CSE 574 Introduction to Machine Learning

Programming Assignment – 2

Classification and Regression

Date: April 12, 2017

Group 12

Venkata Saicharan Kolla

Chaitanya Vedurupaka

Anirudh Yellapragada

Problem 1 : Experiment with Gaussian Discriminators

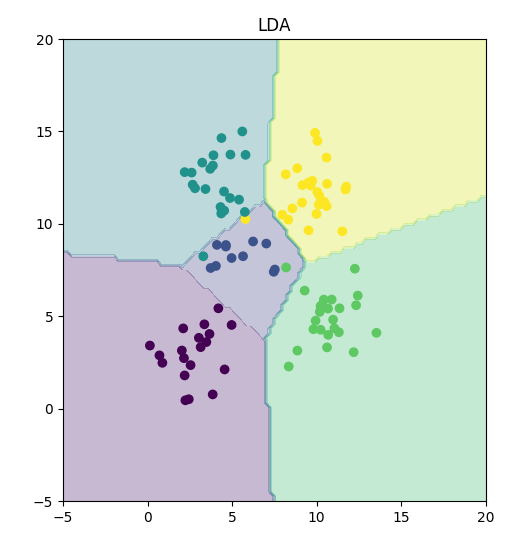
In this experiment, we implement Linear Discriminant Analysis and Quadratic Discriminant Analysis. Both the models have been trained on sample\_train data to get the mean and co-variance matrices.

Below are the accuracies obtained on the sample\_test data in both the scenarios:

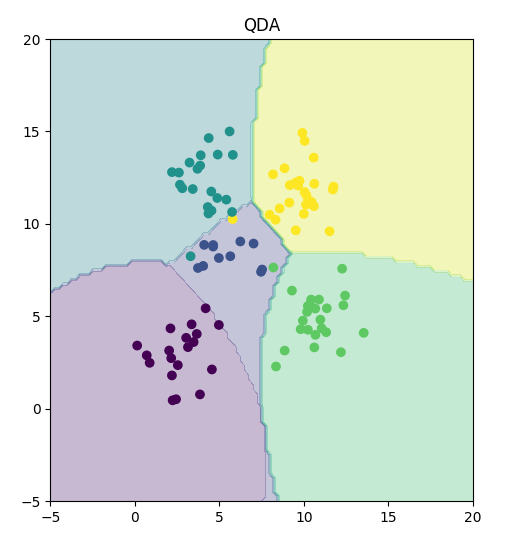
LDA Accuracy = 97.0

QDA Accuracy = 96.0

Plot of discriminating boundaries in LDA :



Plot of discriminating boundaries in QDA :



The decision boundaries in Quadratic Discriminant Analysis(QDA) are determined by functions/equations which are quadratic in nature, which is not the case in Linear Discriminant Analysis(LDA). This is why the LDA can learn linear boundaries whereas QDA can learn quadratic boundaries. This makes LDA a much less flexible classifier when compared to QDA. Because of this higher flexibility, QDA can accurately model a wider range of problems in comparision to LDA.

Problem 2: Experiment with Linear Regression  
  
In this experiment, we have implemented ordinary least squares method to estimate the regression parameters by minimizing the squared loss. Below are the results for the mean square error in different scenarios.

Mean squared error (MSE):

1. Training Data :

MSE on training data without intercept is 19099.44684457.

MSE on training data with intercept is 2187.16029493.

1. Test Data :  
   MSE on test data without intercept is 106775.36155789.  
   MSE on test data with intercept is 3707.84018132.

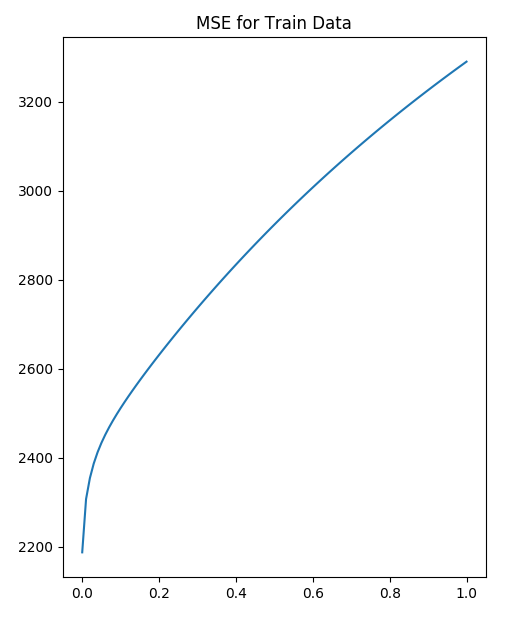
Interpretation of MSE results on training and test data:

1. Clearly, the mean square error seems to be low in the case where we estimated the regression parameter with the intercept/bias term. This is true for both training and testing data. A significant decrease of mean square error (almost equivalent to 29 times decrease) has been observed in case of test data. In case of training data, the decrease in mean square error (almost equivalent to 9 times decrease) is not as significant as in the case of test data.
2. Moreover, in the case where we do not use intercept, the MSE on test data is 5.6 times the MSE on training data. We can clearly observe the importance of using the intercept/bias term, in which case, the MSE on test data is 1.7 times the MSE on training data.

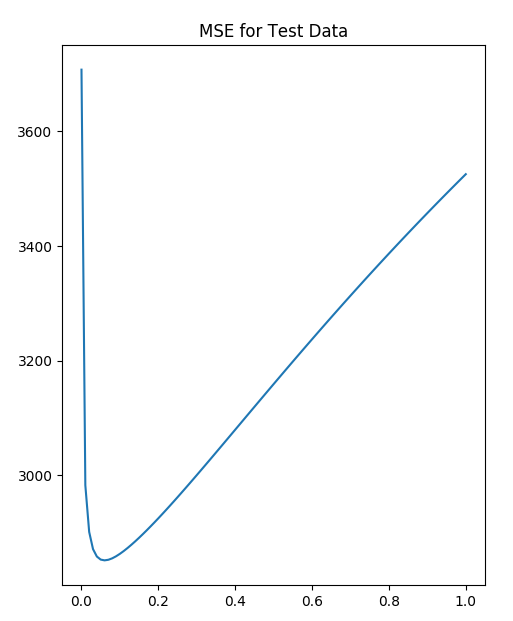
Problem 3 : Experiment with Ridge Regression  
  
In this experiment, we have implemented regularized square loss method in estimating the regression parameter. As we have seen in the previous experiment, we can obtain best results when the intercept/bias term is included. Hence let us consider this experiment on the data with intercept.

Below are the plots of MSE on training and test data for lambda values in the range 0 and 1 in steps of 0.01.

Training data : Lambda(X- axis) VS MSE(Y-axis)



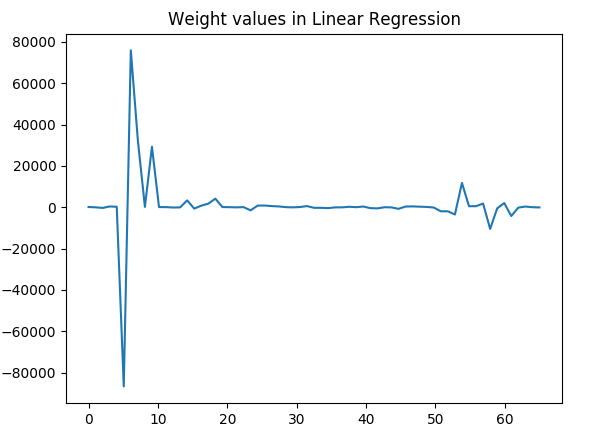
Test data : Lambda(X- axis) VS MSE(Y-axis)

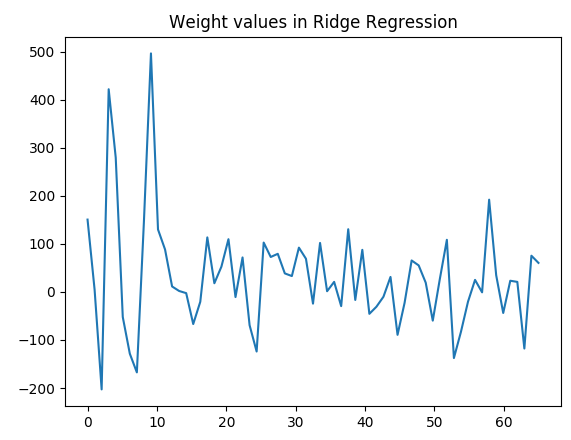


The MSE on training data is in the range of [2187.16029493, 3289.7612813]. The optimal value of lambda (the value for which the MSE on test data is minimal) is 0.06. The value of MSE on test data for the optimal value of lambda is 2851.33021344. In order to avoid the problem of over-fitting, we introduce l2-regularization concept to the existing Linear Regression model making it a Ridge Regression. Ridge regression also helps in reducing the impact of correlated inputs.

Comparing linear regression model with ridge regression :

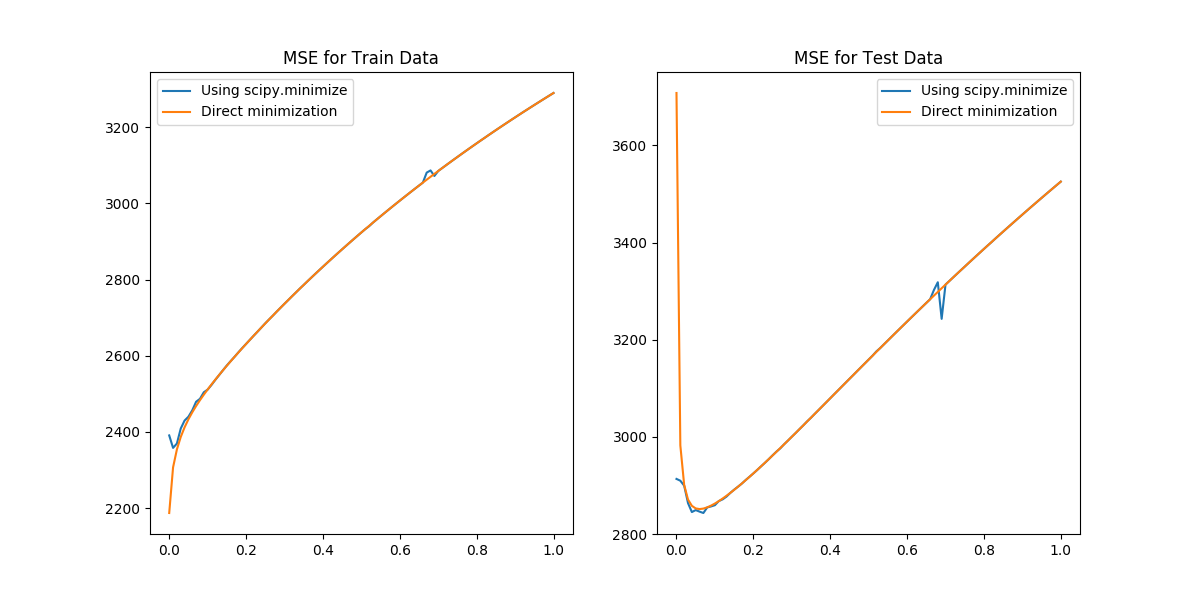
The MSE values on training data calculated using ridge regression seems to go higher than the MSE obtained using linear regression, as we increase the value of lambda, a ridge regression parameter. However, the advantage of ridge regression is evident on test data. The lowest MSE value obtained in ridge regression is relatively lesser than the one obtained in the linear regression model. The optimal value of lambda is the one where the MSE on test data is minimum. Below are the plots of weight vectors obtained in linear and ridge regression.

**

**

The magnitude of weight vector in case of ridge regression are very low in comparision to the one obtained in the linear regression.

Problem 4 : Using Gradient Descent for Ridge Regression Learning



The Gradient Descent method is used when there is difficulty in computation of inverse of matrix. This involves minimizing the loss function for estimating the weights. The plots of MSE vs λ in both cases are almost similar. This is because we are just maximizing the log-likelihood function in case of the gradient descent that is similar to that of the ridge regression.

Problem 5 : Non-linear Regression

For the non-linear regression we are finding out the effect of using the higher order polynomials as the input features. In this case, we consider only the third variable in each of training and test set. And, each of these attributes are converted into a p-attribute vector. The N X (p+1) matrix output is used to train the ridge regression weights as defined in the LearnRidgeRegression method. The p value is varied from 0 to p=6. The mean squared error for test and train data is plotted against p value for λ = 0 and optimal λ value obtained from problem 3.



Consider the plot of MSE in case of the test data. The optimal p value as observed in the graph is at p = 1. Comparing the results in case of “No Regularization” and “Regularization” when applied for test data, the overall performance is much better when there is regularization. But in both the cases, the optimal value is observed at p = 1. As the p value is increased, the error on training data decreases slightly even without regularization, but the error in case of test data increase abruptly after p = 1 as the model is over-fitted. So, when regularization is done we can see a decrease in the error, even when p value is increased.

Problem 6 : Interpreting Results

The Mean Squared Error (MSE) can be used as an error metric to compute the performance of the regression models. MSE is the mean of sum of the residuals and is used as a metric for both training and test data. It gives the fit of the model that has information on how close the test data is to the model data and how far the model is from achieving good accuracy of prediction. The MSE can be calculated for various values of λ, p and this information can be used for model optimization. In problem 3, the MSE is plotted against various lambda values to get the optimal lambda value where the MSE value is minimum. MSE is a good metric in accessing the model based on the prediction of data. The best setting or model varies based on the data but the model that gives least MSE, can be used for that data.

In problem 5, when we plotted MSE against p-values for optimal lambda and λ = 0, the least MSE is observed at p = 1 which is same as the linear ridge regression. So, for predicting the diabetes level using the input features gives best results when using linear ridge regression.