CSE 574 Introduction to Machine Learning

Programming Assignment 2

Classification and Regression

Group 11:

Ahut Gupta (50169273)

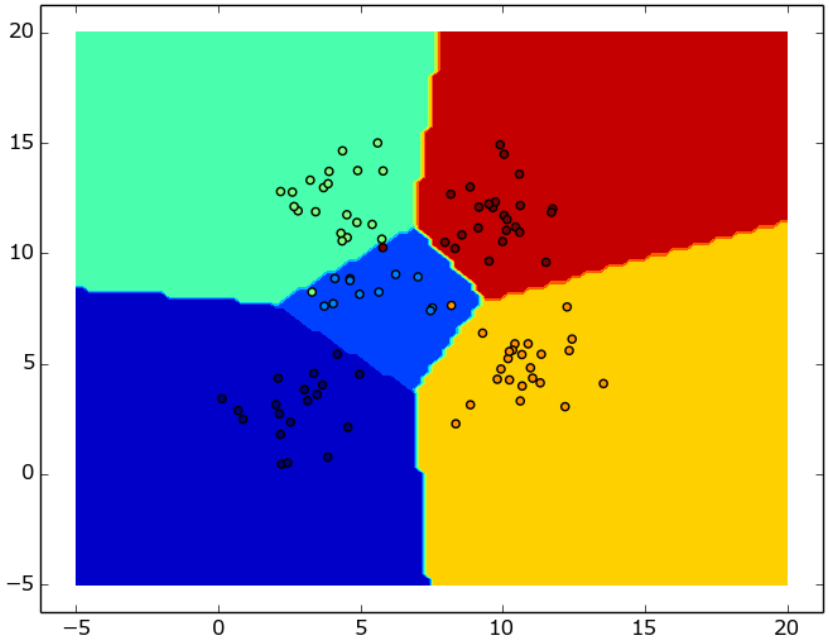
Rachna Shivangi (50169516)

Vandana Chokkam (50167945)

Problem 1 – Experiment with Gaussian Discriminators

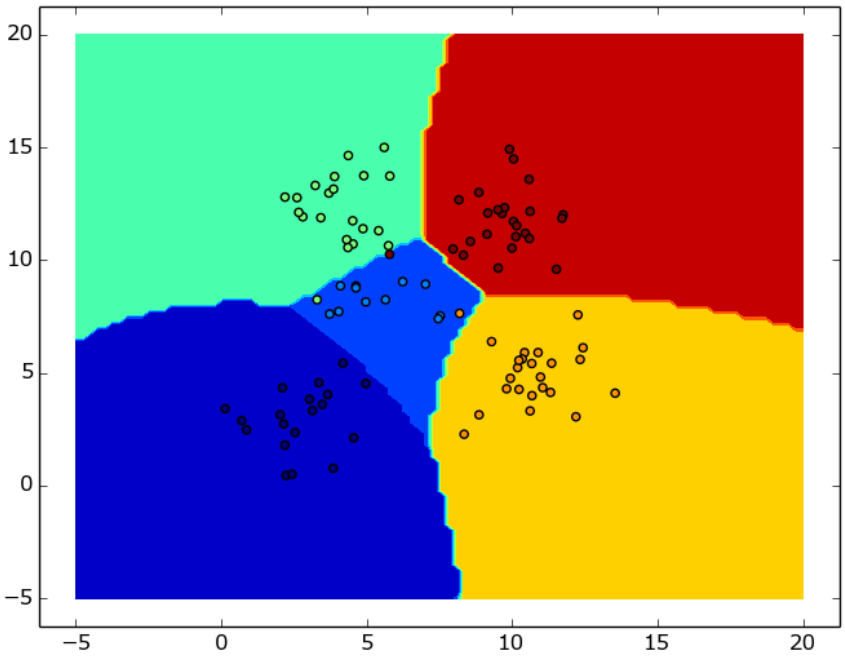
Linear Discriminant Analysis (LDA)

For LDA, we achieved an accuracy of 97% on the test data set. The plot for LDA classification is as shown below:



Quadratic Discriminant Analysis (QDA)

For QDA, we achieved an accuracy of 97% on the test data set. The plot for QDA classification is as shown below:



The plotted boundaries in LDA and QDA differ because in LDA, the covariance is independent of the k number of classes and is the same for each class. Therefore, the classifier follows a linear behavior.

Whereas in QDA, every k class has a different covariance matrix. Thus, the shape of the decision boundary for QDA is determined by a quadratic function.

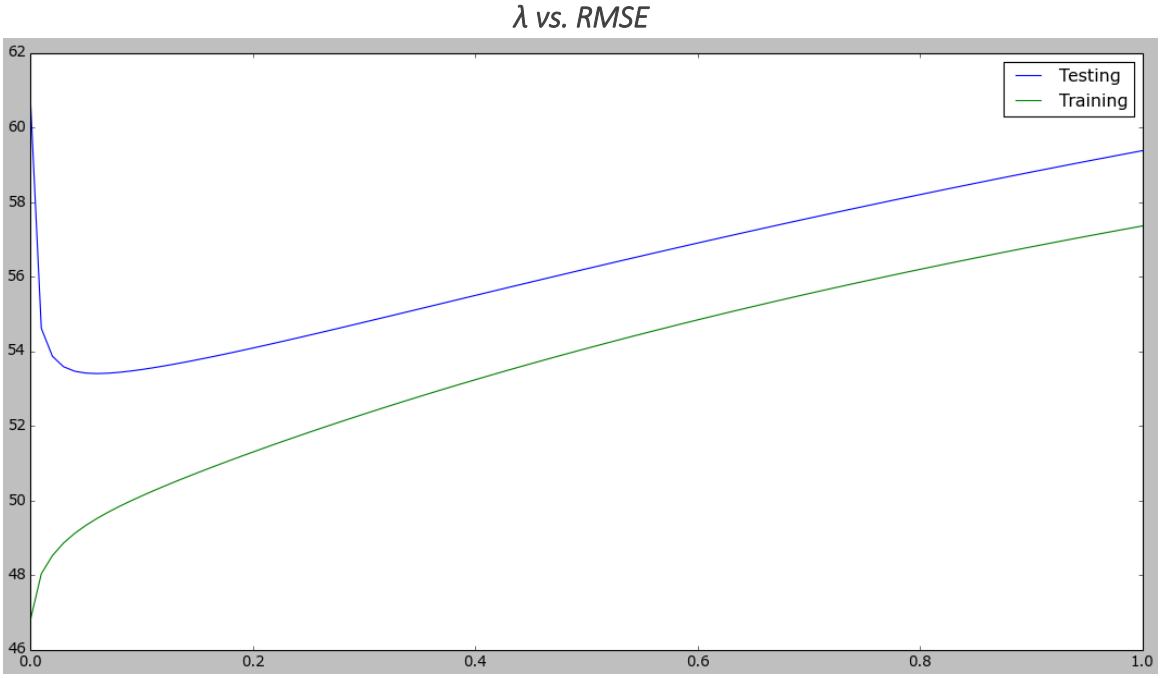
So, in LDA the decision boundaries are lines, but in QDA the boundaries are curves as can be seen from the graph.

Problem 2 – Experiment with Linear Regression

Data Set	**Intercept**	**RMSE**
Training	Without	138.20074835
Training	With	46.7670855937
Test	Without	326.764994365
Test	With	60.8920370955

The RMSE value is lower in the case when an intercept is used for both training as well as testing data, so using intercept is better. Without intercept, the hypothesis line must pass through origin and is restricted to rotation only. But, when an intercept is used the hypothesis line can rotate as well translate. Thus, the hypothesis learnt is much closer to the true concept, hence the error is lower when we use an intercept as can be seen from the table.

Problem 3 – Experiment with Ridge Regression



	OLE	**Ridge**
Min Weight	-86639.45	-111.67
Max Weight	75914.47	203.31

As we can see from the table, the magnitude of weights obtained using ridge regression is much lesser than those from OLE. Ridge regression uses a regularization parameter (λ) which dampens the size of regression coefficients (weights) unlike OLE in which every point in training data affects the weights of the solution curve. Hence, we get very large magnitude of weights for OLE.

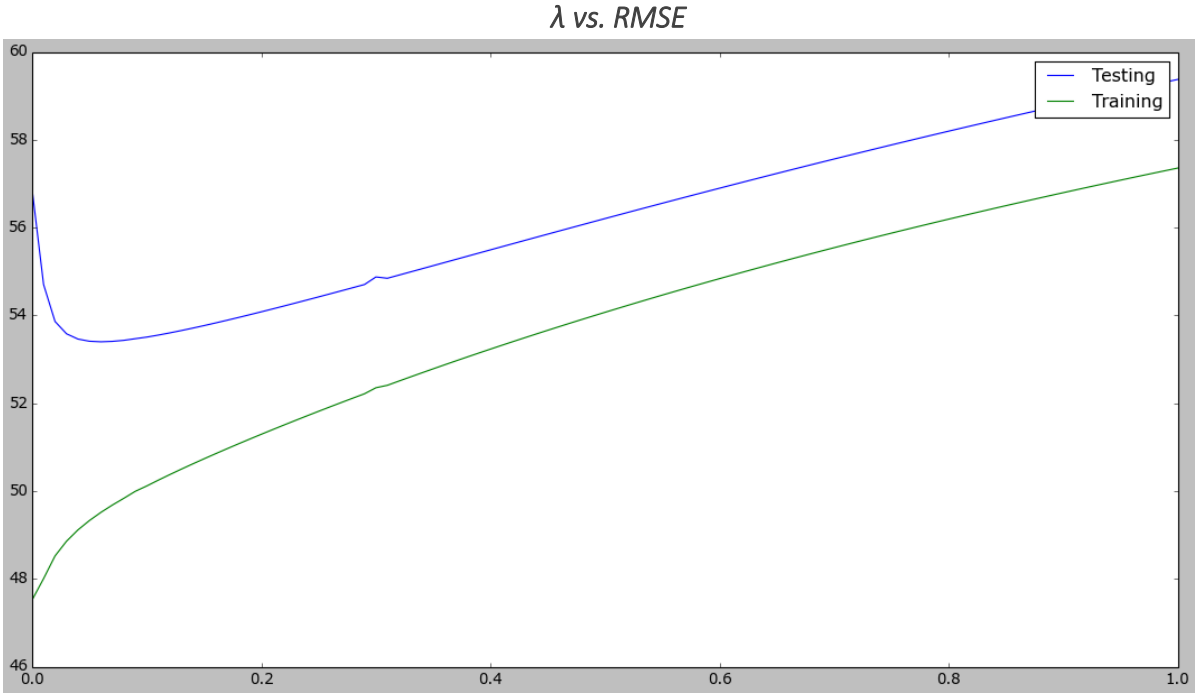
RMSE for OLE with intercept is equal to ridge regression without regularization for both training and test data. RMSE for Ridge regression can however be further improved by varying the λ.

For training data, the regression line is over-fitted when there is no regularization. Hence, we get the lowest error for λ=0. As we increase the λ, the regression line starts to under-fit on the data set, giving higher error values.

For test data, we get high error when λ = 0, as the line is over-fitted to the training data. But as we increase the λ up to 0.06, the weights are regularized and they fit the test data better, giving lower errors. On increasing the λ further, the line starts to under-fit on the data set and we observe higher values of error.

Hence an optimum value of λ for testing data would be 0.06 and that for the training data would be 0, as it gives minimum error for the corresponding data set.

Problem 4 – Using Gradient Descent for Ridge Regression Learning



The parameter estimation by using analytic equations in Problem 3 and the gradient descent procedure gives almost the same values for weights and RMSE for the range of λ values.

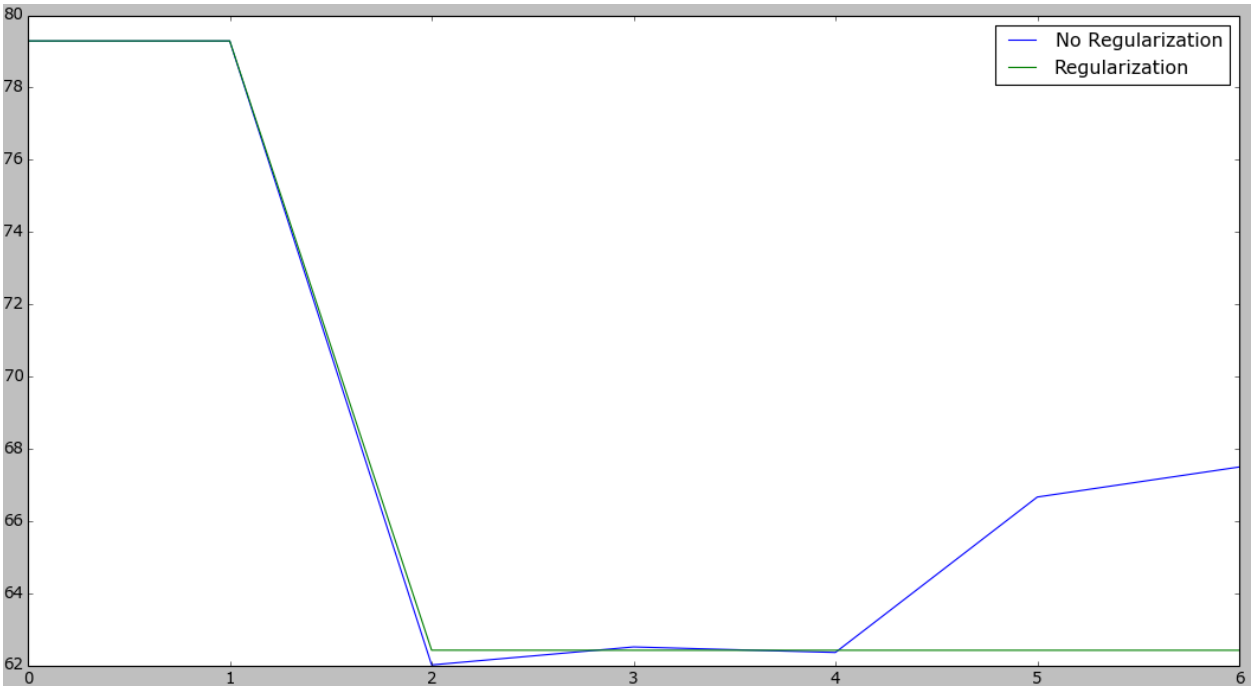
The optimum λ obtained for this problem is the same as that from Problem 3.

As the number of training examples and attributes is small, solving the equations on the matrices is fast. Had the size of the matrices been larger gradient descent method would have been faster and better conditioned.

Problem 5 – Non-linear Regression

Testing Data Set

RMSE vs. P for Testing Data Set



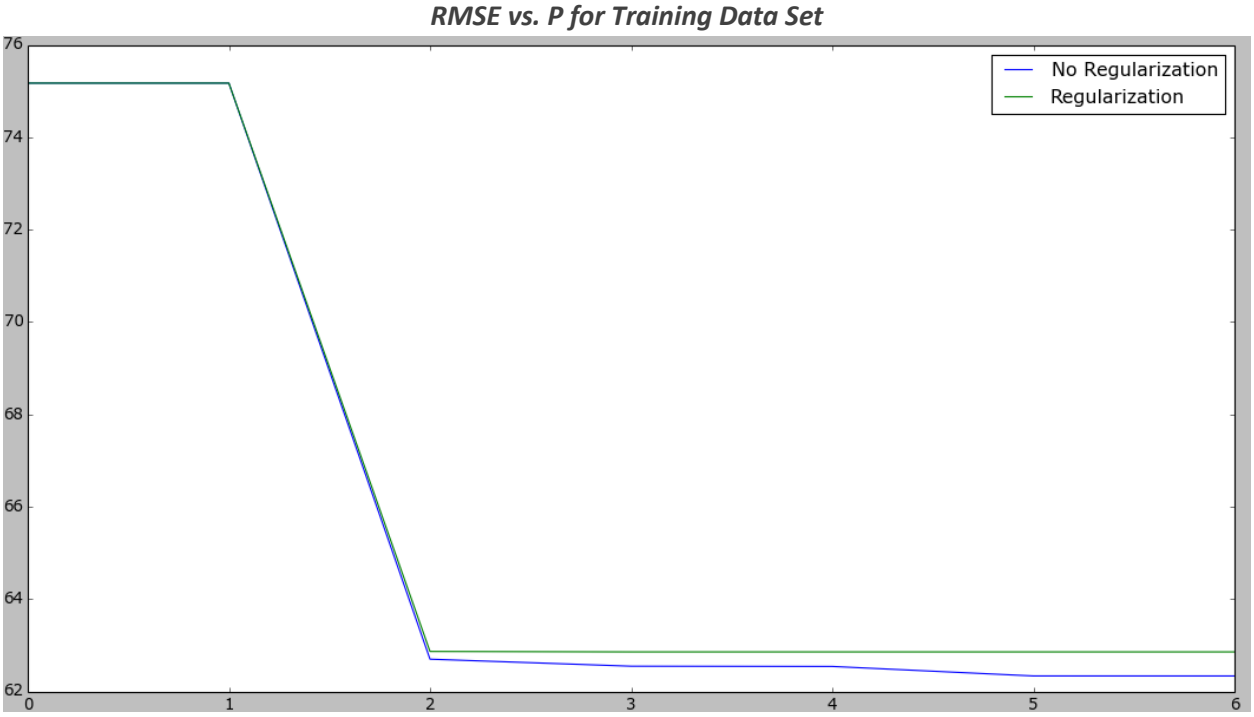
Without regularization (λ=0), we see that the error decreases upto p=2, remains constant upto p=4 and then increases beyond that.

For p=2 the regression line is a quadratic curve. Because of its shape, it can follow the data better than a line in case of p=1. Hence we get lower error for p=2. As we increase p, the regression equation attains higher degree terms and becomes curvier. We observe that the error starts to increase. This is because our curve mimics the training data very closely and overfits to that dataset. Thus performing poorly on the testing data.

With regularization (λ = 0.06), we observe that error decreases till p=2, but it remains constant after that. The reason for lowering of error from p=1 to 2, is the same as in the previous case. But unlike the previous case, we don't observe a rise in error beyond p=2. This is because we use regularization, thus avoiding the overfitting problem.

Thus, we get the lowest error for p=2, irrespective of regularization.

Training Data Set



Without regularization, we see that the error decreases as we increase p.

This is because as we increase p, we get higher order terms in our regression line, and it can bend and conform very closely to the training data. Since we are testing it against the training data, we see lower error each time we increase p.

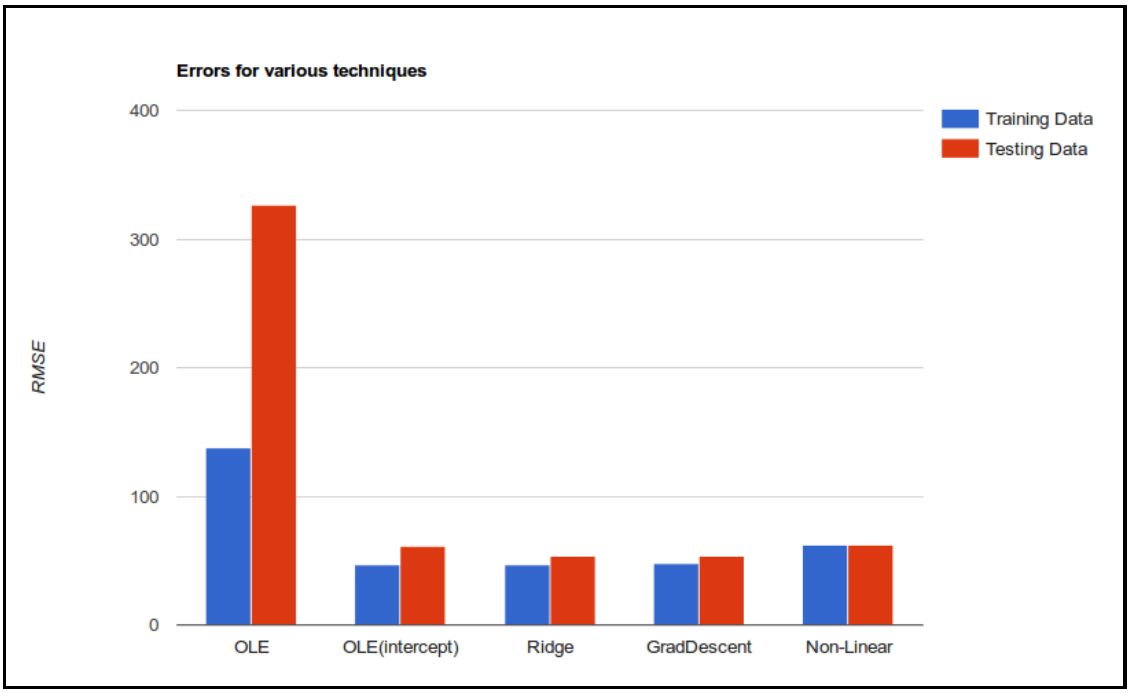
With regularization, the same graph as that of testing data is observed. The error decreases upto p=2, beyond which it remains constant.

Problem 6 – Interpreting Results

We use RMSE to compare the various techniques of Linear Regression used here. As the name suggests, Root Mean Square Error is the squared error of the various input points from our model. Lower values suggest that the points are very close to the model and thus a better fit. Whereas a higher value suggests that the points are far apart from our model. Thus a technique which gives the lowest RMSE is preferred.

The following graph presents the best RMSE values from the different techniques used in problems 2 to 5.

RMSE vs. Error for Various Techniques



From the graph, we can see that lowest errors were obtained from Ridge regression. Two methods were used to compute the weights:

- Matrix arithmetic and inverse formula
- Gradient Descent

Since both these methods give the same error on testing data, we would recommend Ridge Regression to anyone who uses Linear Regression on the diabetes data.

Also, for this case, the number of dimensions is low. Hence, computing the matrix inverse is much faster than using gradient descent. However, if a data set with a very large number of dimensions is used, we would recommend using gradient descent based ridge regression as computing the inverse of the input matrix will be very time consuming and might not be very stable.