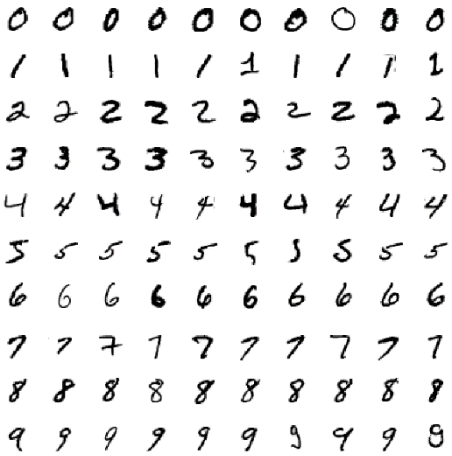
Muhammed Göktepe
2017400162

Assignment 3 Report

Before running my code, you need to add train and test set to code directory. Otherwise, will not run.

Step 1

In step one I load the train dataset and traverse it until I print 10 images of each digit. Resulting plot is as follows.

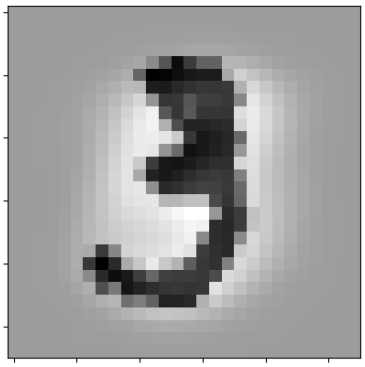


Sample digit images from the MNIST datab 1

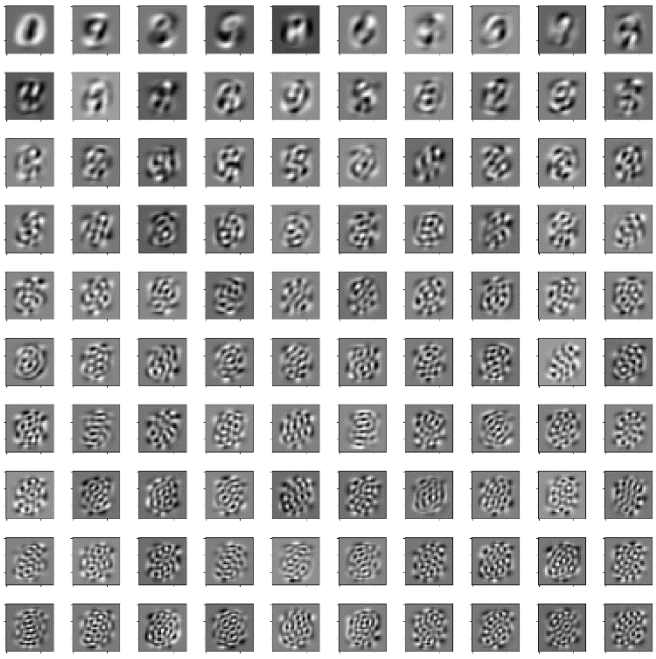
Step 2

In second step I calculated the mean of the columns of the whole dataset and subtract it from each row of the dataset. Thus, I can get the centered dataset. After this step I plotted one digit from dataset. Then I calculated the covariance of the centered dataset. I calculated eigenvalues and eigenvectors for 100 dimensions. I sorted eigenvalues and eigenvectors. Then plotted eigenvectors and eigenvalues.

Muhammed Göktepe
2017400162

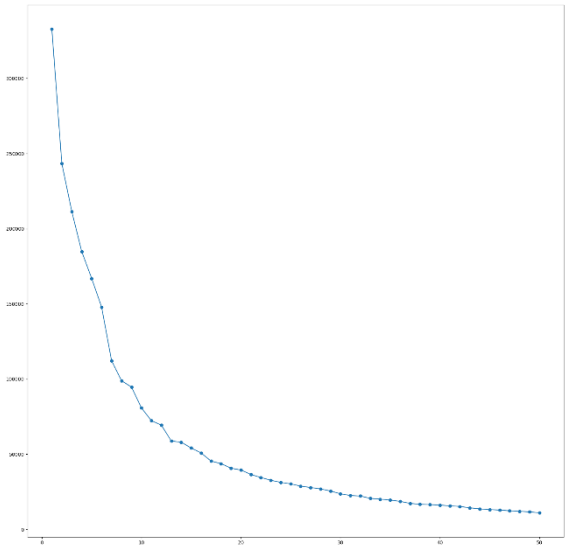


Mean digit image 1



Largest 100 eigenvectors. 1

Muhammed Göktepe
2017400162

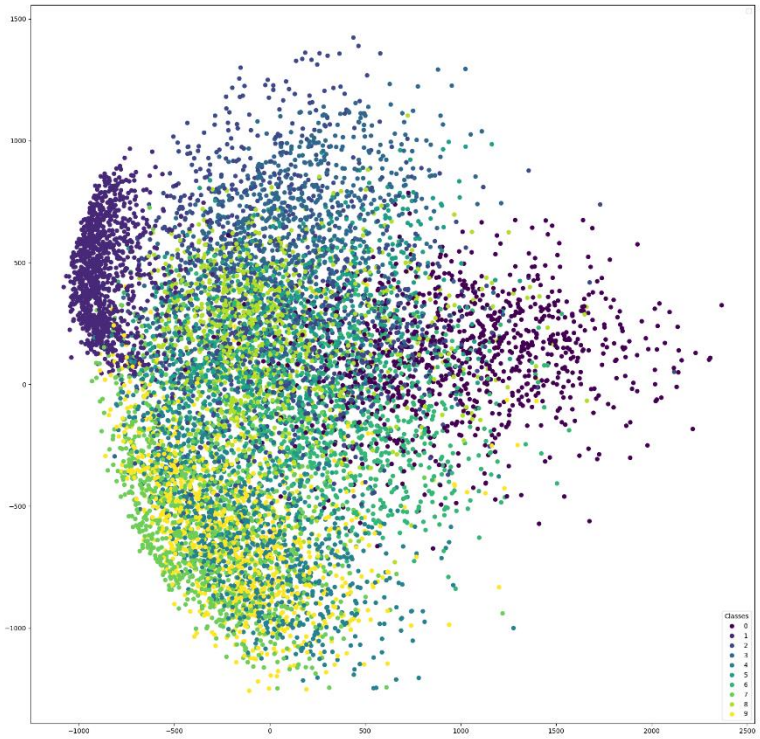


Scree plot (Largest 50 eigenvalues). 1

Step 3

In this step I followed the same procedure in step to and calculated eigenvectors and values. I used top two eigenvectors since the question asked us to reduce the dimensionality to two. Then I plotted the resulting 2D cartesian result. I added the legend to ease which point represent which number. PCA is a technique for reducing the number of dimensions in a dataset whilst retaining most information. It is using the correlation between some dimensions and tries to provide a minimum number of variables that keeps the maximum amount of variation or information about how the original data is distributed. In the plot I see some clusters especially in 1. Other digits are a bit separated but digit 1's cluster is very close to each other.

Muhammed Göktepe
2017400162

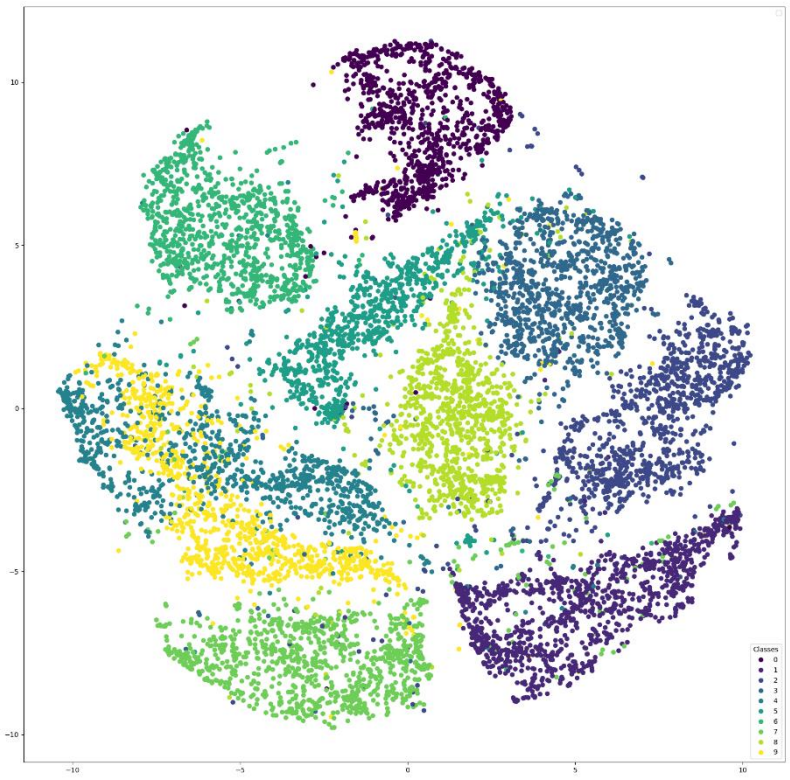


2D PCA visualization of a subset of the 1

Step 4

In this step I used scikit learns library to compute tsne visualization. You can see the plot in the following.

Muhammed Göktepe
2017400162



2D t-SNE visualization of a subset of th 1

Step 5

t-Distributed Stochastic Neighbor Embedding (t-SNE) is another technique for dimensionality reduction and is particularly well suited for the visualization of high-dimensional datasets. Contrary to PCA it is not a mathematical technique but a probabilistic one. The original paper describes the working of t-SNE as:

"t-Distributed stochastic neighbor embedding (t-SNE) minimizes the divergence between two distributions: a distribution that measures pairwise similarities of the input objects and a distribution that measures pairwise similarities of the corresponding low-dimensional points in the embedding".

In t-SNE it looks original data and looks at how to represent this data using less dimension with matching these two distributions. But it uses heavy mathematical computations so in high dimensional dataset takes longer time. Complexity of this algorithm is quadratic. To

reduce the computation time people reduces dimension with other algorithms and uses t-SNE with reduced dataset.

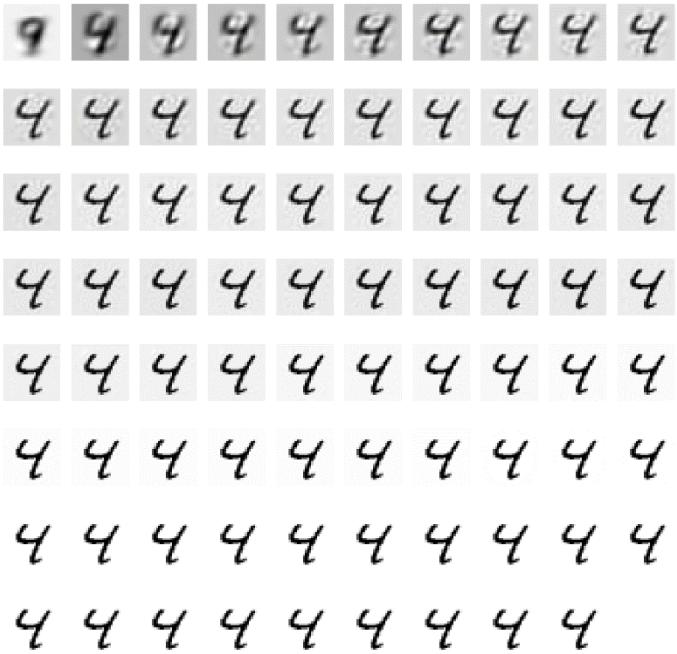
Unlike PCA it tries to preserve the Local structure of data by minimizing the Kullback–Leibler divergence (KL divergence) between the two distributions with respect to the locations of the points in the map. This technique finds application in computer security research, music analysis, cancer research, bioinformatics, and biomedical signal processing.

PCA is a linear dimensionality reduction technique, but t-SNE is non-liner. PCA tries to preserve global structure of data, but t-SNE preserves local structure of data. PCA highly affected by outliers, but t-SNE can handle outliers. PCA is a deterministic algorithm, but t-SNE is non-deterministic. PCA rotates vectors to preserve variance. T-SNE minimizes the gaussian distance between the points.

Step 6

In this step I reconstructed two images. One from the dataset and the other one is Pokémon. While reconstructing I started the dimension from two and increased by ten in each iteration. So the dimensions in my implementation will be 2, 12, 22, …., 782. I have two different plots in this step. First plot includes all iterations until 782 and plots each iteration. Second plot stops when explained variance reached a threshold. In Pokémon example explained ratio reached the threshold in 182nd iteration and it only plots 18 images. After this iteration there will be no significant change in reconstructed image.

Muhammed Göktepe
2017400162

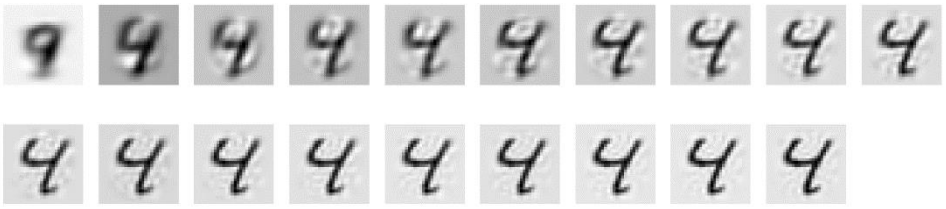


PCA reconstruction examples. 1 Sample from dataset

PCA reconstruction examples. 2 Pokemon sample with all ierations

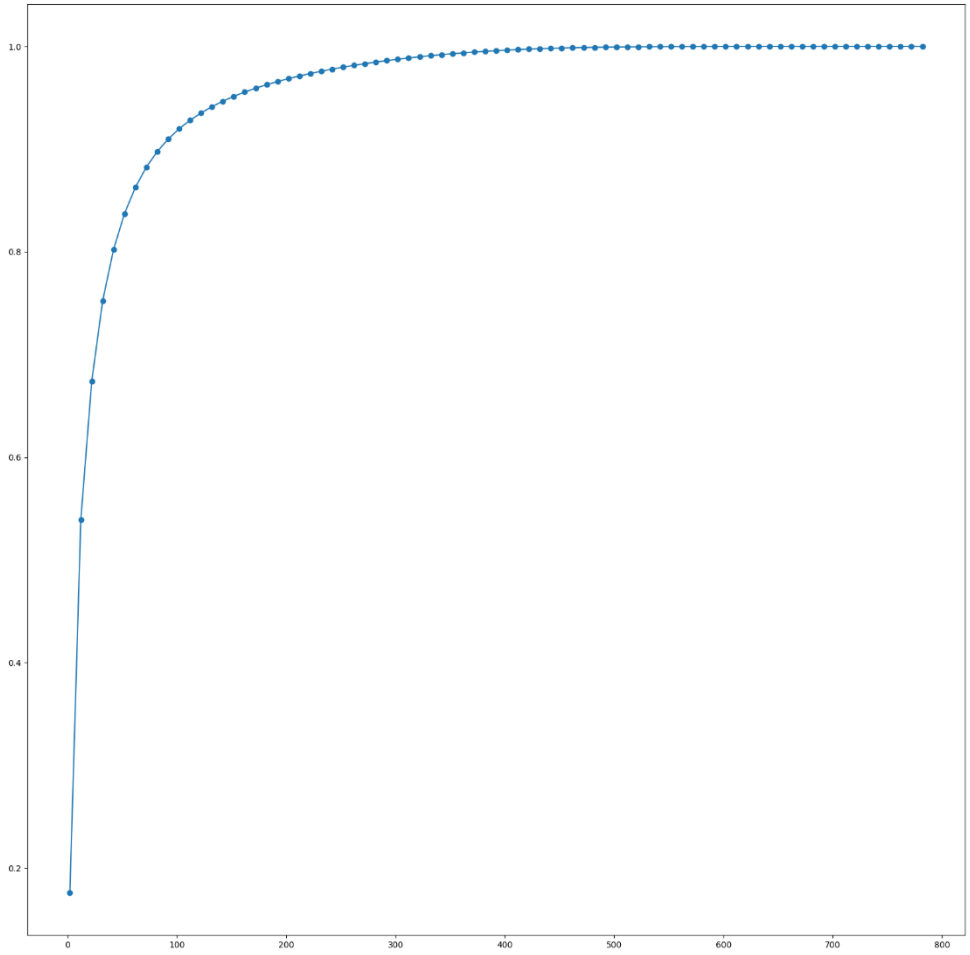
Muhammed Göktepe
2017400162

PCA reconstruction examples 1 pokemon. Using explained variance raito. 182nd iteation



PCA reconstruction examples 2 Using explained variance raito. 182nd iteation

Muhammed Göktepe
2017400162



Explained Variance Plot 1

Muhammed Göktepe
2017400162

Reference

https://towardsdatascience.com/visualising-high-dimensional-datasets-using-pca-and-t-sne-in-python-8ef87e7915b

https://www.geeksforgeeks.org/difference-between-pca-vs-t-sne/