SENG474 Assignment 2

Jonathan Skinnider V00207396

February 28 2020

Note: a reduced size data set was used throughout the assignment. Only 3000 data points from either category where used.

1 Logistic Regression

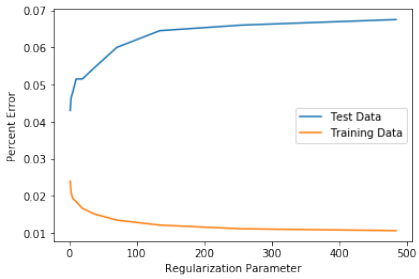

Figure 1: Training and Test Error as Regularization Parameter Varies

Figure 1 shows training and test error varying with different regularization parameters using logistic regression. As the regularization parameter C increases, the amount of regularization occurring during the regression decreases. From $C = 70$ to $C = 484$, the error of the algorithm on the testing data set increases from 6% to 6.75%. All the while, the training error is constant or decreasing. This is evidence of over-fitting, since the algorithm is getting better at predicting the training data (latching onto spurious correlations) however while doing so is actually getting worse at predicting the unseen testing data.

Also observe in the figure when C, the regularization parameter, is set very low. This corresponds to a large amount of regularization, and thus a very simple model. It is at this point that the test and training error are the most similar. Underfitting is defined high bias and low variance. The low variance is demonstrated by the test and training errors being so similar. The high bias is demonstrated by the *average* error over the total 8000 samples comprising the test and training sets being the highest.

2 Linear SVM

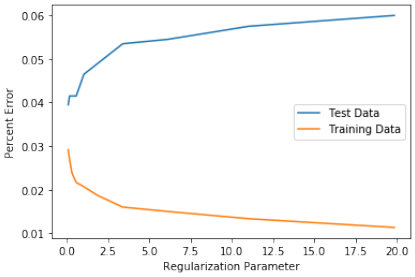

Figure 2: Training and Test Error as Regularization Parameter Varies

We again varied the regularization parameter and reported the test and training error while using a Linear SVM. Figure 2 shows the results.

Again for larger values of C, the figure shows test error increasing as training error decreasing. This demonstrates the algorithm becoming too focused on spurious correlations found only in the training data, becoming to specialized and loosing accuracy on the random test data.

This figure shows underfitting much better than Figure 1. For values C closer to 0, the test and training data approach one another. This corresponds to a lot of regularization creating a very simple algorithm that produces similar results on either data set, but is not as accurate.

3 K-Fold Cross Validation

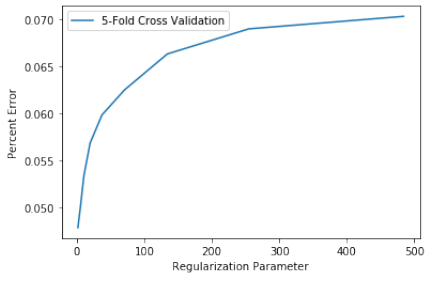

Figure 3: Logistic Regression

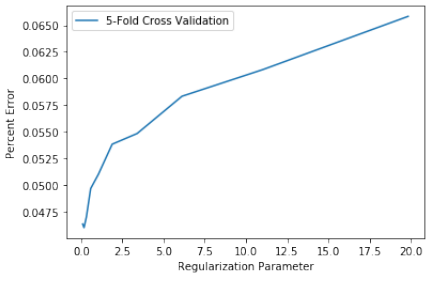

Figure 4: Linear SVM

We completed 5-fold cross validation on the training data set, using both Logistic Regression and the Linear SVM algorithm, in order to determine what regularization parameter to use. The logistic regression, the results where very

similar to testing on only the test data. This might imply that all the data is very uniform, and the training data is actually quite similar to the test data. Similarly the SVM's performance with 5-Fold cross validation was similar to its performance on the test data.

Therefore we will pick $C_{lr} = 1.5$ for logistic regresion, and $C_{SVM} = 0.1$ to compare head to head.

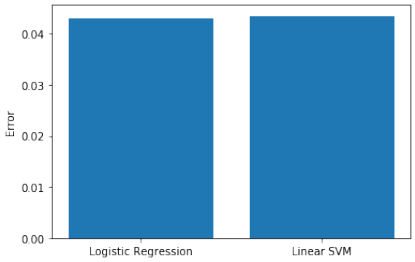

Figure 5: Test Error Using Optimal Hyperparameters

After setting these hyperparameters we trained both models on the entire training set, then tested on the test set. The final results are shown in Figure 5. The SVM did slightly worse, achieving 4.35% error. Logistic regression did slightly better with 4.3% error on the test set.

Since we have 2000 samples in the test set, our 95% confidence interval for these results are

$$\frac{1.36}{\sqrt{n}} = 0.0304 = 3.04\%$$

Therefore both the results lie within the confidence interval of the other, we are unable to conclude that one method is definitively better than the other.

4 Non Linear SVM

We completed the 5-Fold cross validation to find an optimal C in the same range as for our linear SVM. Figure 6 shows one such plot, for $\gamma = 3^{-6}$:

Four more such tests where conducted up to $\gamma = 3^{-1}$. The optimal value for C was always found to be high, 11.01996. This is in contrast to the optimal value for the linear SVM being very low ($C = 0.1$).

Using these optimal parameters we then created an SVM on the full data set, and tested on the full test set. Figure 7 summarizes the results.

In summary, with $\gamma = 0.00541$, the classifier achieves a test error of only 2.55%. This is a 2% improvement from the linear SVM and logistic regression.

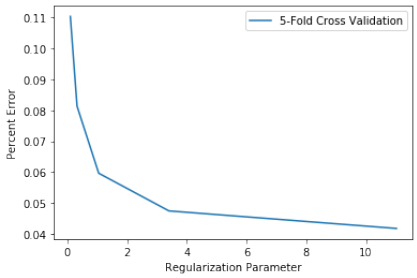

Figure 6: 5-Fold Cross Validation with $\gamma = 3^{-6}$

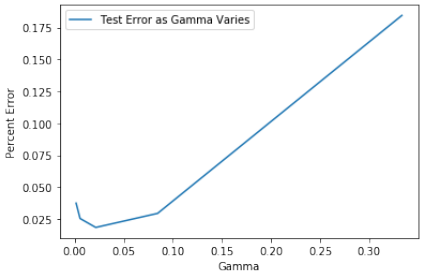

Figure 7: Test Error with Optimal (γ, C_γ)

However this is still within the 3% confidence interval defined earlier. So though the non-linear SVM *appears* to do a better job in classification than the linear SVM, one cannot say with much confidence.