CSE 574 - Introduction to Machine Learning

Classification and Regression

Group Number: 48

Kirti Hari - 50208065

Shruti Kulkarni - 50207124

Shashank Suresh - 50208025

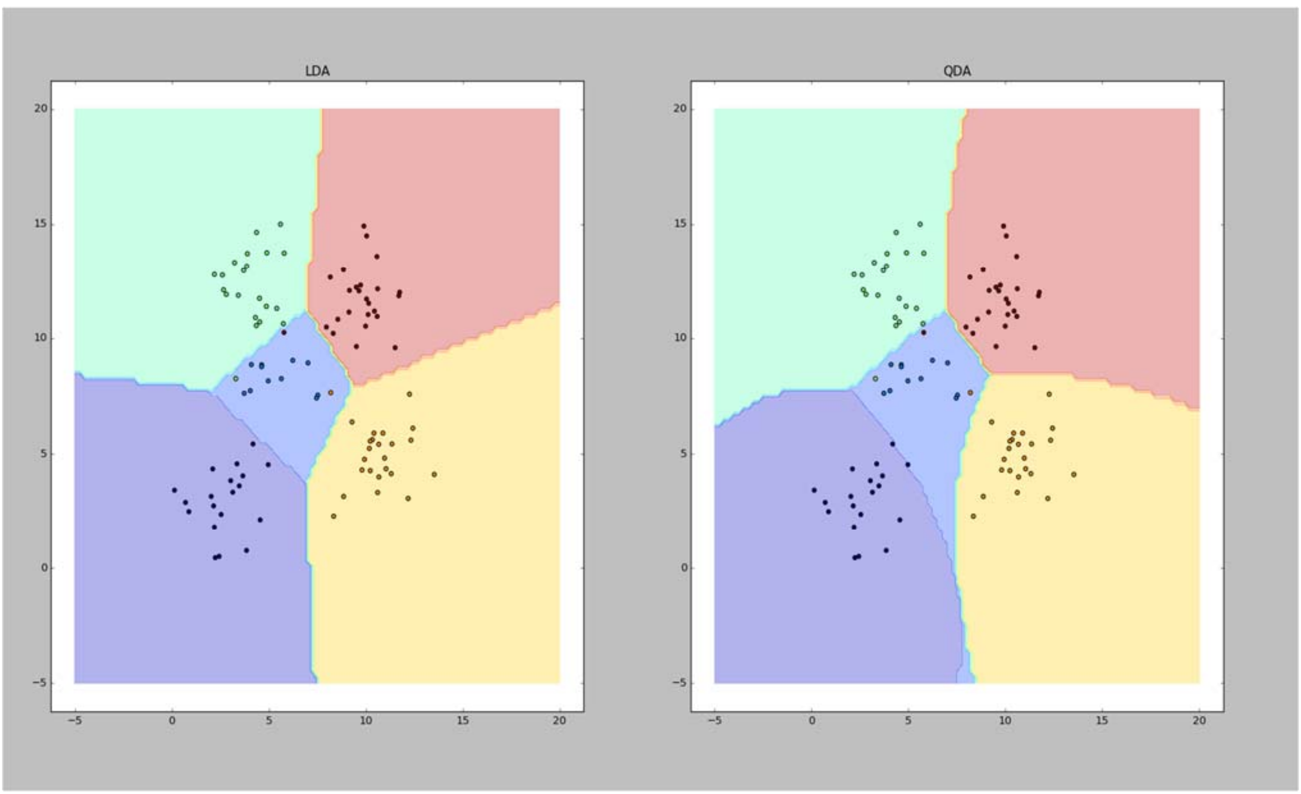
PROBLEM 1 REPORT: Gaussian Discriminators

Problem Statement: Train both methods using the sample training data (sample train). Report the accuracy of LDA and QDA on the provided test data set (sample test). Also, plot the discriminating boundary for linear and quadratic discriminators. The code to plot the boundaries is already provided in the base code. Explain why there is a difference in the two boundaries.

Observations: The accuracies for LDA and QDA observed are as follows:

LDA Accuracy --- 97%

QDA Accuracy --- 96%



LDA tends to give better accuracy than QDA on this dataset. Though QDA tends to fit the data better than LDA, the number of parameters increases significantly with QDA. It requires more training data to get reliable values of mean and covariance. LDA relaxes this constrain and gives better accuracy for a smaller data set.

The difference in boundary of LDA and QDA is seen because of the way the covariance matrix is computed. LDA produces a linear boundary because the same covariance matrix is used for the entire data set. In QDA, however, the decision boundaries are quadratic in x as for each output class k a different covariance matrix is computed.

PROBLEM 2 REPORT: Linear Regression

Problem Statement: Calculate and report the MSE for training and test data for two cases: first, without using an intercept (or bias) term, and second with using an intercept. Which one is better?

Observations: For both the data sets, the recorded MSE error values with and without intercept are:

For Test data:

MSE without intercept for Test Data = 106775.36155644

MSE with intercept for Test Data = 3707.84018157

For Train data:

MSE without intercept for Train Data = 19099.44684457

MSE with intercept for Train Data = 2187.16029493

From the above values, we know that MSE error with intercept is lower for both test and training data. Hence, we can say that the **weights are learnt better with the intercept** for both the data sets. In general, the regression line is forced to go through the origin without an intercept term. If the fitted line does not go naturally through the origin, the weights learnt may be biased when we do not include the intercept.

PROBLEM 3 REPORT: Ridge Regression

Problem Statement: Calculate and report the MSE for training and test data using ridge regression parameters using the "testOLERegression" function that you implemented in Problem 2. Use data with intercept. Plot the errors on train and test data for different values of λ. Vary λ from 0 (no regularization) to 1 in steps of 0.01. Compare the relative magnitudes of weights learnt using OLE (Problem 2) and weights learnt using ridge regression. Compare the two approaches in terms of errors on train and test data. What is the optimal value for λ and why?

Observations:

MSE values for Train Data with λ = 0: 2187.16029493

MSE values for Train Data with λ = 0.06: 2851.33021344

The ridge regression is used to overcome some of the problems faced by linear regression. A regularization factor λ is used to solve this problem.

In Ridge regression, the weight vector is learnt using the below formula:

$$\widehat{\mathbf{w}}_{MAP} = (\lambda \mathbf{I}_D + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

The error function with the regularization factor is given below:

$$J(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \frac{1}{2} \lambda ||\mathbf{w}||_2^2$$

Once the weight vector is learnt, we use the same "testOLERegression" for the prediction. To check the difference between the weights learnt by Linear Regression and Ridge Regression, we have calculated the average of the weight coefficients learnt and we have observed that the weights learnt by the Ridge regression are smaller in magnitude when compared to that of the weights learnt by linear regression for the same data.

For Test Data, on checking the MSE values obtained by Linear Regression and Ridge Regression, we have observed that the MSE values are low in case of the Ridge regression. Hence, we can conclude that the Ridge Regression approach is better in terms of the error. For Train data, it is the opposite.

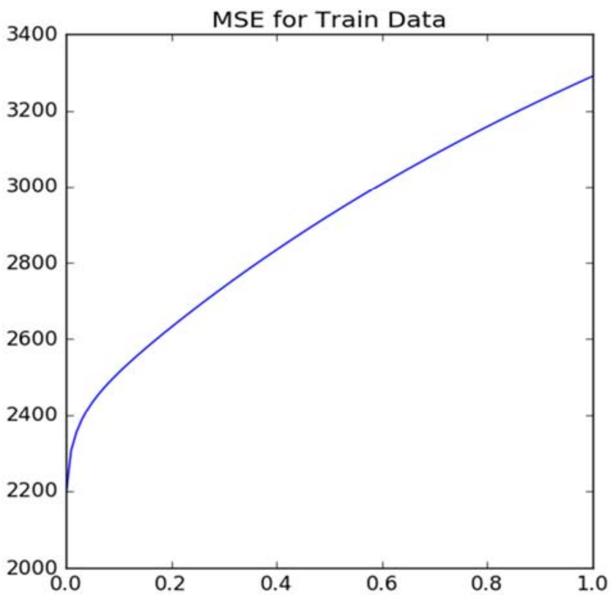
We have plotted the graphs for different values of λ against the obtained MSE values to see how the λ value affects the learning of weight vectors in Ridge Regression.

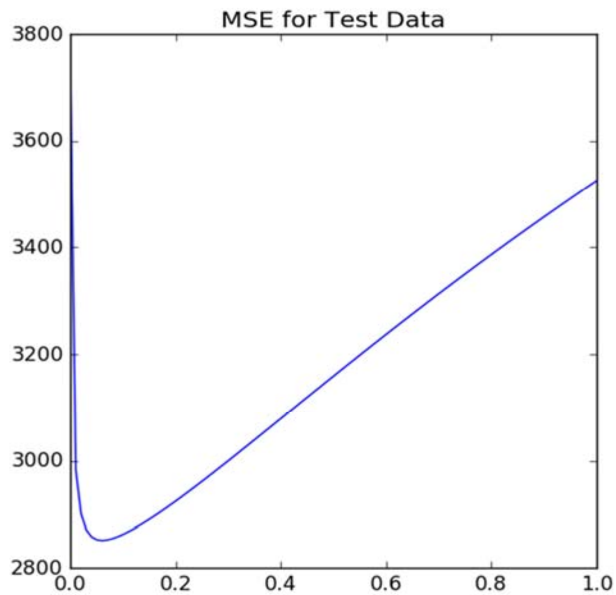
The graph for Train Data shows an increase in MSE in training data as λ increase. The higher value of λ will force the coefficients of weights to be minimum resulting in more MSE.

The graph for Test Data shows that MSE in testing data reduces with increase in the λ, initially. After a certain value the MSE starts to increase again. This is because, the higher penalty by λ will force the model to ignore important elements of the data.

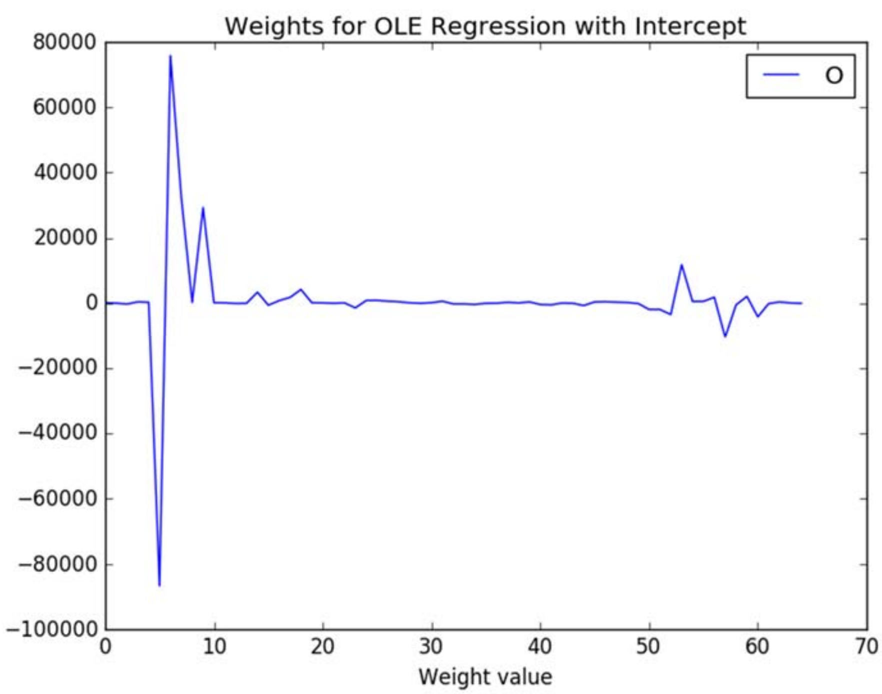
The value of λ with lowest train/test error is its optimal value.

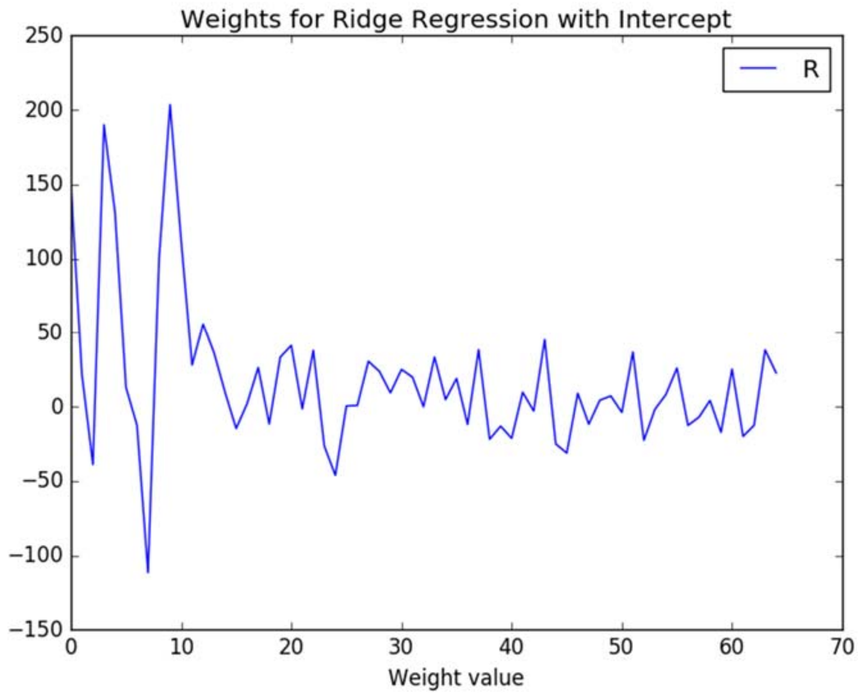
Training data: λ = 0.00
Test Data: λ = 0.06





Comparison of relative weights learnt using OLE and Ridge Regression:



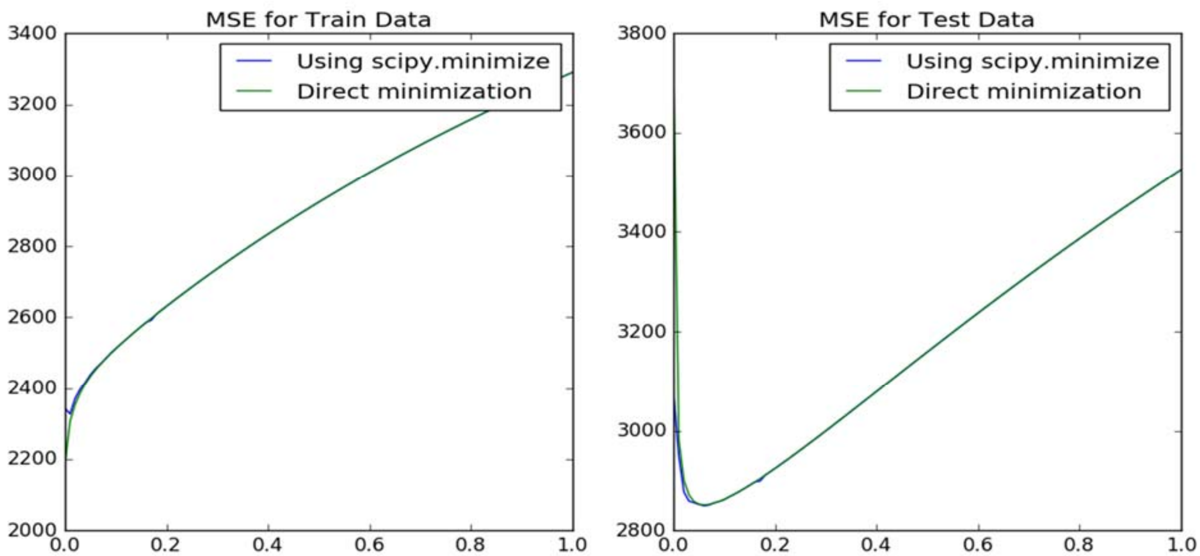


Weights for Ridge Regression with Intercept
R
250
200
150
100
50
0
−50
−100
−150
0
10
20
30
40
50
60
70
Weight value

PROBLEM 4 REPORT: Gradient Decent for Ridge Regression

Problem Statement: Plot the errors on train and test data obtained by using the gradient descent based learning by varying the regularization parameter λ. Compare with the results obtained in Problem 3.

Observations: Gradient descent can be used in learning weights of the Ridge regression instead of minimizing regularized squared error loss. This reduces the computation involved in learning weights. This also bypasses the computation of calculating the inverse of the matrix, which many times is not feasible.



This graph is similar to the graph in problem 3. There is a slight increase in MSE value. This can be seen as a trade-off between the computation and accuracy. The optimal lambda value calculated from either Ridge Regression method or Gradient Descent algorithm is approximately same.

PROBLEM 5 REPORT: **Non-Linear Regression**

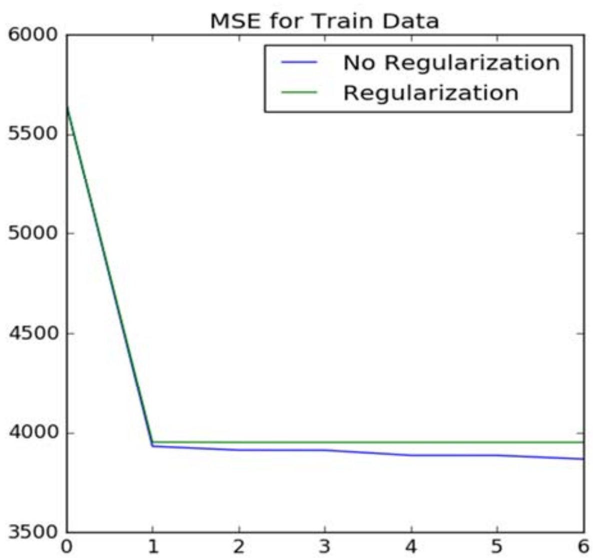
Problem Statement: Using the λ = 0 and the optimal value of λ found in Problem 3, train ridge regression weights using the non-linear mapping of the data. Vary 'p' from 0 to 6. Note that p = 0 means using a horizontal line as the regression line, p = 1 is the same as linear ridge regression. Compute the errors on train and test data. Compare the results for both values of λ. What is the optimal value of p in terms of test error in each setting? Plot the curve for the optimal value of p for both values of λ and compare.

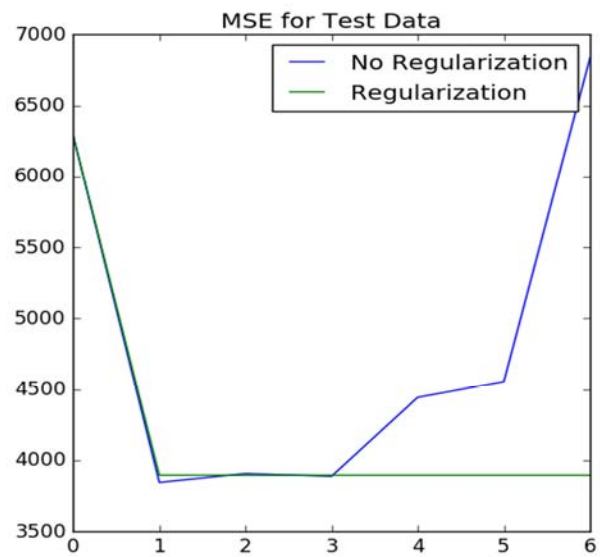
Observations:

If λ = 0, it implies that regularization term is zero and hence no regularization. In problem 3, we have calculated the optimal lambda values for both training and test data which are

Training data: λ = 0.00

Test Data: λ = 0.06





Observing the results for Test data, we can see that the MSE error with regularization is lower as the value of 'p' increases and then it becomes stable. And for training data we do not see any improvement with Regularization.

For a model without a regularization, the MSE value decreases up to p = 3. After that increase in p value causes the MSE value to increase. This is because the model tends to overfit on the training data resulting in higher error in testing data. The flexibility of the model will increase with increase in p, causing the training error to reduce, but test error will increase. In this model, we would prefer p = 1 as this results in lowest MSE value.

For a model with regularization, the MSE error is lower as the value of 'p' increases and then it becomes stable. And for training data we do not see any improvement as well. The MSE almost stays stable with increase in p. Even in this model, we prefer p = 1.

PROBLEM 6 REPORT:

Problem Statement: Compare the various approaches in terms of training and testing error. What metric should be used to choose the best setting?

Observation:

In problem 1, we have used is the Linear Discriminant Analysis and Quadratic Discriminant Analysis. By looking at the accuracies we come to a conclusion that Linear Discriminant Analysis is a better approach when we are doing a Discriminant Analysis of the data. However, this is due to the small size of the test data provided. On having a large test data, Quadratic Discriminant Analysis is a better approach.

The other types of approaches that we used were:

1. Linear Regression

2. Ridge Regression

3. Non-Linear Regression

Comparing the error values obtained on the training and test data, we concluded that Linear regression provided better results on the Training data. Whereas, Ridge regression provided better results on the Test data.

There are lot of problems we may face when we use the linear regression in a real-time scenario, in terms of Impact of Outlier, under fitting problem, unstable when the data has Correlated Attributes, etc.

So, we may tend to use the Ridge regression to overcome the under fitting problem. Also, the λ value should also be selected accordingly, as it impacts the output precision.