Kirstin Megga Ramos

COMP257 – Homework1 – Written Report

PART 1:

A black screen with blue lines

Description automatically generated

A screenshot of a computer program

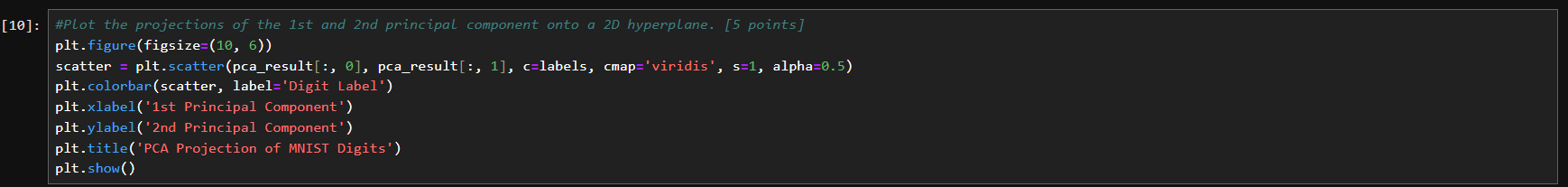
Description automatically generated

A screenshot of a computer

Description automatically generated

Analysis:

Applying PCA to reduce the dimensionality of the MNIST dataset extracts the first two principal components. The explained variance ratio tells us how much information is captured by each principal component. The first principal component is 9.75% and the second principal component is 7.16%. Reducing the dataset to its first component only means that there’s 10% of the variance and then the second component captures 16.91% of the variance. This suggests that the first two components alone only capture a small fraction of the total variance in the dataset, meaning that there are more components contributing to the variability of the digit images. But with the small amount of variance captured, PCA is useful for dimensionality reduction and helps reveal structure in high-dimensional data by projecting it into lower-dimensional space.



A diagram of a colorful explosion

Description automatically generated with medium confidence

Analysis:

This project suggests that the distinction between digits has been successfully captured. However, some digits are more complex than others such as 3, 5 and 8 where the results overlap. I believe this is a good 2D visualization but to fully understand all digits, more dimensions are needed as PCA reduces the complexity of data.

A black rectangular object with a black border

Description automatically generated

Analysis:

Using Incremental PCA, the dimensionality of the MNIST dataset has been reduced from 784 to 154 dimensions. Partial\_fit is applied to each batch of dataset, allowing IncrementalPCA to learn the principal components of the dataset. And after fitting the model using partial\_fit, the entire dataset (X\_train\_flattened) is transformed into a reduced form, resulting in X\_train\_reduced. So now, we still have 70,000 images but each image is represented by only 154 dimensions which contains most of the meaningful variance in the data.

Discuss the challenges encountered and solutions to overcoming them.

* The MNIST dataset is large, containing 70,000 instances. And the solution to overcome this is to use Incremental PCA which allows the data to be processed in smaller batches without running out of memory. It’s also good that the principal component used for Incremental PCA is already given, otherwise, I would encounter difficulty in choosing it as it may lost important variance. Another challenge I encountered is the reconstruction of the data so that we can visualize again the original dataset to the reduced one. I definitely encountered mismatched array shapes while trying to reshape and display the reconstructed images. It comes to my attention that PCA requires careful attention when transforming and reconstructing the data, particularly when working with reshaped data from 28x28 images to flat vectors and then back again.

PART 2:

A screenshot of a computer

Description automatically generated

A graph of colored dots

Description automatically generated

A computer screen with text on it

Description automatically generated

A black screen with colorful text

Description automatically generated

A diagram of a triangle with dots

Description automatically generated with medium confidence

Analysis:

The visualization above accurately represent the data transformations through different kernels. Generating a Swiss roll dataset, I already know this is going to be a non-linear dataset and using Linear Kernel which is basically applying the traditional PCA, it would not capture the non-linear relationships. On the other side, RBF Kernel and Sigmoid, can both capture the non-linearity of the dataset. RBF, I believe is the most effective way of unfolding the Swiss roll’s complex 3D shape to 2D representation. Sigmoid is also good but it doesn’t perform as well as RBF in this case.

A screen shot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Analysis:

For more proof that RBF is the best or optimal hyperparameter for kPCA, binary classification is used for this problem in order to apply Logistic Regression for classification. And then, we implemented GridSearchCV to confirm for the optimal hyperparameter and as the result showed, rbf outperforms all with the gamma value of 1. The high cross-validation accuracy of 0.91 and test accuracy of 0.945 indicates that the model generalizes well to unseen data. Additionally, the use of kPCA allows us to apply Logistic Regression in a reduced space that captures non-linear relationships in data. A screen shot of a computer

Description automatically generatedA graph with a line

Description automatically generated