CSE 151: Introduction to Machine Learning

Programming Assignment1

Instructor: Kamalika Chaudhuri **Due on:**

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.

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 on the training data) and the validation errors (error on the validation data). Which of these classifiers performs the best on validation data? What is the test error of this classifier?

   [Hint: As a check for your code, the training error for $k = 3$ should be about .04.]

2. In the first few lectures, we talked about preprocessing 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?

Solutions

1. The error values were:

k	Training Error	Validation Error
1	0	0.082
5	0.0555	0.100
9	0.0695	0.106
15	0.0915	0.104

   Based on the validation error, the best classifier is the 1-NN classifier. Its error computed on the test data is 0.094.

   (Your computed errors may be different).

2. The error values for the projected data are:

k	Training Error	Validation Error
1	0	0.320
5	0.1945	0.296
9	0.2305	0.290
15	0.2550	0.292

Based on the validation error, the best classifier is the 9-NN classifier. Its error computed on the projected test data is 0.291.

(Your results may differ.)

The classification accuracy is significantly reduced when using only the projected data. With the full 784 dimensional data, the best validation error rate was 0.082. However, using the 20 dimensional projections of the data, the best validation error was 0.29.

The running time was significantly improved when using the projection. Using the full data took about 8 seconds to perform 3-NN classification of the validation set, while the projected data only took 2 seconds to perform the 3-NN classification of the validation set.