Homework 4

1) Implement the Naive Bayes Classifier using your choice of programming language. Please comment your code.

2) Use the training and testing data from the file digits.zip. Calculate the percentage of correctly classified digits. Provide the confusion matrix.

3) Prepare a report on the results of your project. Include your results, your code and instructions how to run it.

Comments on the data.

In the file data.zip:

math.gmu.edu/~igriva/data.zip

you will find 20 files:
10 for training: train0.txt, train1.txt,…,train9.txt
10 for testing: test0.txt, test1.txt,…,test9.txt

Each training file has more than 5000 digits. Select an appropriate number of training examples. Each testing file has about 1000 digits. Try to use all testing digits.

Each line from the files corresponds to each hand written digit. The first entry on each line is the digit it represents: 0, 1, 2 etc... 9 followed by 28*28 = 784 dimensional grayscale intensity vectors with the entries from 0 to 255. Scale each intensity vector with dividing it by the largest entry. All the entries will be between 0 and 1. Replace all nonzero entries with 1 to have 0 and 1 entries only.