|  |
| --- |
| University at Buffalo |
| CSE574 - Introduction to Machine Learning |
| Programming Assignment 2 - Report |
|  |
| Aayush Shah (50207564)  Siddharth Shah (50205787)  Haril Satra (50208283) |
| 4/12/2017 |

|  |
| --- |
|  |

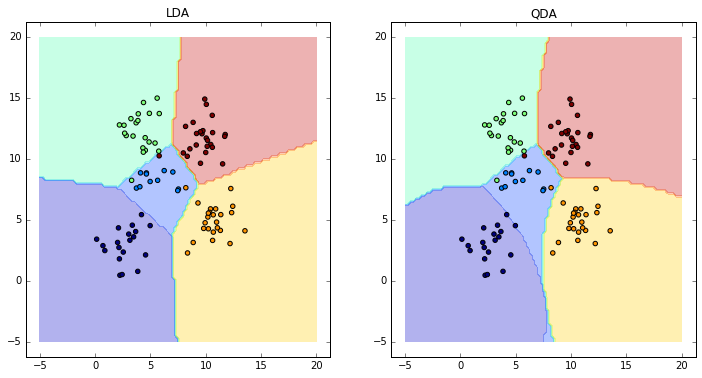
Experiment with Gaussian Discriminators

# Accuracy

LDA: 97 %

QDA: 96 %

# Plot



# Analysis

In LDA (Linear Discriminant Analysis) we assume that the covariance matrix for all the different output classes is the same. This results in a linear classifier. On the other hand, in QDA (Quadratic Discriminant Analysis) the covariance matrix for each output class is different.

This quadratic discriminant function in QDA is similar to the linear discriminant function except that the covariance matrix is different for different classes. Hence the quadratic terms cannot be ignored. Thus the discriminant function in QDA results in a quadratic function containing second order terms. Hence there is a difference between the boundaries of the two classifiers. LDA has linear boundaries while QDA has non-linear boundaries due to the second order terms in the discriminant function.

Experiment with Linear Regression

# MSE (Mean Square Error)

**Training Data:**

* Without Intercept: 19099.44684457
* With Intercept: 2187.16029493

**Testing Data:**

* Without Intercept: 106775.36155355
* With Intercept: 3707.84018134

# Analysis

If we do not add the intercept, we impose a restriction on the regression line to pass through the origin leading to the fit getting heavily pulled down (Systematically shifted towards larger or smaller values).

Thus usually linear regression with intercept will be a better fit and provide a smaller MSE as compared to linear regression without intercept. This can also be observed in the above results that are obtained.

The only reason to use Linear regression without using a intercept is when it is known that the process has a model with zero intercept.

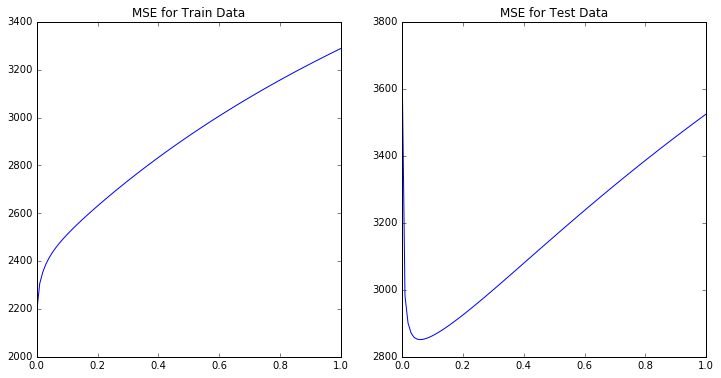
Experiment with Ridge Regression

# MSE (Mean Square Error) for Ridge Regression for optimal lambda (0.06).

**Training Data (With intercept):** 2451.52849064

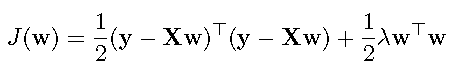
**Testing Data (With Intercept):** 2851.33021344

# Plot



# Comparison of relative magnitudes of weights learnt using Linear Regression and Ridge Regression

In the case of linear regression for weights learnt, we get extreme values (large positive or large negative values). Whereas in case of ridge regression or l2 norm regression the values are regularised and have lesser variance.



Considering the above equation, we can see that if weights are very large then the sum of square error term will minimize but the penalty will increase. However, if weights are less, then the penalty term will minimize but square error will increase.

We can validate this with our results as the extreme weights in linear regression are regularized.

# Comparison of Linear Regression and Ridge Regression in terms of errors on train and test data.

|  |  |  |
| --- | --- | --- |
|  | **Train Data** | **Test Data** |
| **Linear Regression** | 2187.16029493 | 3707.84018134 |
| **Ridge Regression** | 2451.52849064 | 2851.33021344 |

Linear Regression often faces the issue of overfitting on the training data. But this fit might not necessarily be good for the test data.

This can be overcome by introducing a penalty term. Ridge regression introduces this penalty term as l2 norm of the weights, leading to generalization of the fit or in other terms reduces the complexity of the fit. This may increase the error on the training data but will usually perform better on the testing data as compared to Linear Regression.

# Parameters

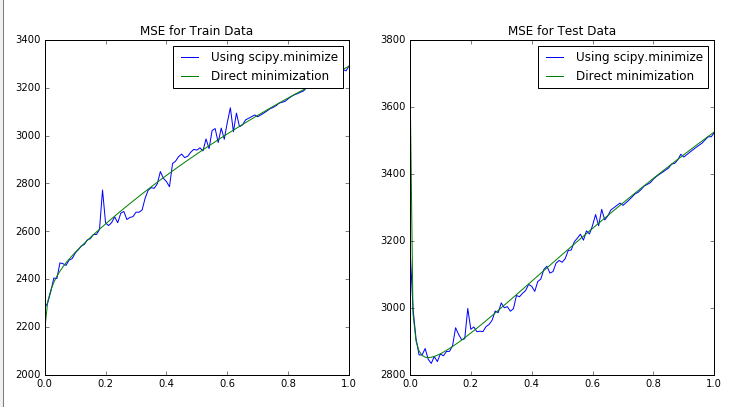
**Lambda:** The optimal value of lambda which we obtained is 0.06.

With growing lambda, the training error increases as the residual sum of squares become larger. However, in case of test data the training error reaches a minimum for a specific value of lambda after which it again increases. The optimal value of lambda suggests that it is a good generalization fit.

There is no fixed value of lambda which provides the minimum error for any data. It varies from data to data and the best way to obtain is to vary it in small steps starting from 0 (no regularization) up to some higher value.

Using Gradient Descent for Ridge Regression Learning

# Plot



What gradient descent basically does is that it starts from some point and moves downhill to find the point with the lowest error. The output of gradient descent will eventually (if enough iterations are provided) be the same as that of methods used in problem 2 and problem 3.