CSE 151: Introduction to Machine Learning Summer 2019, Session 2

Programming Assignment 1

Instructor: Joseph Geumlek **Due on:** Aug 15

Instructions

- This is a 20 point homework. The assignment should be done individually.
- You are free to use any programming language that you wish.
- The programming assignment should be submitted as a single pdf file to Gradescope. Enter your answers to the questions first, followed by a copy of your code.

Problem 1 (20 points)

In this problem, we look at the task of classifying images of digits using k-nearest neighbor classification. Download the files `pa1train.txt`, `pa1validate.txt` and `pa1test.txt` from the class website. These files contain your training, validation and test data sets respectively.

For your benefit, we have already converted the images into vectors of pixel colors. The data files are in ASCII text format, and each line of the files contains a feature vector of size 784, followed by its label. The coordinates of the feature vector are separated by spaces. In other words, each line has 785 numbers, of which the first 784 are coordinates of a feature vector, and the 785th is the label.

1. For $k = 1$, 5, 9 and 15, build k-nearest neighbor classifiers from the training data. For each of these values of k, write down a table of training errors (error rate on the training data) and the validation errors (error rate on the validation data). Which of these classifiers performs the best on validation data? What is the test error rate of this classifier?

   [Hint: As a check for your code, the training error for $k = 3$ should be about 0.04, rounded.]

   Due to random tie-breaking, your results may vary a little when you rerun your code. For this assignment, you only need to run each classifier once, and report the error rates seen. Do not worry about the fluctuations.

   You should randomly break ties when it comes to taking the majority vote. You do not need to worry about breaking ties on distances for this assignment.

2. In the first lecture, we talked about processing data with projections. In this part of the assignment, we will look at how using a projection as a pre-processing step affects the accuracy and running-time of nearest neighbor classification.

   Download the file `projection.txt` from the class website. This file represents a projection matrix P with 784 rows and 20 columns. Each column is a 784-dimensional unit vector, and the columns are orthogonal to each other.

   Project the training, validation and test data onto the column space of this matrix, and repeat part (1) of the problem. For $k = 1, 5, 9, 15$ write down a table of the training and validation errors, as well as the test error of the classifier which performs best on the validation data.

   [Hint: As a check for your code, the training error for $k = 3$ after projection should be about 0.16.]

   How is the classification accuracy affected by projection? How does the running time of your program change when you run it on projected data? It is okay to only talk about rough estimates of your running times.