CSE 574: Programming Assignment 2
Classifcation and Regression

Group No. 51
Ajay Kumar Davuluri, UBIT ID: adavulur, Person No: 50168851
Lalith Vikram Natarajan, UBIT ID: lalithvi, Person No: 50169243
Srinath Goud Vanga, UBIT ID:srinathg, Person No:50169176

April 5, 2016

1 Objective

Implement Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), Linear Regression, Ridge Regression, Ridge Regression using Gradient Descent, Non Linear Regression and interpret the results.

2 Datasets

- A 2D sample data set in sample.pickle file with 150×2, 150×1 dimensions.
- Diabetes data set in diabetes.pickle file with 242×64, 242×1 dimensions.

3 Observations

3.1 LDA vs QDA

- Accuracy of LDA = 97%
- Accuracy of QDA = 96%

From the plots shown (1 & 2), it can be seen that the LDA and the QDA differ with the QDA having curves for bounds. This is because of the inherent difference in the way LDA and QDA are generated. In the LDA, we use a pooled mean whereas in the QDA, we separate this mean into two, thus generating a quadratic equation.

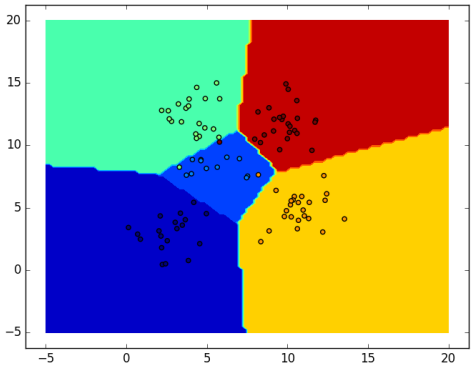

Figure 1: LDA Plot

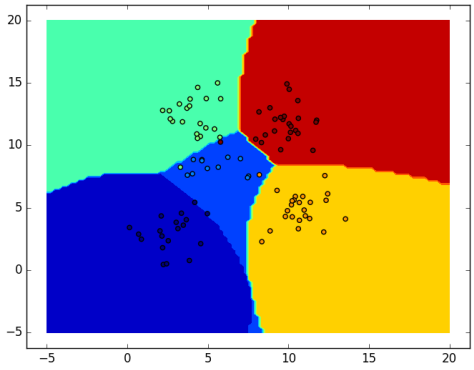

Figure 2: QDA Plot

3.2 RMSE Comparison

- RMSE without intercept 326.76499423180127
- RMSE with intercept 60.892037089024484

The RMSE value becomes very less when an intercept is used and is thus a preferred method of computation.

3.3 Ridge Regression

The plots 3 & 4 show us that RMSE is much lower for the train data and more for the test data. The optimal value of λ is found to be 0.055

3.4 Ridge Regression using Gradient Descent

From the plot 5, we can find that the RMSE obtained from this method has not changed much from the computation in problem 3.3. With the gradient descent methodology, an iterative approach of finding the minimum is carried out.

3.5 Non Linear Regression

The plot 6 show the data tested with λ values equal to 0 and equal to 0.055 (the optimal one obtained from the problem 3). Using p values from 0 to 6 better results than that of linear regression were found. It was also found that p = 3 is giving best results, although it doesn't change much after p = 3 and fairly stabilized with RMSE.

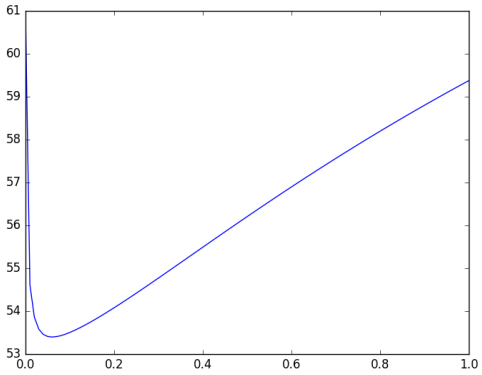

Figure 3: RR Plot

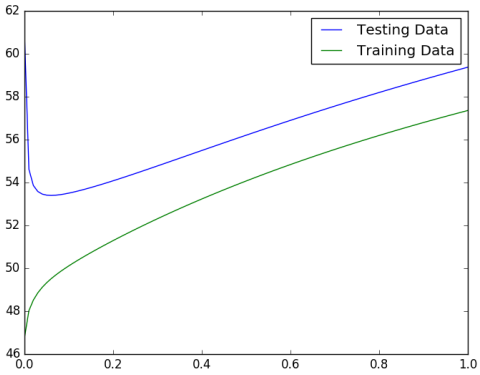

Figure 4: RR Training vs Test plot

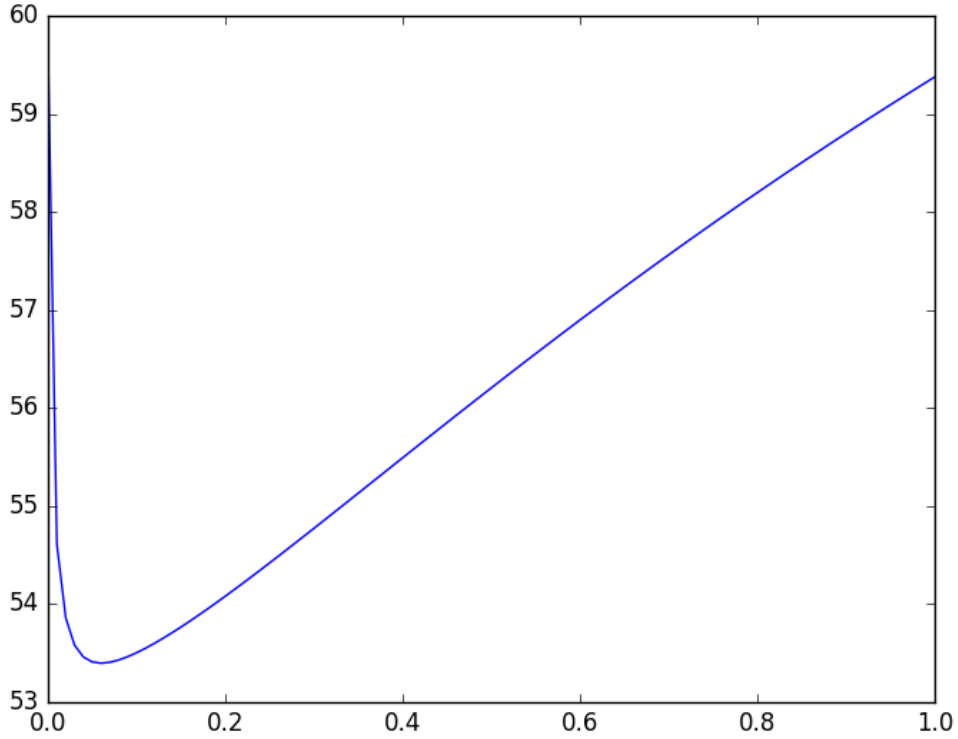

Figure 5: RR plot using GD

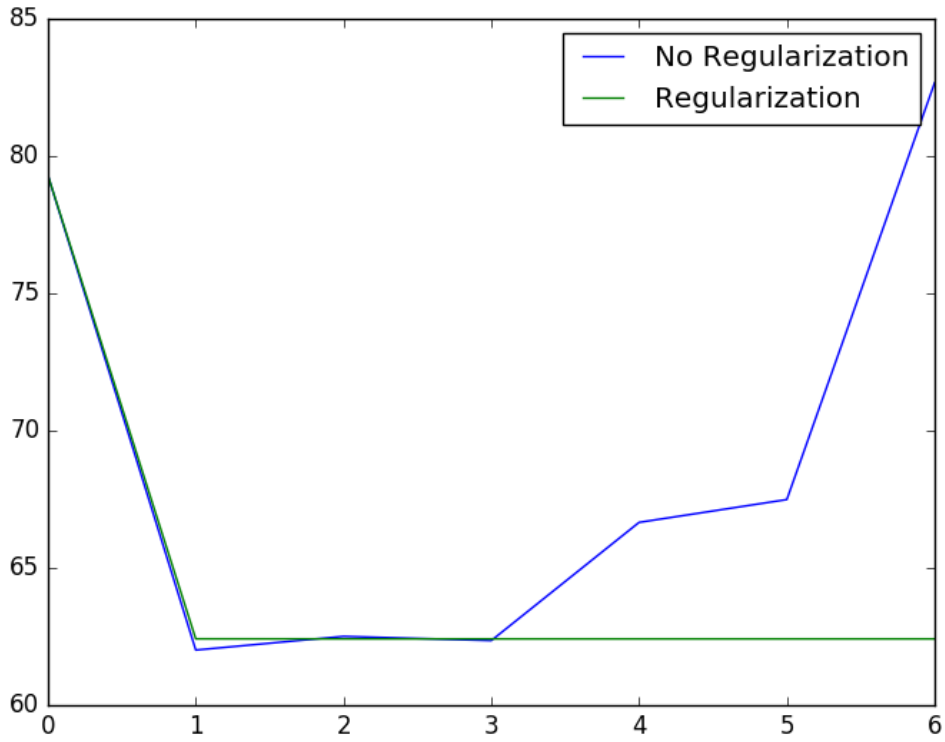

Figure 6: NLR plot

3.6 Interpreting Results

The RMSE and the lambda are in fact good regression measures. RMSE tends to penalize high values and better fit with lower values. λ is used to figure out the optimal regularization value that penalizes overfitting. When used without intercepts, RMSE is about 326.76 whereas with an intercept it reduced to 60.89, so having an intercept term reduces the error. As regards the λ value, it was found that the RMSE value oscillated between 53 and 60 for various values of λ with experiments conducted on both ridge regression and gradient descent. it is identified that λ value of 0.055 is optimal. Both Ridge regression and Gradient descent methods are giving similar results as mentioned earlier, but it may be too simplistic an approach so adding non-linear features can help the cause. Also it is noted that RMSE actually has lower values when non linear features are used.