Data Mining
IT 270
Homework 3

Dr. Alexander Pelaez
alexander.pelaez@hofstra.edu

$5^{th} April, 2023$

Instructions

Please make sure if you use R you copy and paste it into Word using Courier Font (makes it easier to Read). For each of the problems that are looking for a response (not just a calculation), be sure to explain and interpret the results. If you are not sure. . . ASK PLEASE. Please start each question on a new page and clearly label that start of each problem (Maybe slightly larger font, bold face , underline. . .) anything that will help me find the problem you are working on.

Please remember if you submit a word document, you must copy your R Code and results (generally just copy what the output in the console is and it will include your code). Don't provide me R code with original data output please - it makes it too long. When you copy and paste it to word please make sure the R Code / Output is in courier font (10 pt).

1 Handwriting Analysis

Files Needed:

- Code: `handwriting.R` .
- Data: `mnist_train.csv`

Notes: the R code refers to a file online, but you can download it from Blackboard instead.

Your colleague has come up with a great piece of code to do handwriting recognition.They mentioned that it works perfectly with 100% accuracy and wants to get your opinion. The actual data set is much larger, and your colleague hasn't commented the code.

Your objective:

(a) Comment each section labelled "SECTION " , be sure to provide enough information of what is going on in the section.

(b) The code needs to work on your machine, so you will need to get it to work.There may be a few errors. Indicate what you needed to do to get the code to work.

(Please note that running the code could be limited based on your machine. What you might want to do is reduce the data set to say 1000 observations, get the code to work, then slowly increase the number of observations until you computer cant seem to run it efficiently anymore.)

(c) Analyze what your colleague did and explain their process (from the **technique** perspective) and whether the code could be improved. Indicate the number of hidden layers and number of neurons. Do you think this might be a problem, if so why?

(d) Use the table below to provide level of accuracy of your colleagues, code. If it doesn't not run , make any necessary modifications you feel necessary to get the code to run, while keeping in line with your colleagues "approach".

Observations	Execution Time	# Accurate	% Accurate
1000			
2000			
3000			
...			

(e) Next, modify the code in a way that you think is more efficient and for the same observations, or even more provide a similar table. Indicate the number of hidden layers, you choose and the number of hidden nodes in each layer.

Observations	Execution Time	# Accurate	% Accurate
1000			
2000			
3000			
...			

(f) Provide an interpretative summary of the differences between your final code and your colleagues code. Provide any explanations about the approach and your thoughts about the different methods.