In the name of Allah

دانشگاه یزد

Computer Engineering Faculty of Yazd University
Artificial Intelligence

Pattern Recognition Course

Homework 4
(Feature Extraction)

Instructor: Dr. Yazdian

Shiva Zeymaran

Student ID: 40109434

Spring 2023

PCA

a) First, we should load the dataset that is 10,000 samples with 784 features. In this question we should reshape the dataset to have the (784x10,000) data.
To implement PCA at first, we define a function named 'initial_calc' to calculate mean vector, covariance matrix, eigenvalues, and eigenvector. We extract this calculation from main body of PCA to have less calculation in future parts.

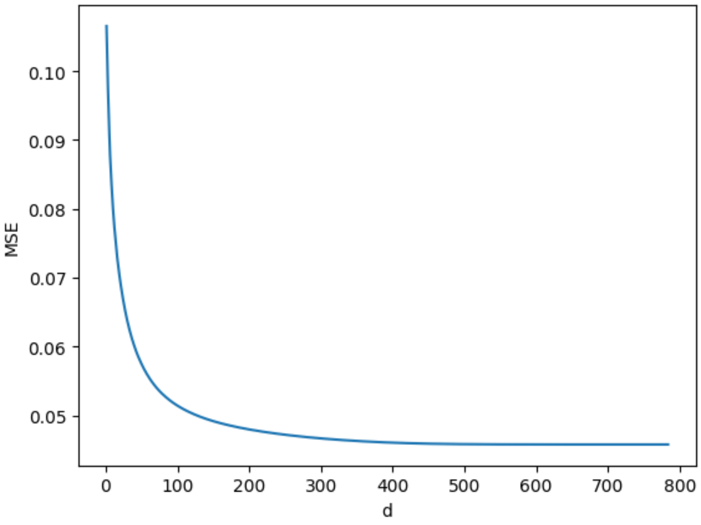
b) The suitable value of 'd' is **150**. This means that if we use PCA with d = 150, we will have a good PCA-transformed data with low variance from real dataset.

c) The formula to calculate 'X_hat' from eigenvectors and PCA result is as follow:

$$\boxed{\text{PCA reconstruction} = \text{PC scores} \cdot \text{Eigenvectors}^{\top} + \text{Mean}}$$

Using this formula, we implemented the 'pca_reconstruct' function.

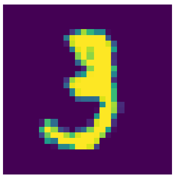
Plot various values of 'd' vs. MSE of original samples and their reconstructions:



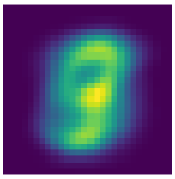
As a result, by increasing the value of 'd', the error of reconstruction will decrease. Also, increasing the value of 'd' up to approximately 200 has more effect on decreasing the MSE and after that there is not a big change up to end. Therefore, we can use 'd' equal 200 to get a good PCA-transformed data.

d) Here, we try to reconstruct the dataset again, with 5 given 'd' values, and then extract the reconstruction result of 10th sample. Here is the original image and results of reconstruction:

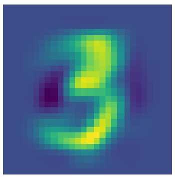
Original



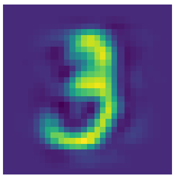
d = 1



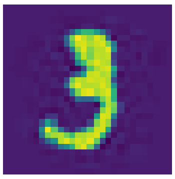
d = 10



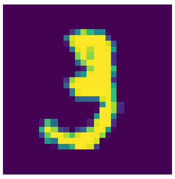
d = 50



d = 250

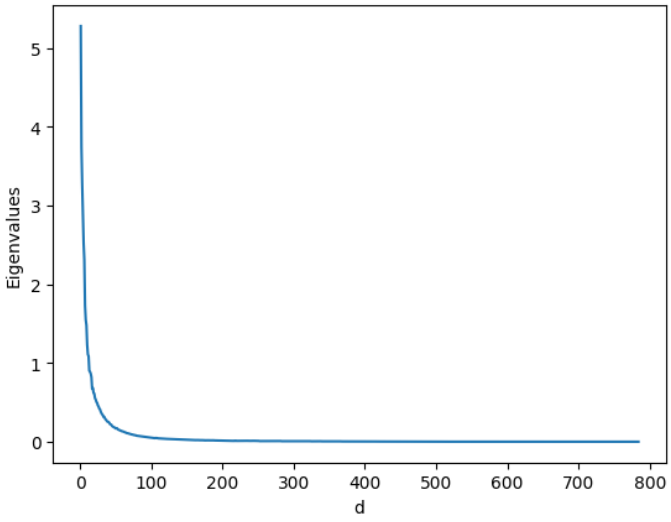


d = 784



As a result, by increasing the value of 'd' we will have more accurate result. It means that we will have more features and this will help us to reconstruct data near to original data. Also, it is clear that with d = 250 we have a good estimate of sample and we again understand that this is an optimal value for 'd'.

e) Plot 'd' vs. eigenvalues:



As a result, to select the best value for 'd' (number of eigenvectors), we need to select eigenvector with large eigenvalue and discard the ones with small eigenvalues. Here, the best value for 'd' is **150** (as it was in part b).