Assignment 2 Report

Anoop (2015CS10265)

March 11, 2018

Text Classification

Multinomial Naive Bayes

Train Accuracy = 0.68448
Test Accuracy = 0.38752

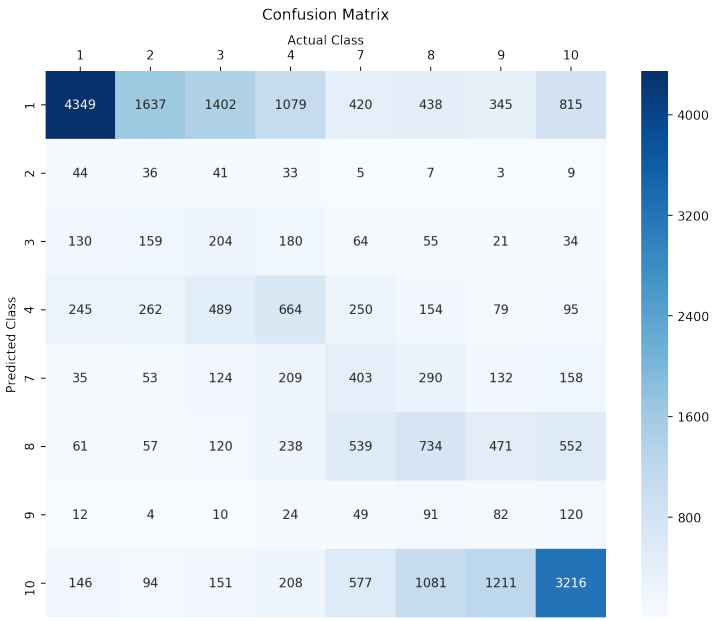
Random Guessing

Test Accuracy = 0.07264

Majority Class

Test Accuracy = 0.20088

Multinomial Naive Bayes - Confusion Matrix



Micro F1 score = 0.2793
Macro F1 score = 0.7800
Class with highest diagonal entry = 1
From the low Micro F1 score, we can infer that there is class imbalance. Also from the Confusion Matrix we can observe that there are more predictions of class 1 and 10. This means that the model is influenced a lot by the class imbalance.

Multinomial Naive Bayes - With Stemming

Train Accuracy = 0.6798
Test Accuracy = 0.38684
There is not much increase in accuracy as compared with the model with no stopword removal and stemming. This is because of class imbalance that we observed in from the confusion matrix previously.

Multinomial Naive Bayes - Feature Engineering

Unigrams + Bigrams
Train Accuracy = 0.40352
Test Accuracy = 0.34868

Normalized TF-IDF
Train Accuracy = 0.384
Test Accuracy = 0.34604

Sublinear TF
Train Accuracy = 0.35912
Test Accuracy = 0.35144

Complementary Multinomial Naive Bayes - With TF-IDF

Train Accuracy = 0.53036
Test Accuracy = 0.35816

Feature Engineering Report

After trying many feature engineering techniques, there is no significant improvement over baseline Multinomial Naive Bayes model. Even with stopword removal and stemming, there is no improvement over the baseline model. This is mainly due to the class imbalance in the training data. After adding more relevant features, there is a reduction in training accuracy. From this we can infer that our original baseline models were overfitting the training data as there was very high difference in train and test accuracies of baseline model.

MNIST Handwritten Digit Classification

One-vs-One Model

Train Accuracy = 0.9341
Test Accuracy = 0.9328

One-vs-All Model

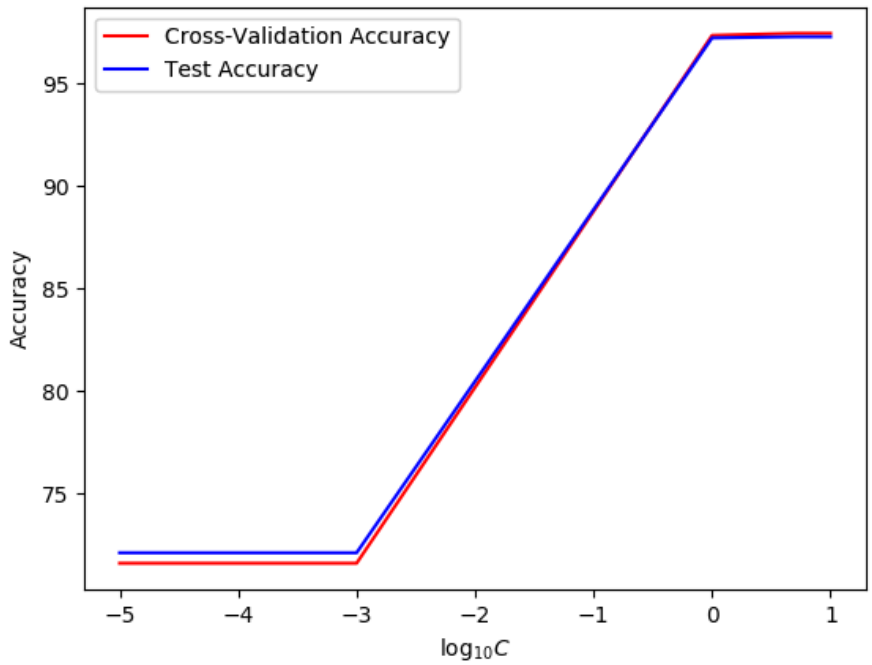
Train Accuracy = 0.87715
Test Accuracy = 0.887

Multi-Class SVM using LIBSVM

Linear Kernel
Train Accuracy = 0.40352
Test Accuracy = 0.34868

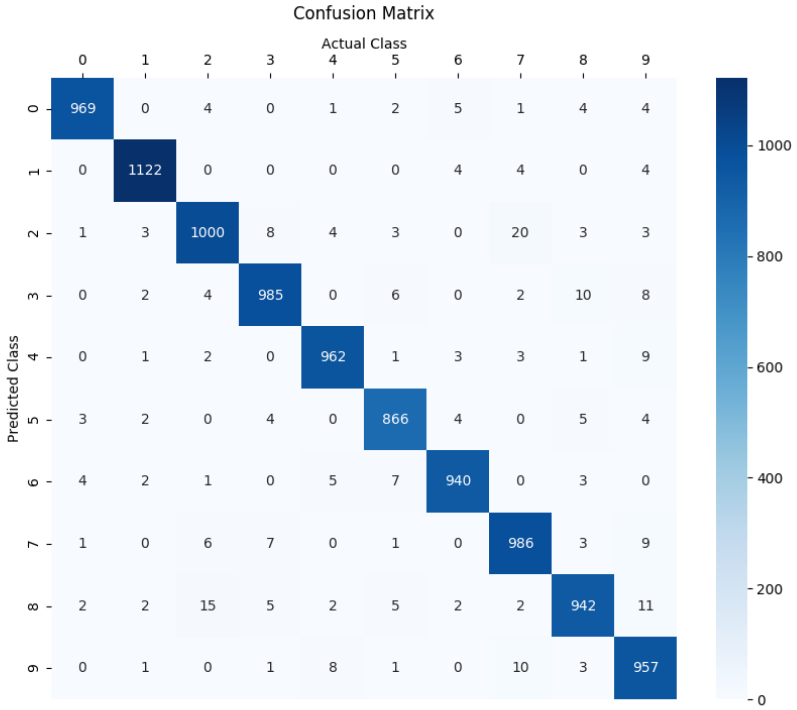
Gaussian Kernel
Train Accuracy = 0.384
Test Accuracy = 0.34604

Cross-Validation using LIBSVM

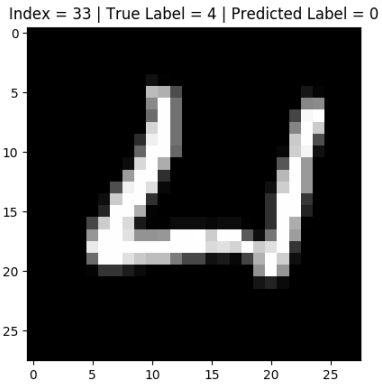


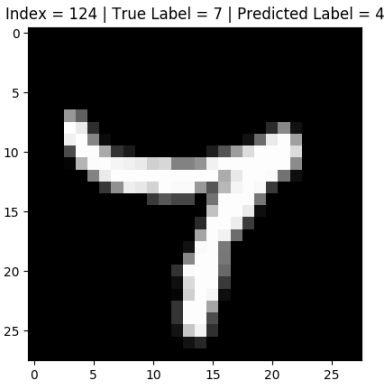
From the plot we can observe that increasing C is increasing the Cross-Validation as well as Test accuracy. This means by choosing a small value of C we were underfitting the data. From the plot we can observe that there is no much difference between Cross-Validation and Test accuracy. This means that we are not overfitting the data at any point.

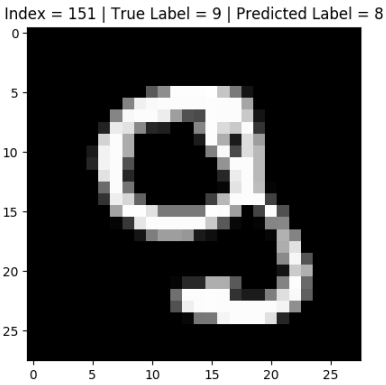
Best SVM Confusion Matrix

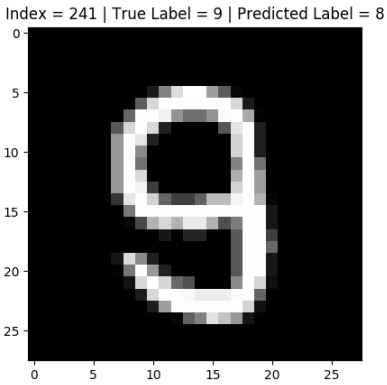


Visualization of Misclassified Examples









From the mispredictions and confusion matrix we can see that the classifier is confused between similar looking numbers like 4-9, 4-7, 8-9.
Easiest class to classify = 1
Most difficult class to classify = 7