Project update of handwritten digit recognition

Xin Wang

October 2020

The first two algorithms, ridge regression and LASSO, have been investigated. All the 60,000 training cases and 10,000 test cases have been used. And Results and code can be found here.

1 Ridge regression

The experiment on ridge regression can be found in ridge_regression.ipynb.

The total number of singular values of training matrix is 784, of which about 100 values are most important. With only the 100 singular values preserved, a comparable accuracy can be achieved. But no improvement in accuracy is observed.

First, we use all the ten digits, 0-9, as the valid label set. The overall error rate is very high, 74.4%, which is slightly better than jsut random guess. And tuning the parameter λ would yield very different choice using training data directly or using a holdout set in advance.

Then we use a series of binary classifier to train the model. The overall error rate is reduced to 14.6%.

The detailed error count for each digit has been shown in Figure.1.

The error rate comparison among digits can be found in Figure.2.

2 LASSO

The experiment on LASSO can be found in lasso.ipynb.

The detailed error count for each digit has been shown in Figure.3.

The error rate comparison among digits can be found in Figure.4.

3 Plan for next step

From the results above, we can see that there are several digits which are particularly hard to predict correctly. Thus, it should be interesting to explore why those digits are more difficult to predict. This may provide some hint as for how to optimize the model.

Besides, the third algorithm, neural network, should be explored.

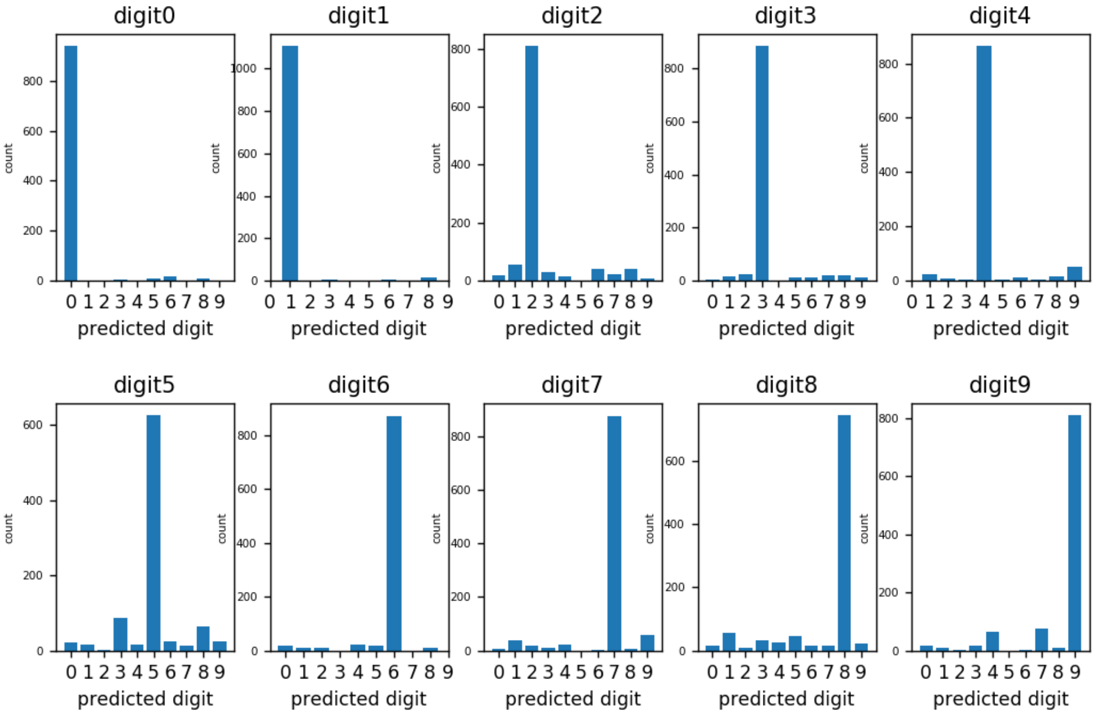

Figure 1: Error count distributions of each digit for ridge regression.

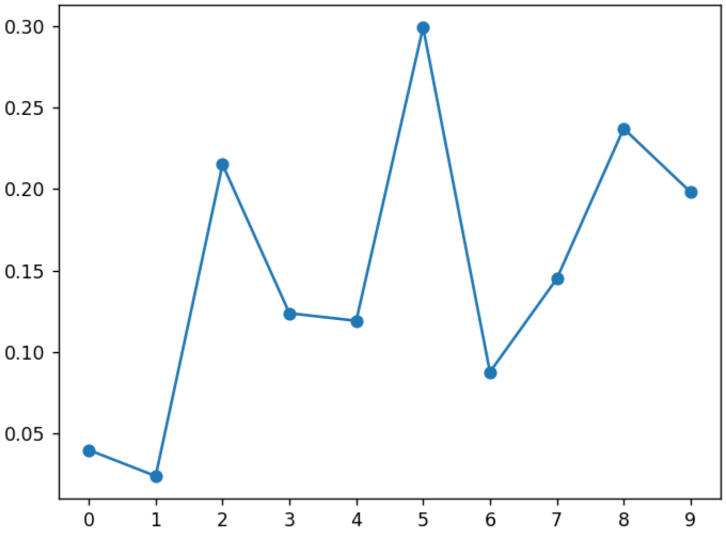

Figure 2: Error rate of each digit for ridge regression.

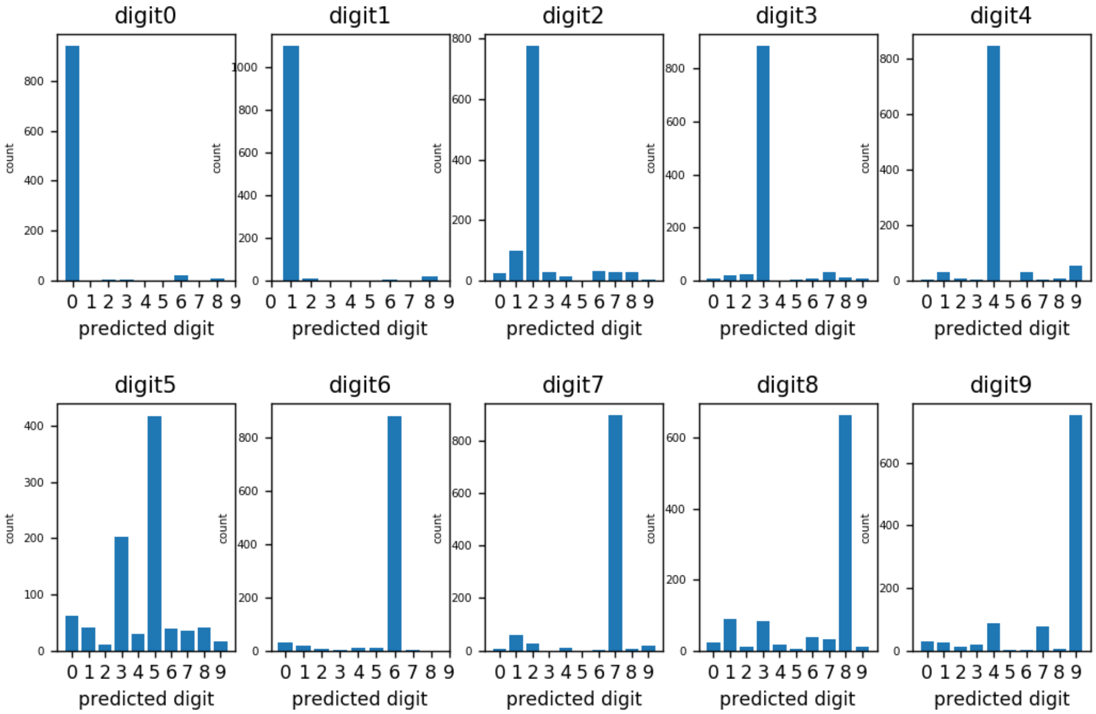

Figure 3: Error count distributions of each digit for LASSO.

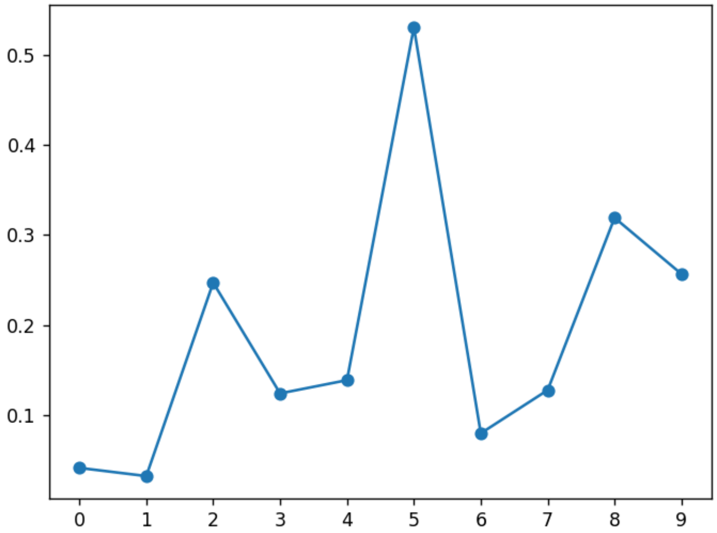

Figure 4: Error rate of each digit for LASSO.