CSL 603 - Machine Learning (Lab 2)

Linear and Logistic Regression

Rishabh singh - 2016CSB1054

September 21, 2018

Introduction

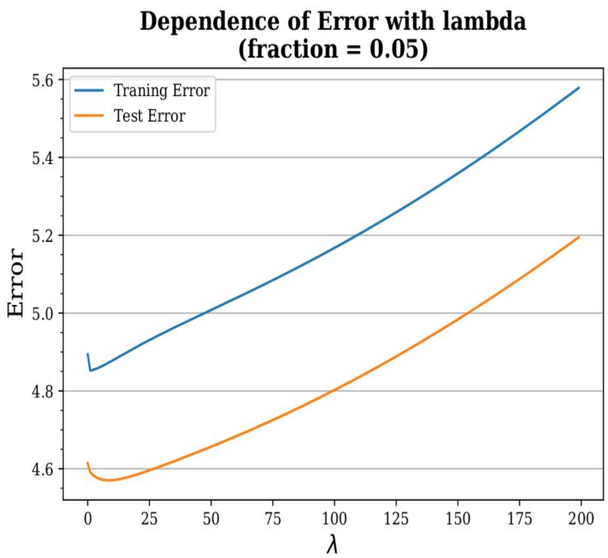
In this lab I learned two different models using Linear and logistic regression. I performed linear regression to predict the age of Abalone (is a type of snail) based on some of its characteristic features like Sex, Length, Diameter etc. In second model I performed logistic using both gradient Descent and Newton Raphson parameter Update Equation. The Logistic Regression is performed to predict whether credit card can be issued to an individual.

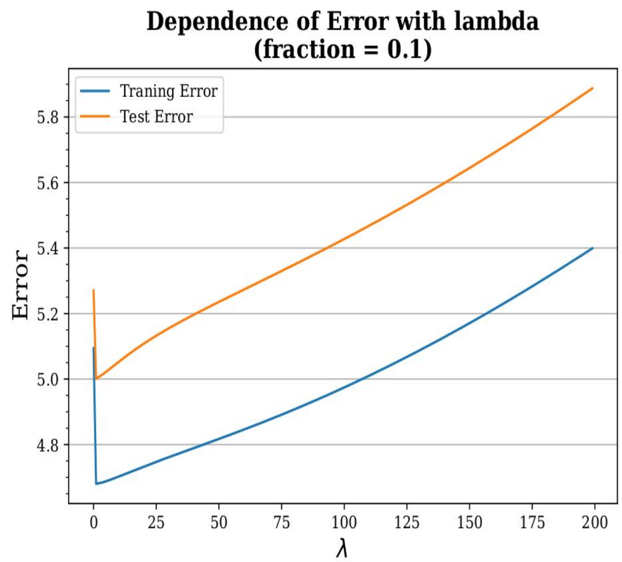
1. Linear Regression

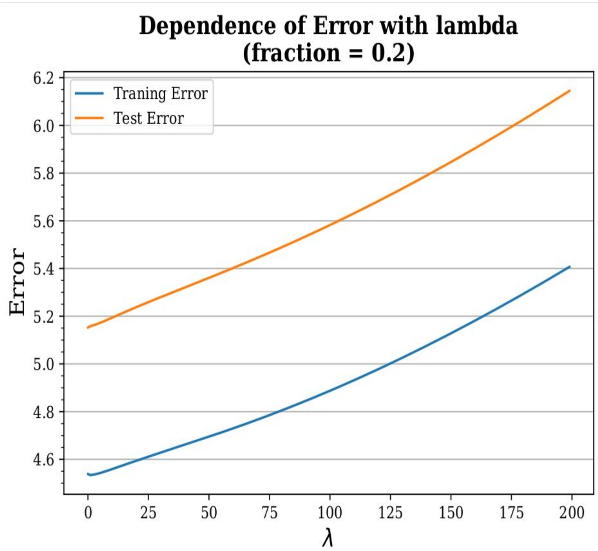
In this assignment our target is to train a model that can learn to find the age of Abalone giving some other parameters. We use linear regression to learn a linear boundary for the prediction and to reduce the complexity of the model I have used regularization with different values of lambda.

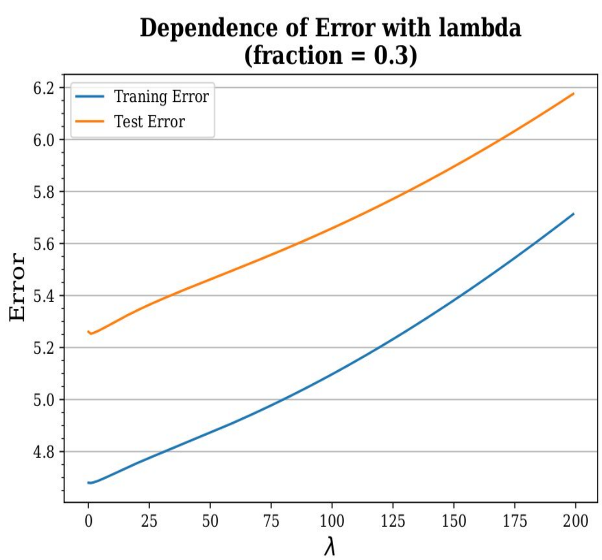
1. For different fraction of (validation dataset)/(training dataset) I standardise the dataset i.e. I converted the training dataset columns to have mean 0 and standard deviation 1.
2. For each fraction I made different lambdas values (100) and applied linear regression on each value of lambda.
3. The graphs obtained for different training set fraction are following.

Dataset 1 Plots:









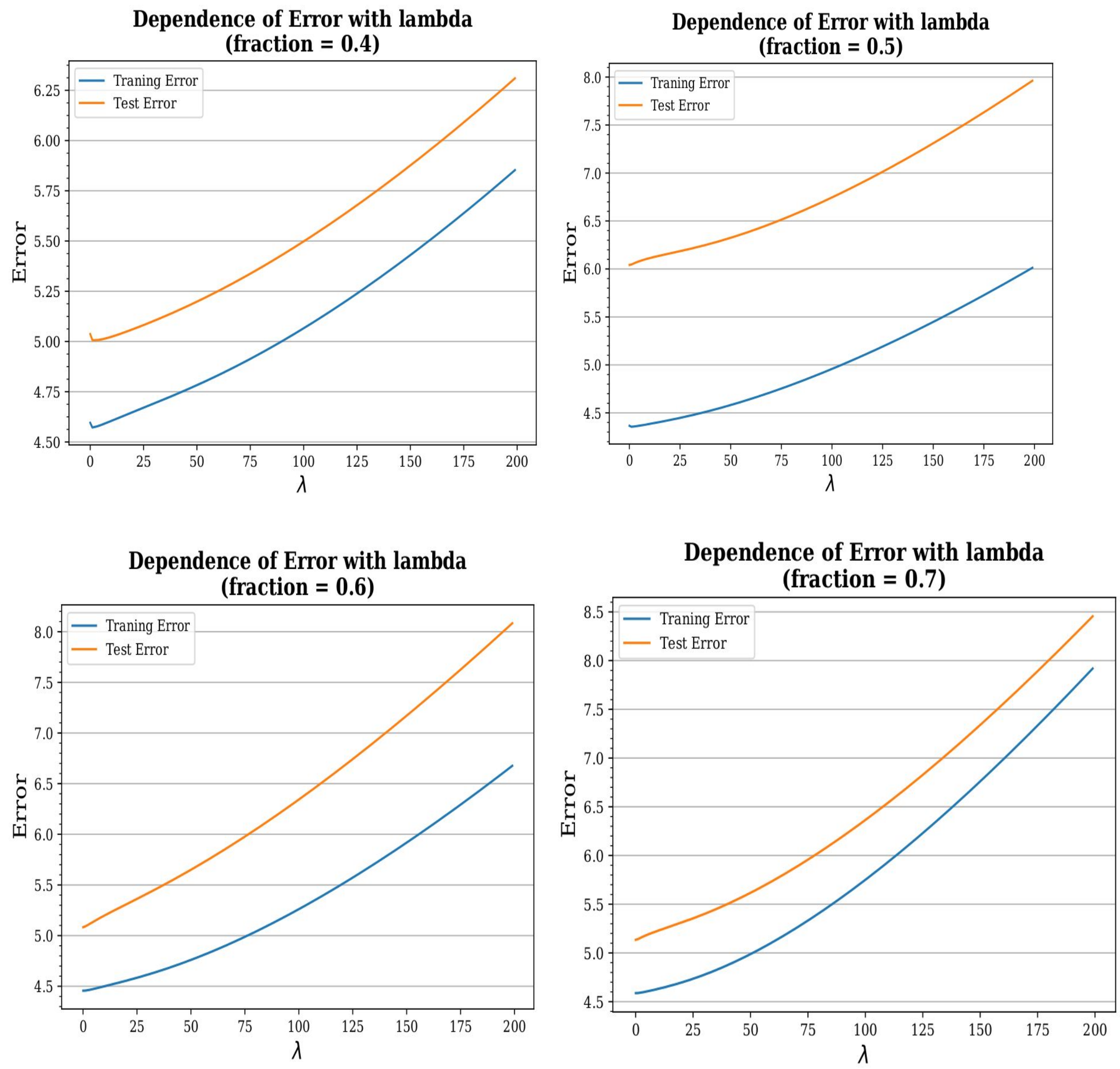
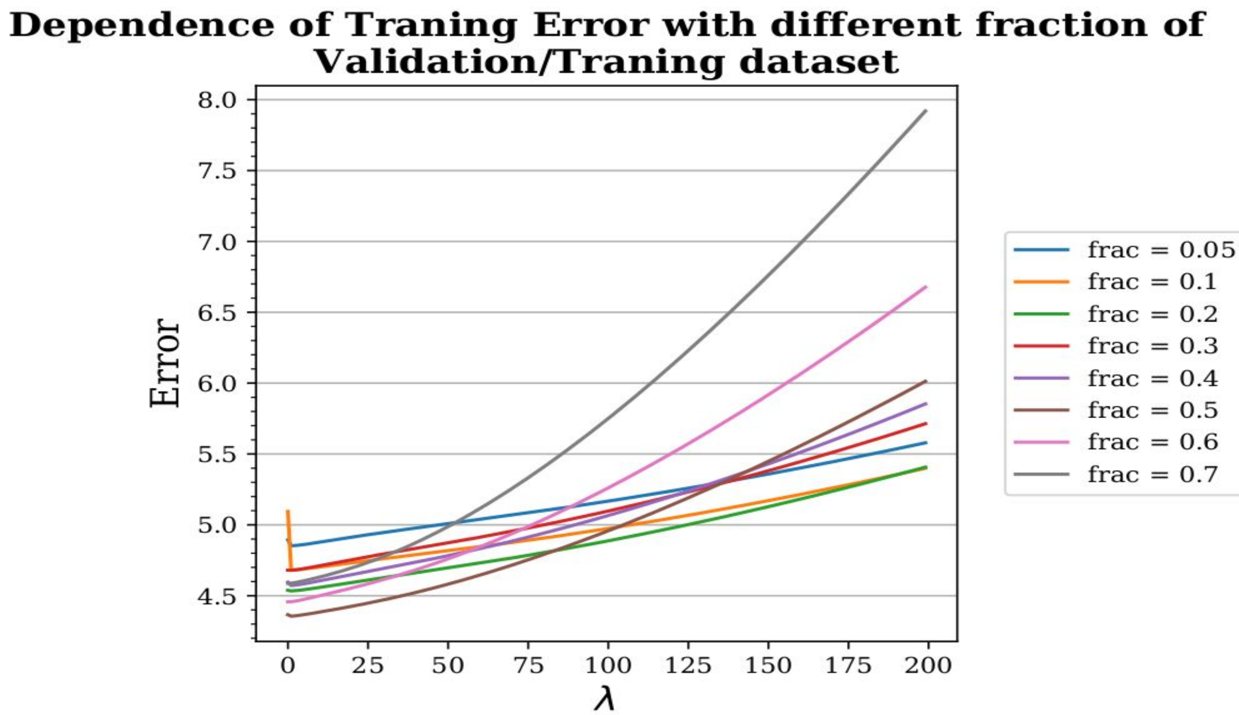


Fig 1 :Showing Effect of λ on Error.



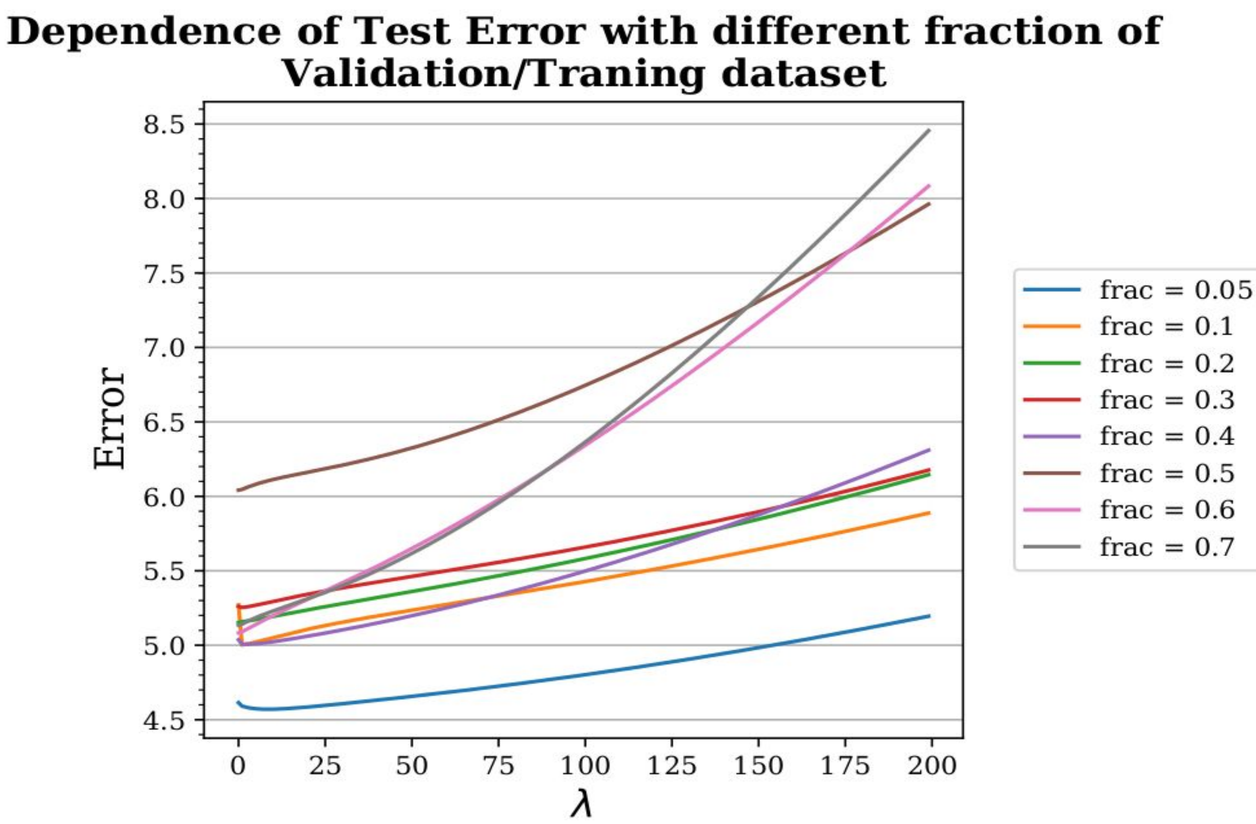


Fig 2 : Effect on Error on increasing λ shown for various fraction.

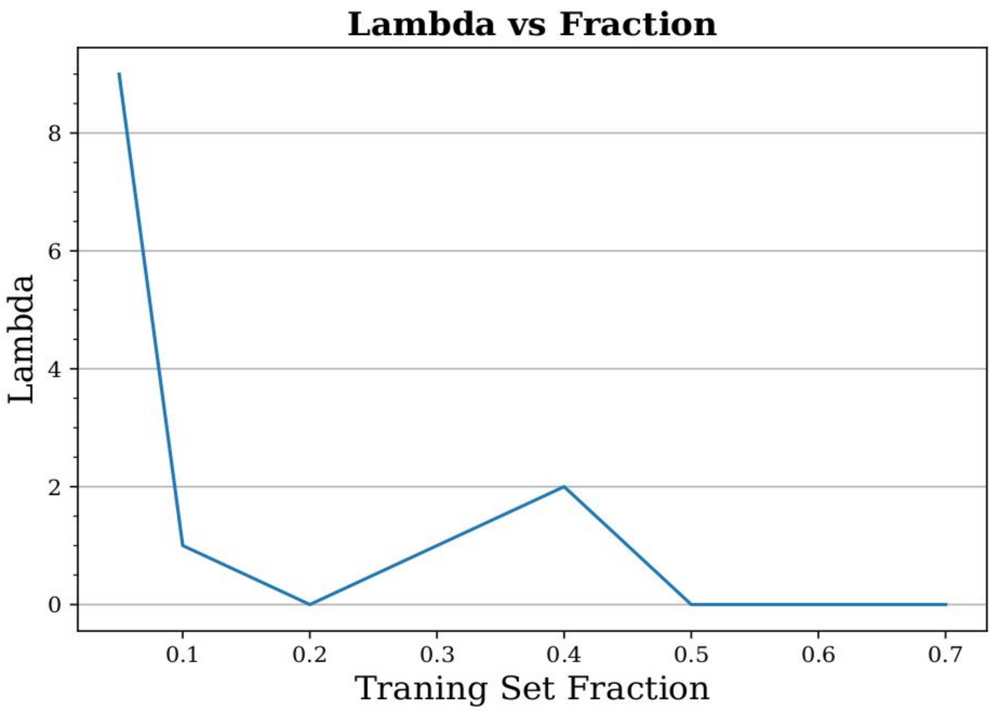


Fig 3 : Plot shows lambda to get minimum error for each fraction.

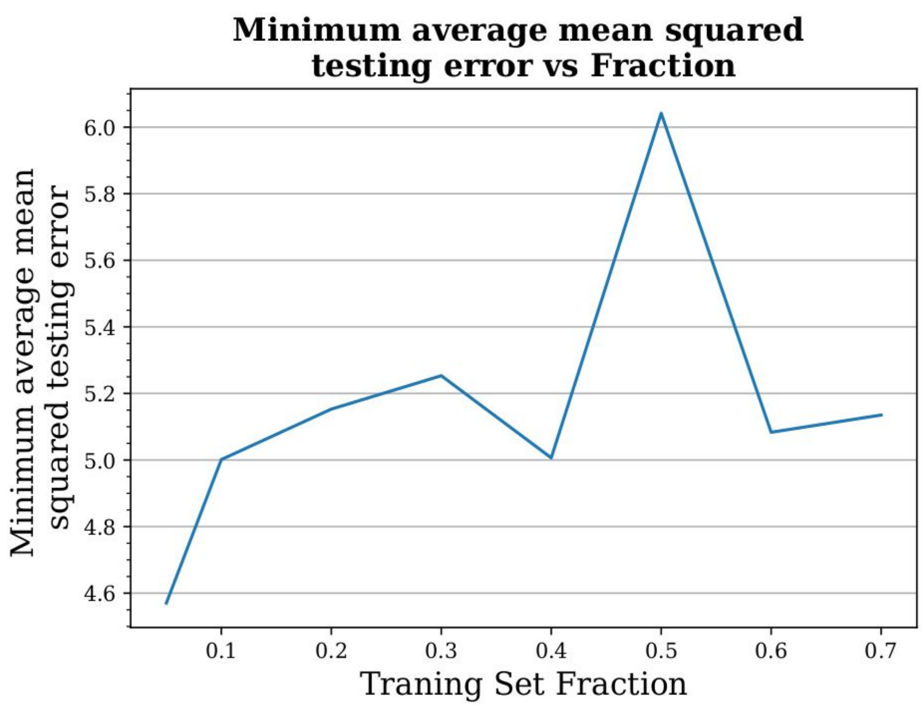


Fig 4: Plot shows the minimum average mean error for different training set fraction.

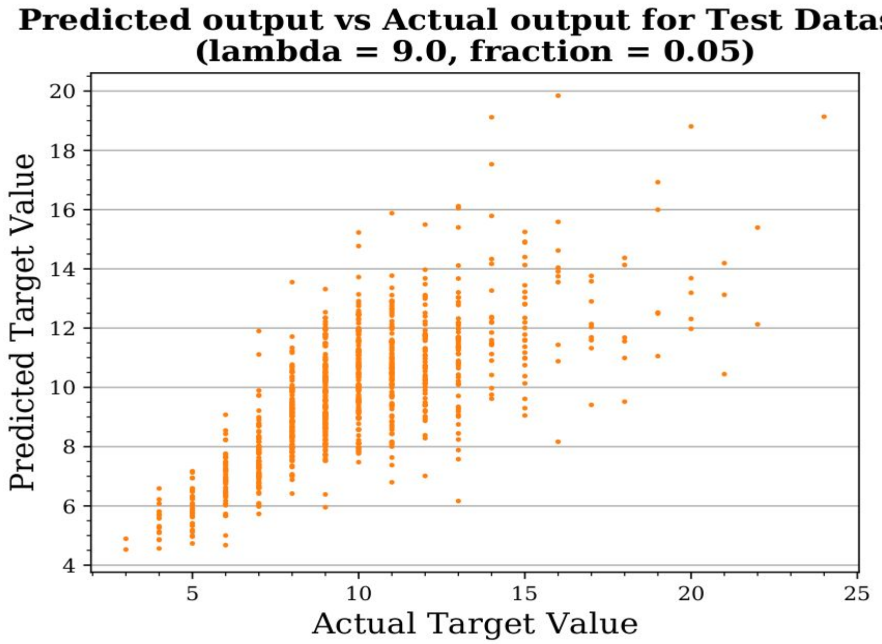


Fig 5: Plot showing the Predicted Output vs Actual Output on Training dataset.

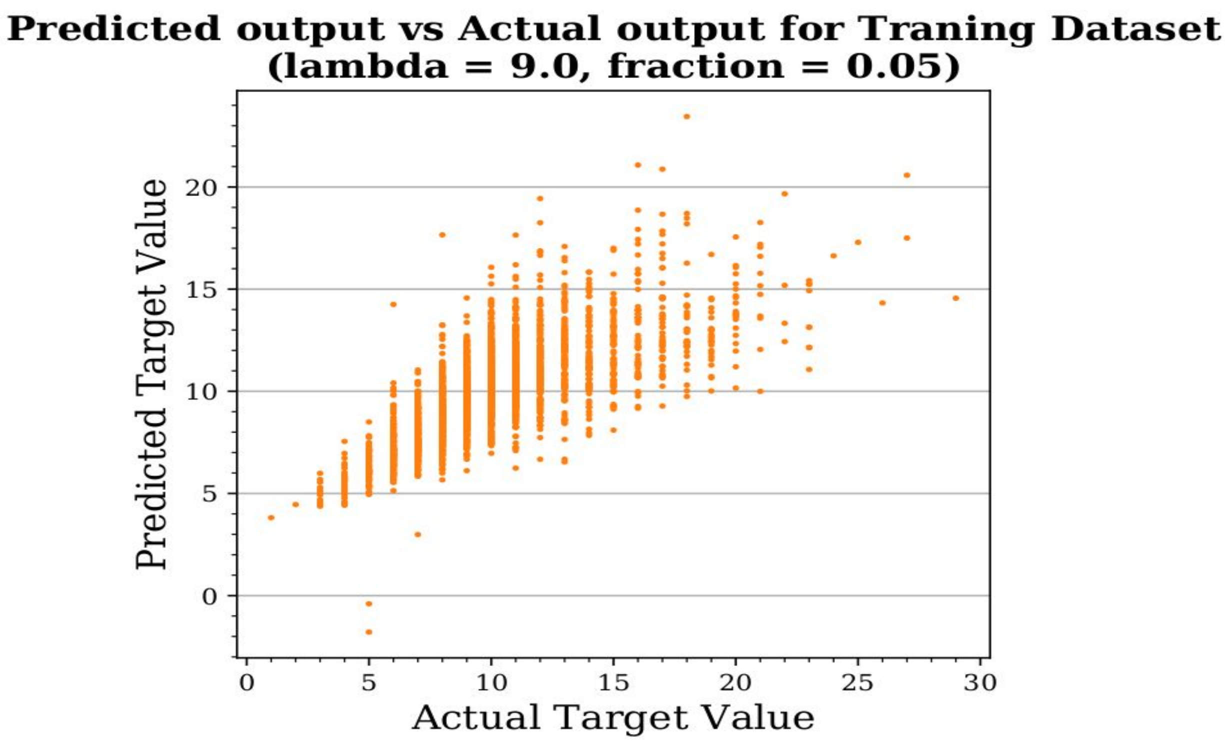


Fig 6: Plot showing Prediction Output vs Actual output on test dataset.

Dataset 2 Plots:

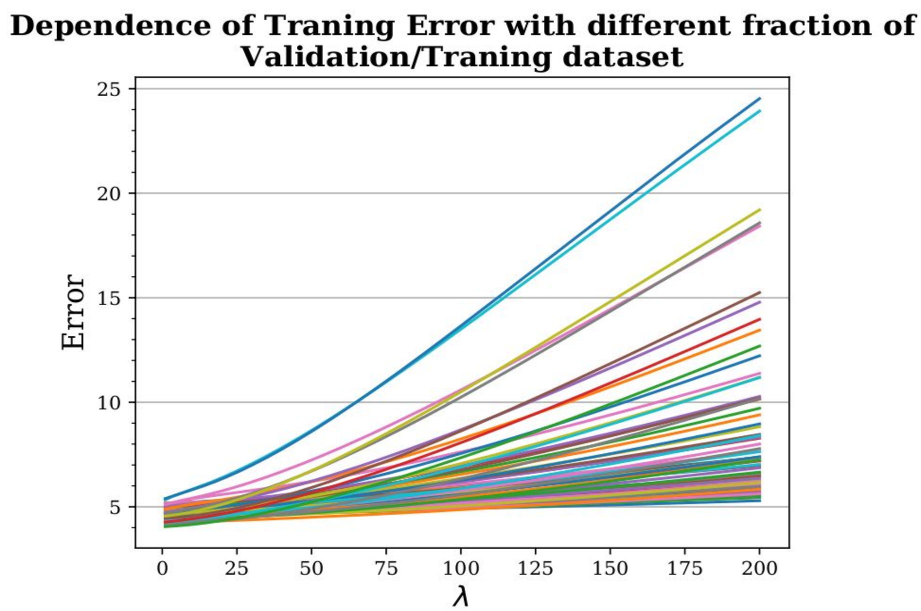


Fig 7: Plot Showing variation of Error on increasing λ for 100 validation/training dataset fraction.

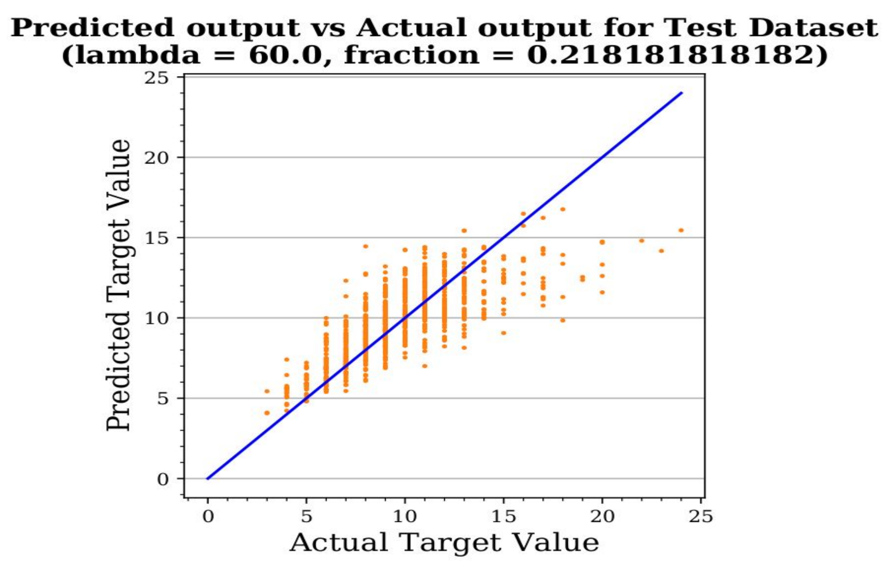


Fig 8: Plot showing Prediction Output vs Actual output on test dataset for the dataset 2.

Observations:

1. On increasing λ I observed that both training and test error increases. This is because the higher order term is penalised which increases bias.
2. The train error will increase with λ because regularization increases bias of the model.
3. The test error for fraction = 0.05 is highest which is because the number of instances is much more in training dataset.
4. Plot 7 and plot 8 shows that there are some of the noises term which are contributing more to the Error on the Test dataset.
5. If lambda is increased after a certain limit (which depends on training set fraction) I observed that train and test errors begin to increase which shows that there is a limit upto which we can regularize the data, above that the model becomes too simple that it is unable to give useful predictions.

a. Does the effect of λ on error change for size of the training set?

From the observation 1 we can say that increasing λ increases the Error on the training increases while the test error first decreases and then increases, which shows that there is a limit upto which we can regularize the data.

b. How do we know if we have learned a good model?

From the plot 5,6 and 8 we can see that the actual value and the predicted value are approximately at 45 degree. It shows we have learned quite a fir model.

2. Logistic Regression

In this Experiment, I used logistic regression to predict whether credit card can be issued to an individual. I used gradient descent and newton raphson method to converge to the minimum error.

1. Plot of the Dataset.

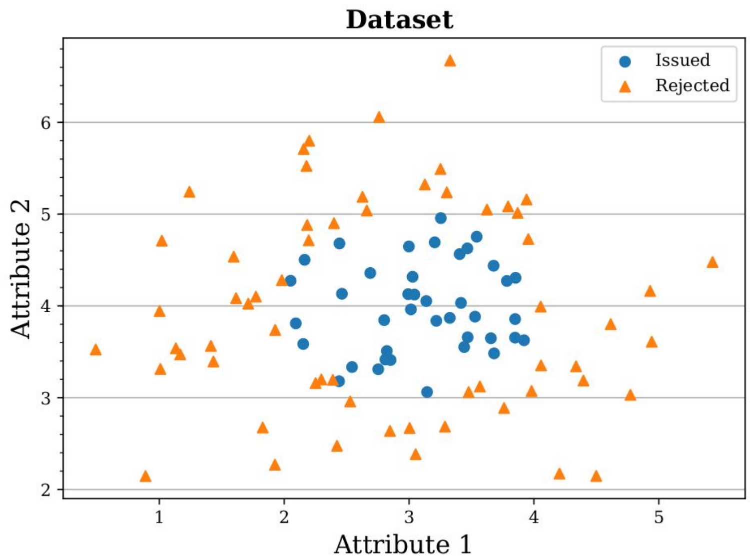


Fig 9: Plot showing the dataset (Different class label are shown with different markers).

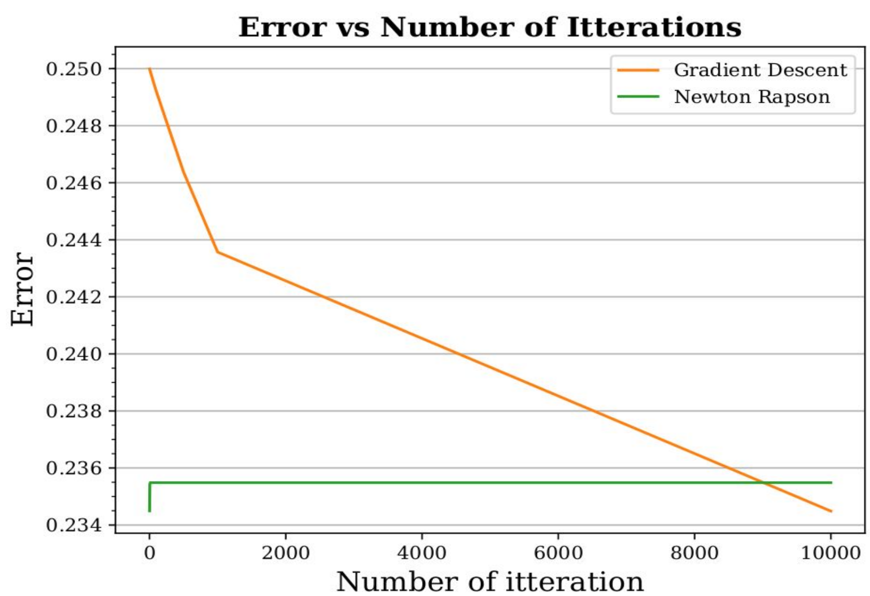


Fig 9: Plot showing the effect of increasing number of iteration on Error.

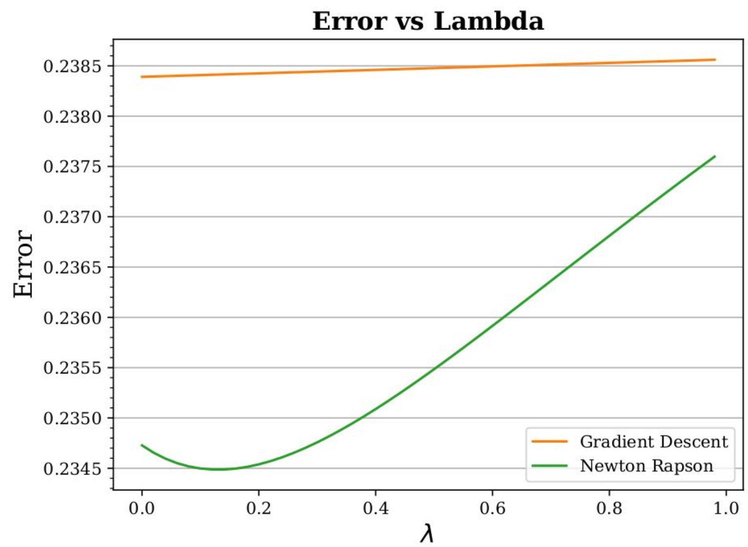
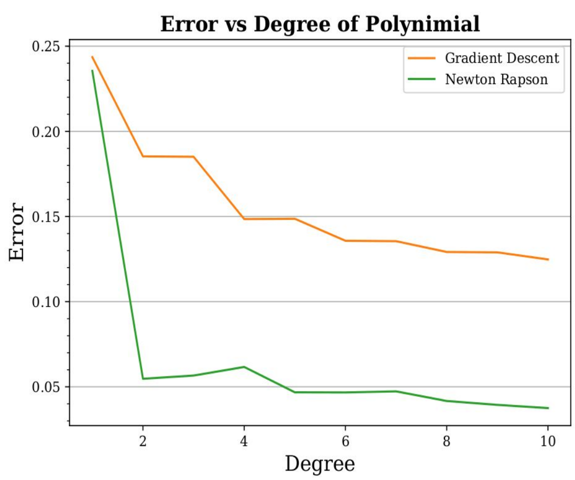


Fig 10: Plot showing effect of error on increasing lambda for fixed number of Iteration.



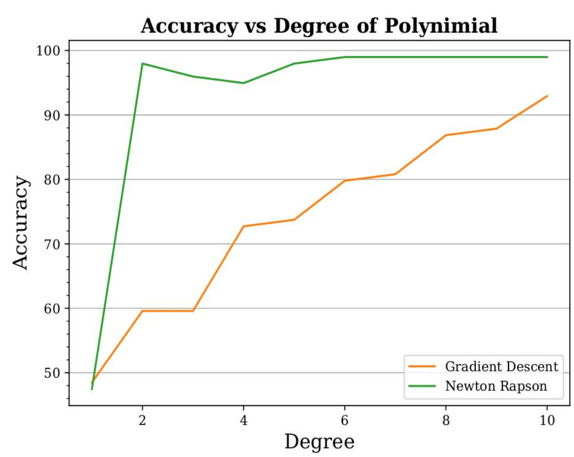


Fig 11: (a) Plot showing Effect of increasing the degree of polynomial on error. (b) Plot showing effect of increasing degree on Accuracy of prediction of class.

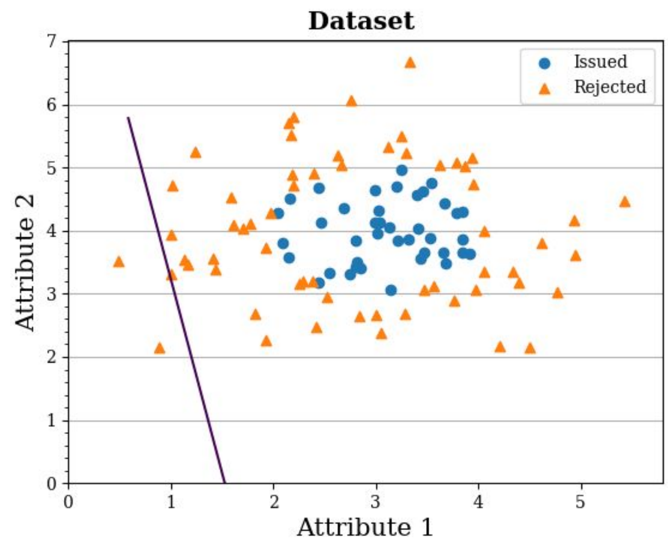


Fig 12: Plot showing the decision boundary for logistic regression.

Observations:

1. The dataset is not linearly separable.
2. From the plot 9 we observe that newton raphson method converges very quickly i.e. in about 100 iteration while the gradient descent takes very large number of iteration to converge.
3. From plot 10 we observe that as we increase λ the value of error continuously increases for gradient descent while it first decreases and then increases for newton raphson method. This can be explained simply by the fact that there is a limit upto which we can regularize the data, above that the model becomes too simple i.e. is unable to give useful predictions.
4. From plot 11 we observe that as we increase the degree of polynomial the error decreases and Accuracy increases. This can be explained by the fact that decision boundary not remains linear and hence it can easily separate the classes.