CSE 574

INTRODUCTION TO MACHINE LEARNING

PROGRAMMING ASSIGNMENT 1

CLASSIFICATION & REGRESSION

ALEKYA KUMAR

(50249052)

TRUPTI JADHAV

(50249177)

Problem 1: Experiment with Gaussian Discriminators

Linear Discriminant Analysis and Quadratic Discriminant Analysis are aimed at performing classification on data.

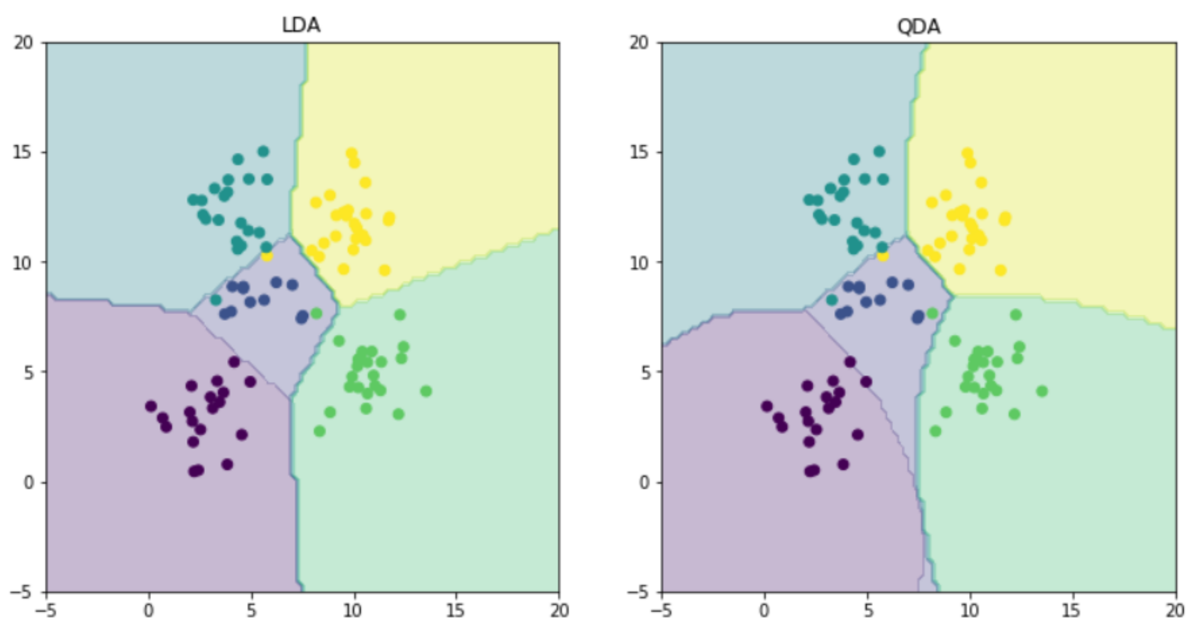
LDA has different means and the variance is the same. For calculating variance, LDA takes the entire Input variable X training data. LDA uses the Mahalanobis distance to calculate the class. It uses linear combination of variables to predict the out.

QDA also uses means and variance to perform classification. It is similar to LDA, but the variance here is not a diagonal matrix. It uses Non linear combination of variables to predict the output.

The Bernoulli parameters, covariance and means are estimated using the normal distribution formula. Once the parameters are obtained, the MLE is estimated for each class of the multivariate normal distribution. After estimating MLE, we calculate the accuracy, which will give us how well our classifier has performed.

The difference in the boundaries between LDA and QDA is because of the difference in variance between the two methods. LDA does not suffer from high variance in the data, but QDA does. As the variance is high for QDA since it uses different variance for each class, the category region also changes. This is the reason LDA gives linear boundary and QDA gives a non linear one. Also, LDA uses the property of Mahalanobis distance that serves as an efficient way of classifying an input.

Accruacy obtained for **LDA – 97%** and **QDA – 96%.**



Problem 2:

Experiment with Linear Regression

Linear regression is a basic and commonly used type of predictive analysis. Good data does not always tell the complete story. Regression analysis is commonly used in research as it establishes that a correlation exists between variables.

In simple linear regression, each observation consists of two values. One value is for the dependent variable and one value is for the independent variable.

Multiple Regression Analysis When two or more independent variables are used in regression analysis, the model is no longer a simple linear one.

The weights that maximized the likelihood of the training data were calculated and the regression model is built.

The prediction is done for both the cases: first, without intercept and second with intercept.

If we don't include the constant, the regression line is forced to go through the origin. This means that all of the predictors and the response variable must equal zero at that point. If the fitted line doesn't naturally go through the origin, the regression coefficients and predictions will be biased.

And the mean square error (MSE) is calculated.

The MSE without intercept: **106775.36.**

The MSE with intercept: **3707.84**

Thus we can notice that MSE with intercept is better as it gives lower measure of the error. The inclusion of intercept does not force the model to pass through origin thereby reducing the error.

Problem 3:

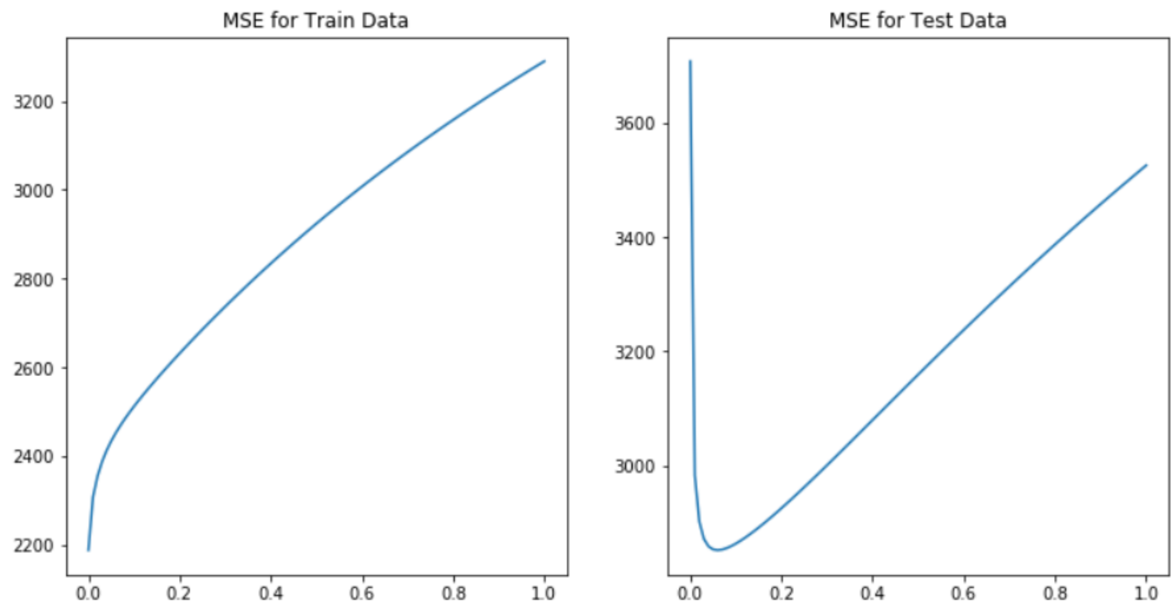
Experiment with Ridge Regression

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. Ridge regression is like least squares but shrinks the estimated coefficients towards zero. It is hoped that the net effect will be to give estimates that are more reliable.

The optimal value of lambda can be found out for the function where the error is the lowest. Here the lowest error in test data is 2851.33 for the lambda value of **0.06.** As we can see from the graph, the error graph reached the minimum at a point and then increases eventually as the value of lambda increases. The point where the error is lowest gives the best value for lambda.

From the previous problem, we can see that the error is 3707.84 for the test data. For the same function adding a regularization parameter reduced the error to a large extent, 2851.33.

Thus adding a learning or regularization parameter has benefitted the prediction.

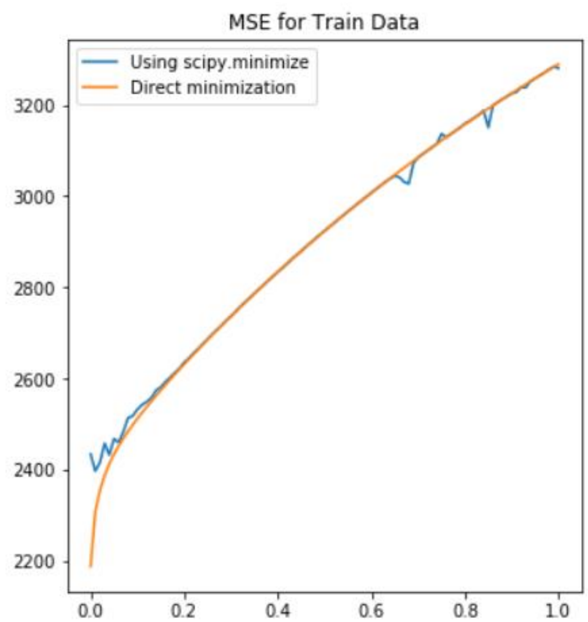


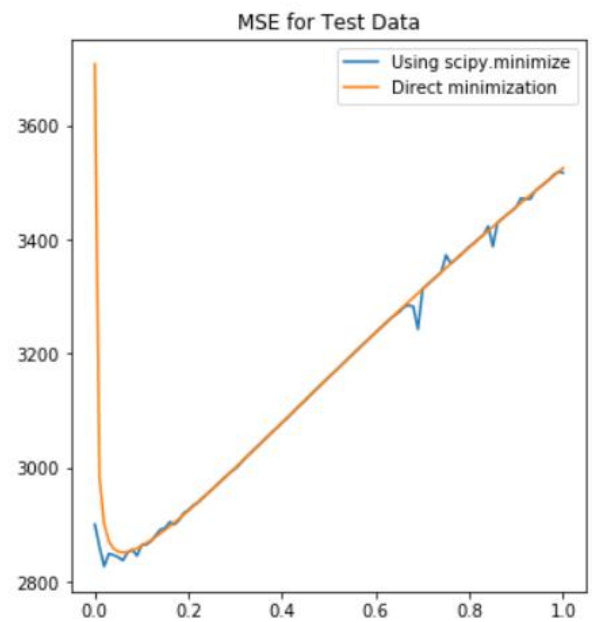
Problem 4: Using Gradient Descent for Ridge Regression Learning

Gradient Descent is used in regression to avoid calculating the inverse term. Inverse calculations are numerically unstable and if there exists a column that can be expressed as a linear combination of any other column(Rank < number of columns), then the inverse can't be calculated and the division by zero error occurs. In other words, it is safer to add a gradient term to avoid such situations.

The gradient descent was implemented to calculate the weights. The Mean Square Error for Train and Test data were calculated.

We are comparing the MSE obtained from Ridge Regression with and without using Minimize function. Different lambda values were used for each iteration and the graph was plotted. The blue line indicates optimization using the Minimize function and the orange one indicates Direct Minimization.



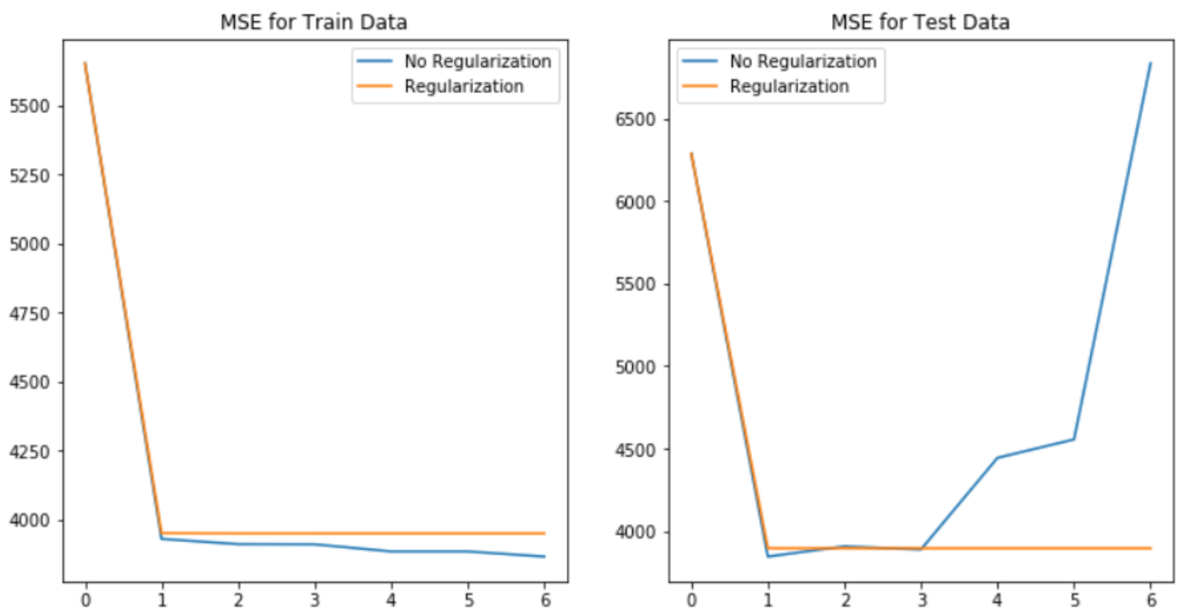


Here, the blue curve has some dips and shoots for some values of Lambda. For the dips in the curve, they mean that the function is trying to find a minima or trying to converge at a point and for others, it means there is an overshoot since the function takes baby steps of 0.1 units.

Problem 5: Non-linear Regression

The non linear regression matrix was generated using the VanderMonde principle. The matrix consists of elements of the form 1 + x^1 + x^2 + …. + x^ n. It is a polynomial non linear function.

Lambda values of 0 and the optimized lambda value obtained from Ridge Regression was passed tot the function to calculate the error for both training and test data. The exponent value was varied from 0 to 6. The test and train error were plotted for both with and without regularization parameter.



From the graph we observe that the function with a regularization parameter performs better than the one without regularization.

Regularization parameter : For the initial exponent of value 0, the error was high. For the first power of x, the error dropped drastically and then the trend continued. The optimal value of p would be 1 here.

No Regularization Parameter : Though the error decreased for the first power of x, it started increasing from the second power and so on. Here, the optimal value of p would be 1.

Problem 6: Interpreting Results

The previous 4 problems helped in predicting the diabetes level by using 4 methods of prediction.

We can compare the methods based on the Test error values.

Ordinary Linear Regression	**3707.84**	
Ridge Regression	**2851.33**	→ Regularization parameter = 0.06
Ridge Regression - Gradient Descent	**2826.95**	→ Regularization Parameter = 0.02
Non Linear Regression	**3845.03**	→ Regularization parameter = 0.02
	6286.40	→ Regularization parameter = 0

From the values, we can find that Ridge regression coupled with Gradient Descent and with regularization parameter value as 0.02, has the lowest error. Thus, we can conclude that Ridge Regression with Gradient Descent can be recommended for predicting diabetes level.