Intro to ML and Big Data
Assignment 4

CS 4731 — Dr. Mihail
Department of Computer Science
Valdosta State University

October 15, 2019

Do not attempt a Mihail homework the night it's due.

1 Introduction

In this assignment, you will implement extend the previous assignment by computing a PCA subspace and reducing the dimensionality of the original data points to improve the accuracy of a nearest neighbor classifier.

2 Background

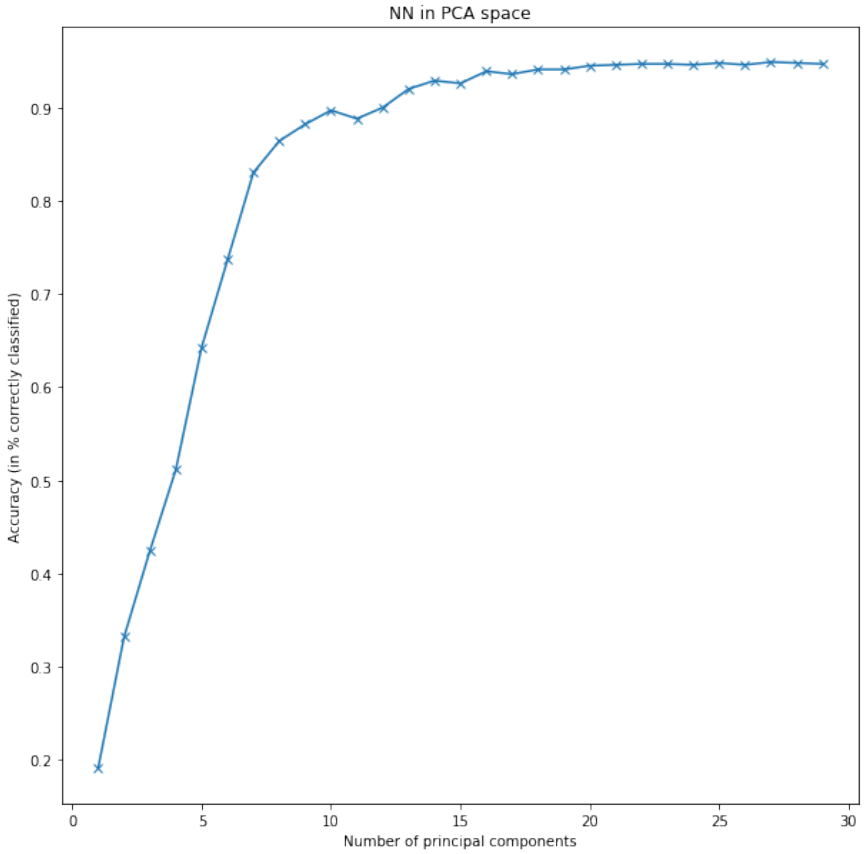
The data points consist of 28x28 grayscale images, linearized as 1x784 row vectors. If you consider each digit as a point in a 784 dimensional Euclidean space, then a memory-based technique such as Nearest Neighbor consists simply of storing the original data points in this space. At test time, a new digit is classified by computing some distance (e.g., Euclidean) between the new datapoint and EVERY point in this space. The closest digit (in terms of *some* distance metric) is then looked up and the class of that digit is then used to classify the new digit.

3 Dataset

You will use a preprocessed dataset from: `https://github.com/daniel-e/mnist_octave/`. The code to load the dataset into Python is as follows:

```
!wget https://github.com/daniel-e/mnist_octave/raw/master/mnist.mat
import scipy.io
mat = scipy.io.loadmat('mnist.mat')
```

Figure 1: Final result. On the horizontal axis you have the number of principal components used in NN classification. On the vertical axis, you have the accuracy (in %).



4 Project requirements

The main experiment in this project consists of fixing the number of training samples (10000) and the number of test samples (1000, from a disjoint set) and computing accuracy of NN classification for multiple choices of principal components. When you successfully implemented this project, you should see a figure, such as Figure 1.

- **Write-up** You will write-up your project in a LaTeX document with all the decision points along the way documented and discussed. This has to have a section titled "What I've learned" where you will discuss what you learned.

5 Due date

The project is due before the end of the day Sunday, October 20th.