Machine Problem 2: Naive Bayes Classifier on MNIST

By Vincenzo Marconi

Problem Statement

On this assignment we are to examine how correct the Naive Bayes Classifier (NBC) is and analyze how much of a difference it is to KNN. We perform the same 3 test protocols as in KNN but in this assignment we are to do each with 2 different cases. With the Maximum likelihood Estimation MLE we are to perform test protocol 1 and 3, and for the MAP rule we are to perform all 3 test protocols.

For this assignment, I performed all 3 test protocols with a hybrid MAP and MLE rule. This hybrid rule is nothing more than just the MAP rule with the following alpha values: I used alpha values starting from 0 to $2^{15}/100$. The alpha values followed from 0 with 1 and then I set it to 2^k for $k = 0,1,\ldots,15$. Because the program was really fast I performed this hybrid test on all test protocols to examine more closely what would happen. If you see below, no big difference occurred but something interesting was discovered.

Results

For the case of classifying handwritten digits, the NBC is not very effective. On average, the classifier generated 20-22% error which is much greater than KNN. The alpha values or the pseudo counts did not make much of a difference as the number of pictures of each class was already somewhat evenly distributed. As for the alpha that generated the smallest value, none could be found; all alpha values generated exactly the same error to 2 decimal places.

Each digit class had about 5000-6000 pictures. The results of said alpha values for the hybrid testing can be seen in the figures on the next page. Figure 1 displays the results of the 10-Fold cross validation and Figure 2 the 5-Fold. Each graph suggests that the data is barely if not at all affected by the alpha values. Therefore I believe the alpha values do not reduce overfitting.

When calculating the Gaussian probability for each pixel it was discovered that adding a very small value to the variance, in particular a value close to the smallest variance decreased the error dramatically. When this value is added, the Gaussian distribution becomes wider and increases the class conditional probability of each pixel processed by the classifier. In this dataset, this feature would be a great thing as the data would be blurred, but when the handwriting style changes many Type 2 errors would occur. This appears to be where the overfitting can occur.

If we were examining one persons constant handwriting then the NBC might be able to beat KNN and produce lower errors. I believe NBC loses correlation between pixels because of the naivety in this particular classifier and KNN did not.

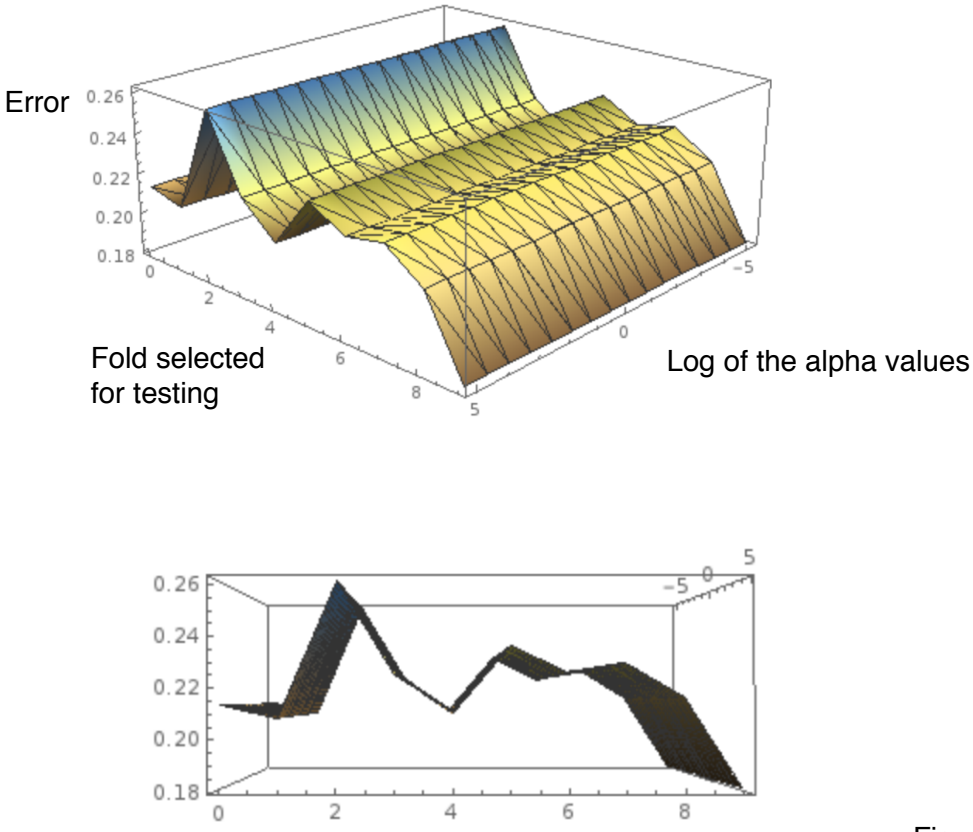

Figure 1

Figure 2