Assignment 2 Data Mining Report

Kushank Gautam
V00843403

Logisitic Regression

Logistic Regression is a statistical model that uses logistic functions to models binary variables. A logistic model has variables with only two possible values such as 0,1 or Pass/Fail etc. Logistic Regression uses two types of regularization for penalizing and in this assignment we are using L2 regularization also known as Lasso Regression, where we add the absolute value of magnitude of coefficient as penalty term to the loss function. This type of regularization shrinks the less important feature's coefficient to zero, therefore removing the weightage of some features altogether. This is usefule in this case such as many pixel values for predicting sandals and sneakers were insignificant to our prediction.
The following table provides with Testing and Training error for Fashion-MNIST data when modeled with Logistic Regression, using L2 regularizatoin and taking C=1, a=1.

Table 1: Training & Test Errors Logistic Regression

Regularization	Training Error	Testing Error
0.1	0.047	0.0585
0.541	0.0471	0.0585
0.877	0.0467	0.0595
1.154	0.0473	0.056
1.321	0.0475	0.0585
1.469	0.0469	0.056
1.615	0.0472	0.058
1.794	0.0461	0.056
1.908	0.0441	0.056
2	0.0434	0.0595

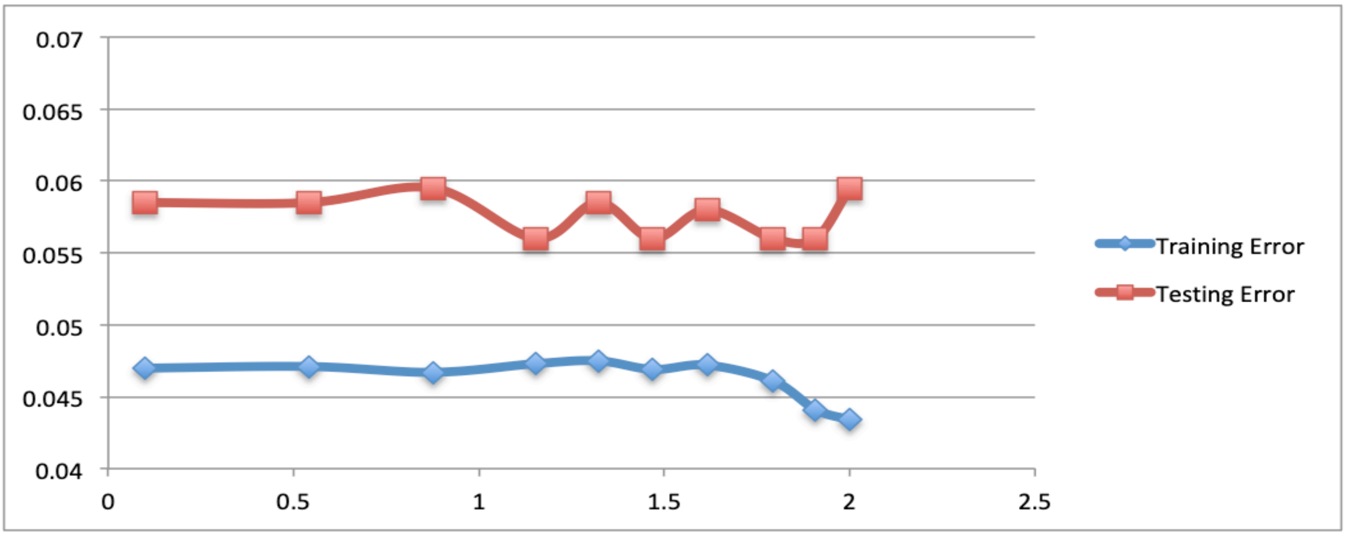

Figure 1: Training & Test Errors vs. Regularization

Analysis

After running the model for training on 12000 samples and testing on a different set of 2000 samples, the result shows a decrease in error for both training and testing as the regularization increases. With very little regularization, there wasn't enough penalty applied and we say overfitting. I kept my regularization between 0 and 2, as advised in the assignment but, it is visible the training and test error shows symptoms of underfitting as the values increase from 1.6 to 2.0 and a similar trend can be expected when going above 2.0.

Support Vector Machine

A support vector machine is a supervised learning method that divideds data into two different kinds and sorts them with a margin in the middle. We are using linear support vector machine as our labels only have two different values, making our data plottable in two dimensions. Based on this we can use a single line(1-D) to distinguish our data plotted on a 2-D plane.

Table 2: Training & Test Errors for SVM

Regularization	Training Error	Testing Error
0.1	0.0199	0.0505
0.541	0.0188	0.057
0.877	0.0172	0.062
1.154	0.0172	0.0595
1.321	0.0172	0.06
1.469	0.0169	0.0605
1.615	0.0171	0.06

1.794	0.0169	0.0595
1.908	0.0169	0.06
2	0.0167	0.06

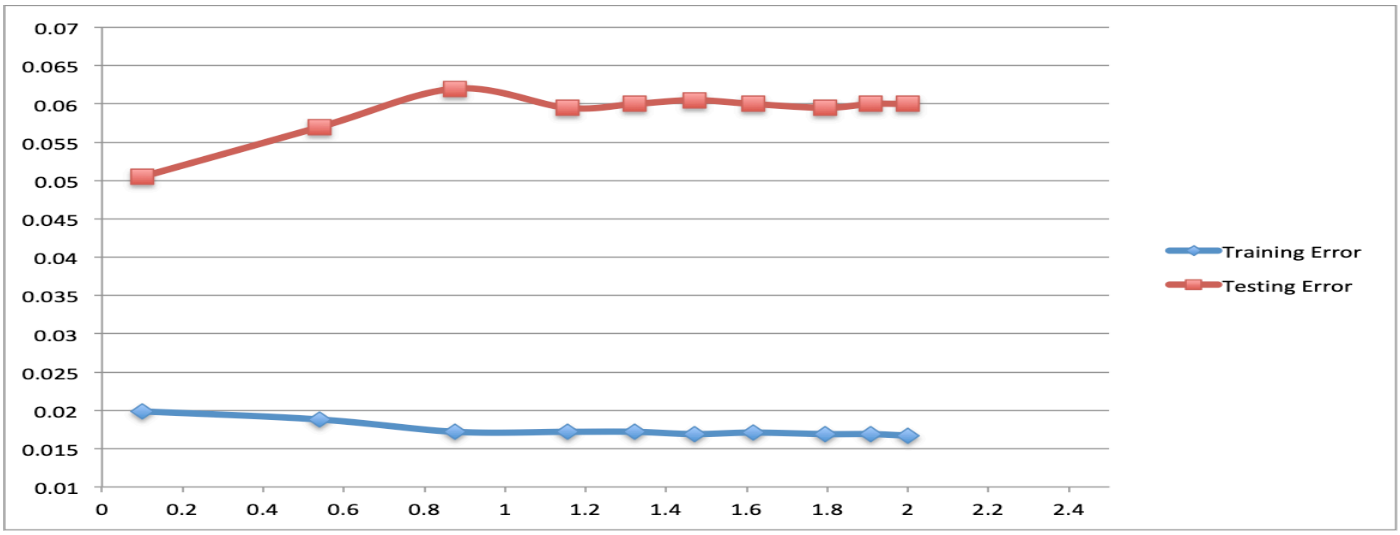

Figure 2: Training & Test Errors vs. Regularization

Analysis

In Support Vector Machine, both training error gradually decreases as the regularization is increased, whereas the test error starts at its lowest point at 0.0505 and then starts increasing but comes to a drop after crossing regularization value of 1.2.

K-Fold Cross Validation

For k-Fold cross validation, I decided to go with 5 folds-splits for my data set as it splits it in the ratio of 4:1, which gives 80% data for training and 20% data for testing, this is believed to be a good estimate.
For both Logistic Regression and SVM using, k-Fold cross validation, I used the *.score* method to find the performance evaluation for every different regularization value.
Provided below are the mean scores of 5-fold cross validation results run with different Regularization values for both Logistic Regression and SVM.

Table 3: Mean Accuracy scores for Logistic Regression & SVM in K-Folds

Regularization	LogReg Mean Score	SVM Mean Score
0.100	0.868	0.95708333333333
0.541	0.8861666666667	0.9535
0.877	0.891	0.9520833333334
1.154	0.8923333333332	0.95216666666666
1.321	0.8925833333333	0.95166666666668

1.496	0.8942500000001	0.951833333333333
1.615	0.8939999999999	0.951833333333333
1.794	0.8947499999999	0.9515
1.902	0.8949999999999	0.95091666666665
2.000	0.8950833333333	0.950833333333334

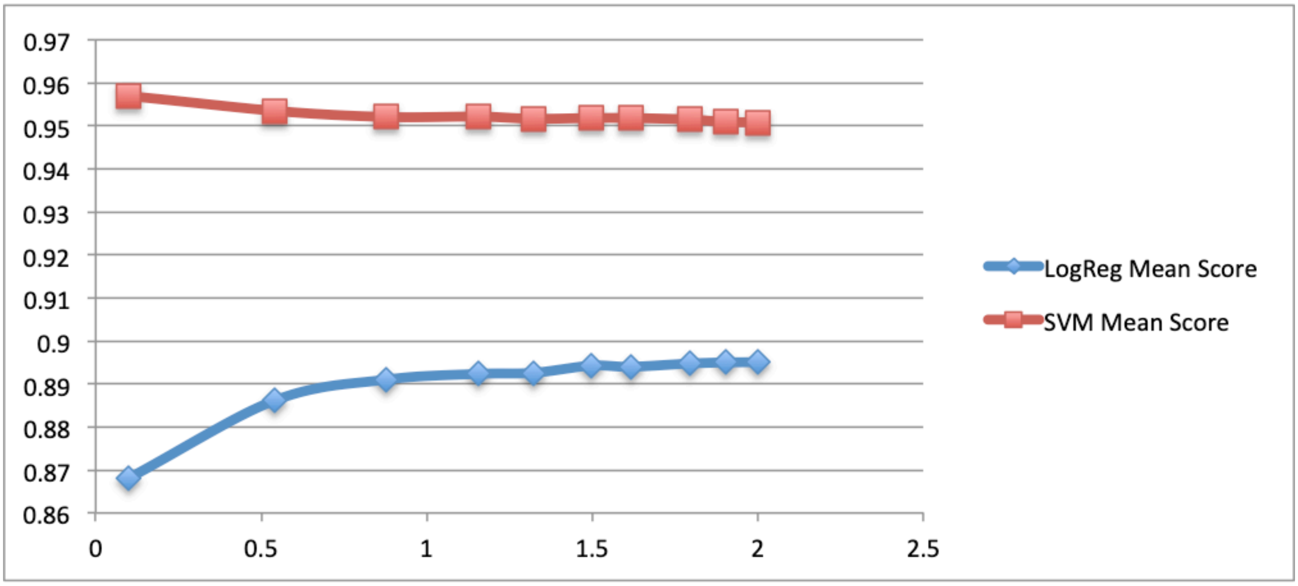

Figure 3: Mean Accuracies of Log Reg & SVM vs. Regularization in K-Fold

Analysis

With different regularization values for both Logistic Regression and SVM a general trend is visible. The accuracy in mean scores for SVM is higher than those of Logistic Regression.

For Logistic Regression the accuracy is low at 0.868 in the beginning but as regularization increases, it is proportionaly with the regularization value and reaches a limit at around 0.895.

For SVM we see a different trend as it starts with the highest value 0.957 and then gradually but slightly decreases as the regularization increases.