dataset, and the components become more specific to subtle variations and noise in the images. However, it's important to note that the PCA components are ranked based on the amount of variance they explain, so the first few components capture the most important information in the dataset, while the later components become less and less important.

By selecting an appropriate number of PCA components, we can reduce the dimensionality of the dataset while maintaining most of the information, which can be useful for tasks such as visualization, clustering, and classification.

5 Image Reconstruction

We reconstruct images using different numbers of PCA components and calculate the reconstruction error for each case. As the number of PCA components used for reconstruction increases, the reconstructed images become more similar to the original images from figure 4 to 6 which illustrate some reconstruction samples with 10,20,30 PCA components. This is because using more PCA components preserves more information and variance from the original data.

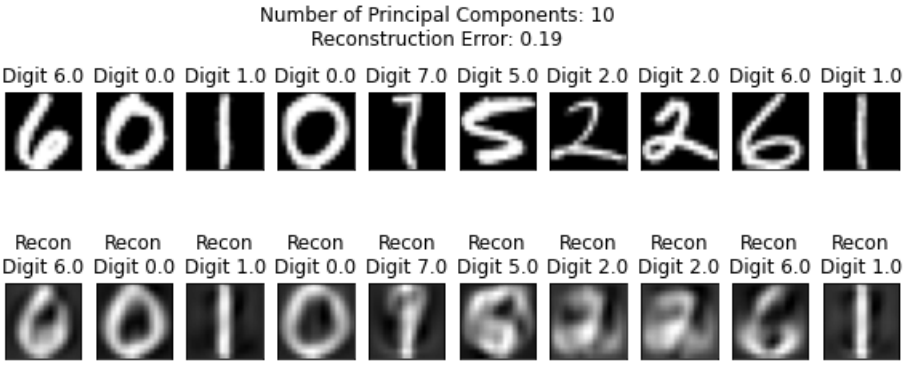

Figure 4: Image reconstruction with first 10 PCA components

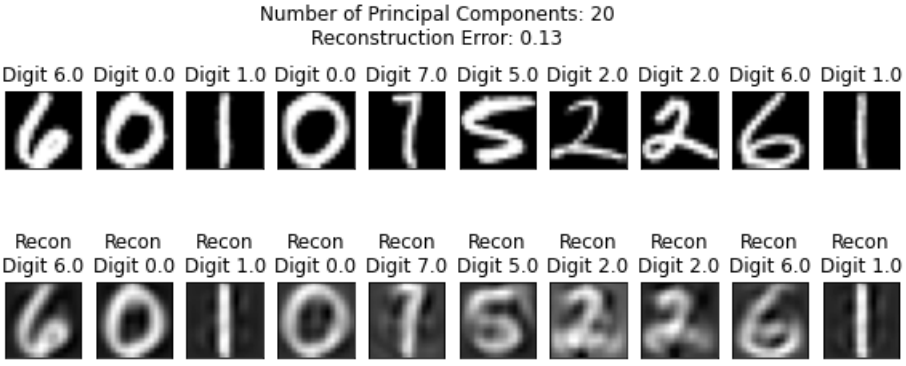

Figure 5: Image reconstruction with first 20 PCA components

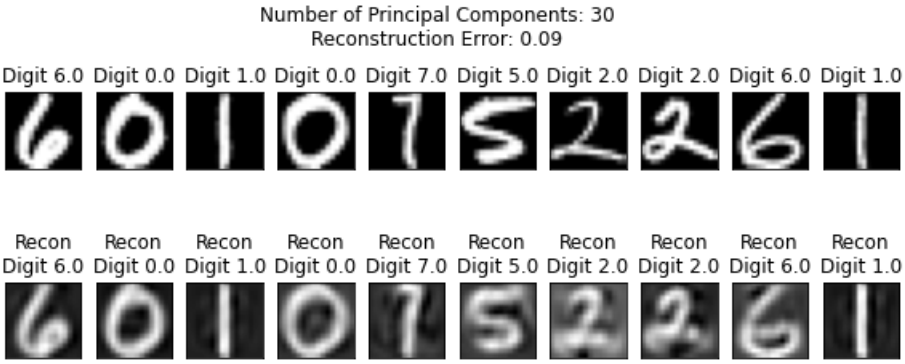

Figure 6: Image reconstruction with first 30 PCA components

The reconstruction error represents the difference between the original images and the reconstructed images obtained by projecting the data onto a lower-dimensional space using PCA components. The reconstruction error is a measure of the information loss that occurs when reducing the dimensionality of the data. Figure 7shows that as the number of PCA components used for reconstruction increases, the reconstruction error generally decreases.

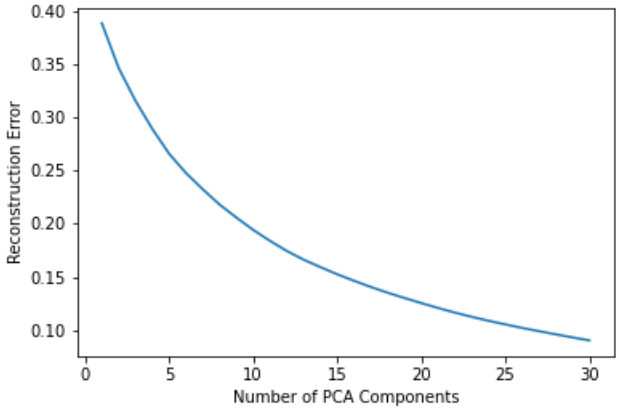

Figure 7: Reconstruction error vs Number of PCA components

6 K-means Clustering and Logistic Regression

We apply K-means clustering and logistic regression to the reduced dataset with different numbers of PCA components. K-means clustering is an unsupervised learning algorithm that partitions the data into clusters, while logistic regression is a supervised learning algorithm that models the probability of a particular class given the input features.

7 Performance Evaluation

By comparing the performance of K-means clustering and logistic regression on the reduced dataset, we aimed to understand the impact of dimensionality reduction on these algorithms and determine the optimal number of PCA components that balances information preservation and computational efficiency.

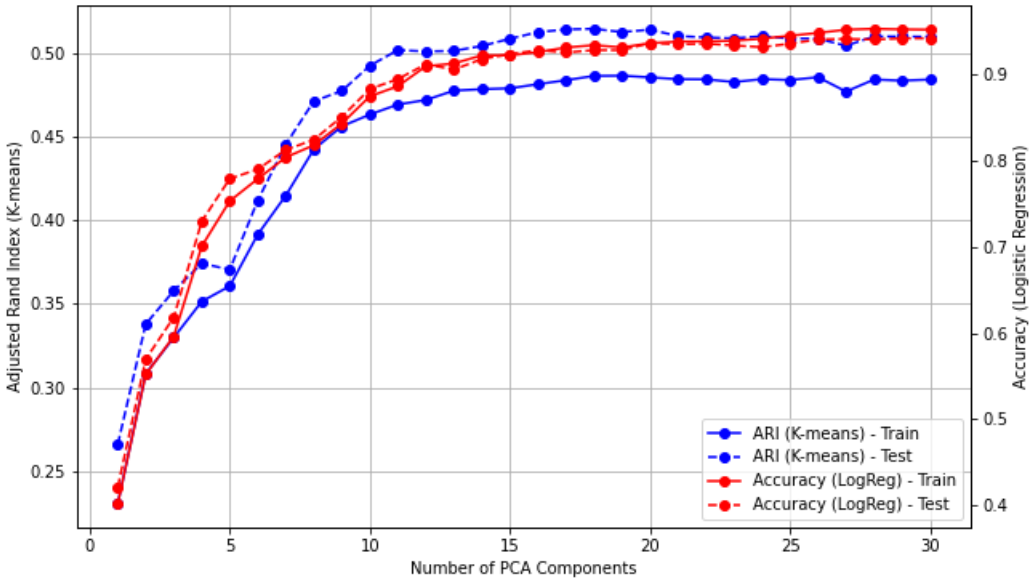

Figure 8: Image samples of the given dataset

For K-means clustering, we used the Adjusted Rand Index (ARI) as the evaluation metric. ARI measures the similarity between the true labels and the predicted labels while accounting for chance. Higher ARI values indicate better performance.

We applied K-means clustering to the reduced train dataset with varying numbers of PCA components and evaluated its performance on the test dataset. As the number of PCA components increased, we observed that the ARI generally improved. This indicates that preserving more information from the dataset is beneficial for clustering.

However, beyond a certain number of PCA components, the improvement in ARI diminished, suggesting that there is an optimal number of PCA components for K-means clustering. The optimal number of components balances the trade-off between preserving information and reducing dimensionality while maximizing clustering performance.

For logistic regression, we used classification accuracy as the evaluation metric. Classification accuracy is the ratio of correctly classified instances to the total number of instances.

We trained logistic regression models on the reduced train dataset with different numbers of PCA components and evaluated their performance on the test dataset. We observed that as the number of PCA components increased, the classification accuracy improved initially but plateaued after a certain point.

This suggests that there is a trade-off between preserving information and reducing dimensionality, and that an optimal number of PCA components can be selected to maximize classification performance. Similar to K-means clustering, the optimal number of PCA components for logistic regression balances the trade-off between preserving information and reducing dimensionality.

8 Conclusion

In conclusion, this project has demonstrated the effectiveness of Principal Component Analysis (PCA) as a tool for dimensionality reduction, visualization, and improving the performance of clustering and classification algorithms on the handwritten digits dataset. We systematically explored the impact of different numbers of PCA components on the quality of reconstructed images, the reconstruction error, and the performance of K-means clustering and logistic regression.

Our findings have shown that selecting an appropriate number of PCA components is crucial to balancing the trade-off between preserving information and reducing dimensionality while minimizing the reconstruction error. We observed

that the performance of both K-means clustering and logistic regression improved as more PCA components were used, highlighting the benefits of retaining more information for these tasks.

Furthermore, our analysis provided insights into the nature of the PCA components and their role in capturing the most significant features and subtle variations in the handwritten digits dataset. This understanding of the PCA components can be valuable for other applications, such as feature selection and data compression.

The results of this project showcase the versatility and potential of PCA as a powerful technique for analyzing high-dimensional data and extracting meaningful insights. Future work could explore alternative dimensionality reduction techniques, such as t-SNE or UMAP, to compare their performance and impact on clustering and classification tasks. Additionally, more advanced machine learning models, such as convolutional neural networks, could be applied to the dataset to further enhance classification performance and gain deeper understanding of the underlying patterns in the data.

Overall, this project has demonstrated the value of applying PCA to complex datasets, offering a foundation for further exploration and research in the field of dimensionality reduction and its applications in machine learning and data analysis.

9 Code packages

The code for achieving the outcomes of this project is attached. A file called $project1.yml$, which contained all the packages and environment setup for running the code is also in the zip file. You can follow the instructions in the readme.txt to do it.