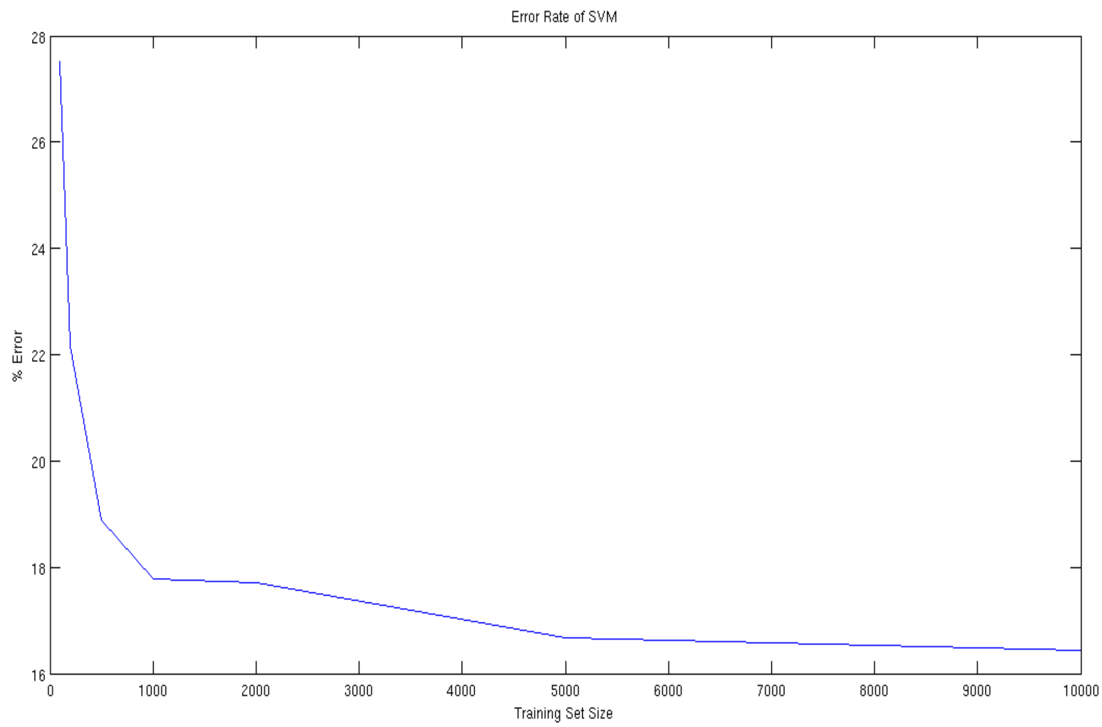
CS 189 Problem Set 1
2/1/14

Thomas Chow
Login: by

Mukul Murthy
Login: dq

Problem 1

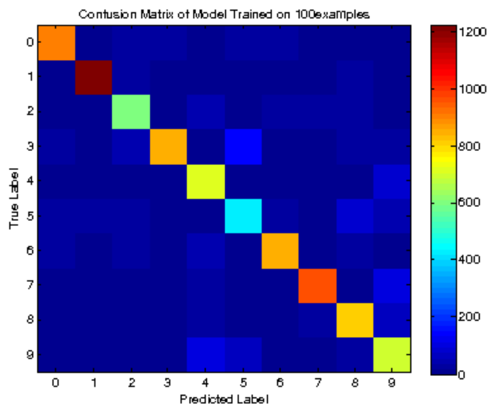
:

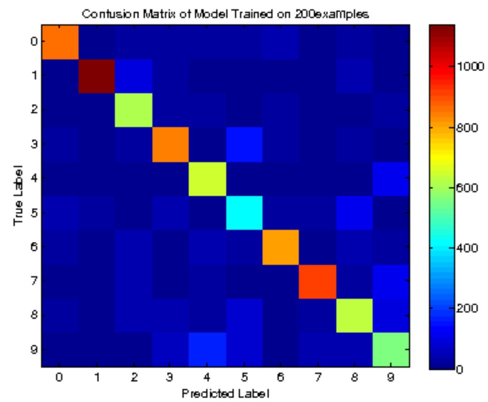


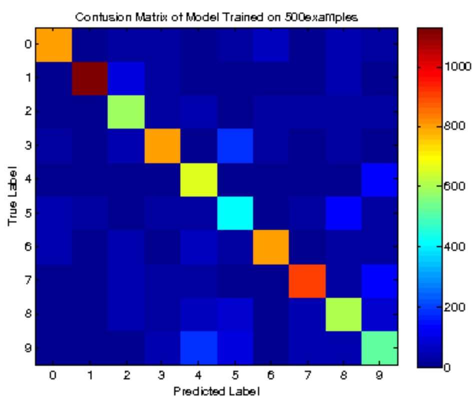
Parameters: We used the default C=1 and ε=0.1. We thought those gave us reasonable estimates until cross-validation to lower the error rate in part 3.

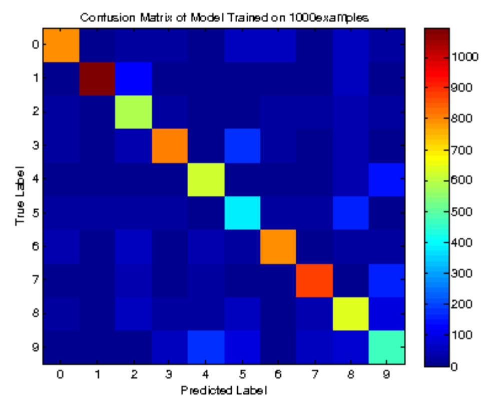
Problem 2

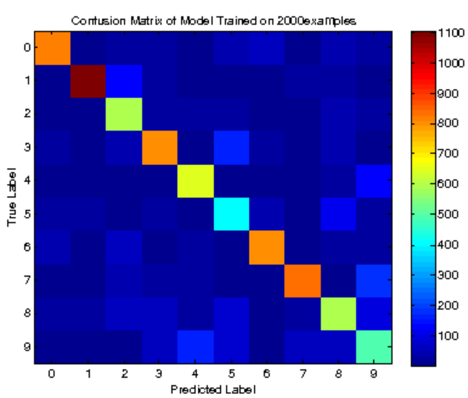
Confusion Matrices:

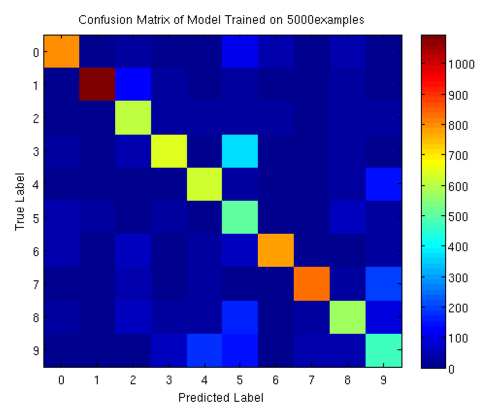


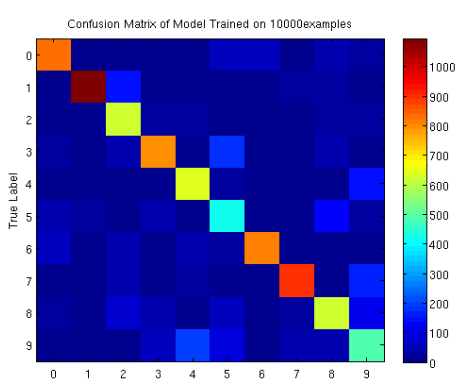












From looking at the confusion matrices, we can determine that our classifier was pretty good but not perfect. In an ideal world where every number was classified correctly, the confusion matrix would be one color down the diagonal (assuming equal quantities of each number) and zero everywhere else. Our confusion matrices showed that there were many cells not on the diagonal with significant quantities, meaning that these were misinterpreted numbers. The confusion matrices also show that the larger the size of the training data, the better the classifier, as non-diagonal cells were generally a darker shade of blue with the higher training data sizes.

In terms of digit classification specifically, it seems that often times digits 9 and 4 are mixed up, as well as 5 and 3 or 5 and 8. This makes intuitive sense as such pairs of digits can have similar handwritten stroke patterns.

Problem 3

Our optimal parameter we found for C was 10^-6. This resulted in an error rate of 9.78%. We experimented with magnitudes starting from 10^0 down to 10^-10, ignoring constant factors.

Cross validation is useful because it helps smooth out any abnormal data. Because part of the training set is held out while training the classifier, that data is tested as a check of how good the SVM is. And because cross validation alternates which subset of the training data is held out and averages it, this helps prevent model overfitting to the training data.