CSE 574 Introduction to Machine Learning

Programming Assignment 2

TEAM 18

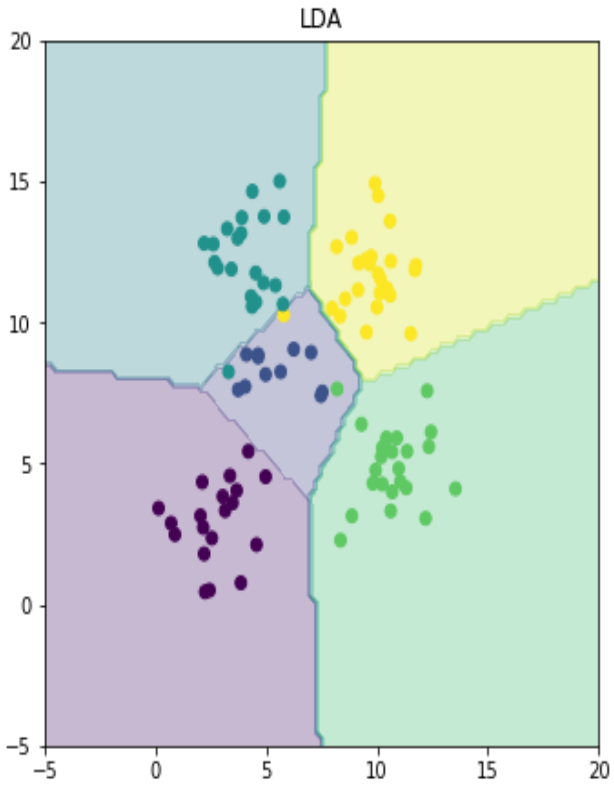
Team Members:

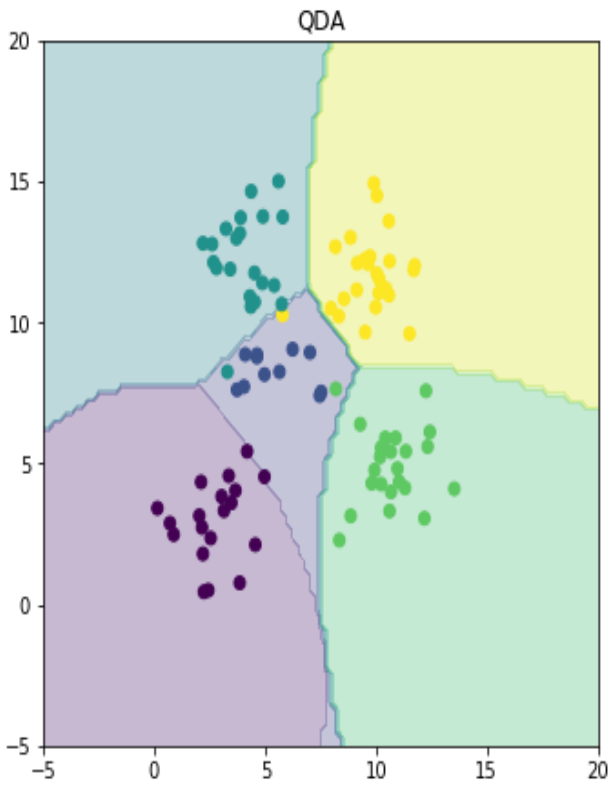
Name	Email	UB ID
Akash Yeleswarapu	akashyel@buffalo.edu	50207826
Maheedhara Achalla	maheedha@buffalo.edu	50207395
Phani Ram Sayapaneni	phaniram@buffalo.edu	50166685

PROBLEM 1: EXPERIMENT WITH GAUSSIAN DISCRIMINATORS

QDA Accuracy = 96.0

LDA Accuracy = 97.0





Conclusion:

Boundary separation in the LDA is mostly linear as we have used covariance matrix of the entire data set in computation of the LDA while in the QDA, boundaries are bit curved as we have used different covariance matrices for different classes in the computation of QDA.

PROBLEM 2: EXPERIMENT WITH LINEAR REGRESSION

S.No	Input	With Intercept	Without Intercept
1	Training Data	2187.160	19099.449
2	Test Data	3707.840	106775.362

Conclusion:

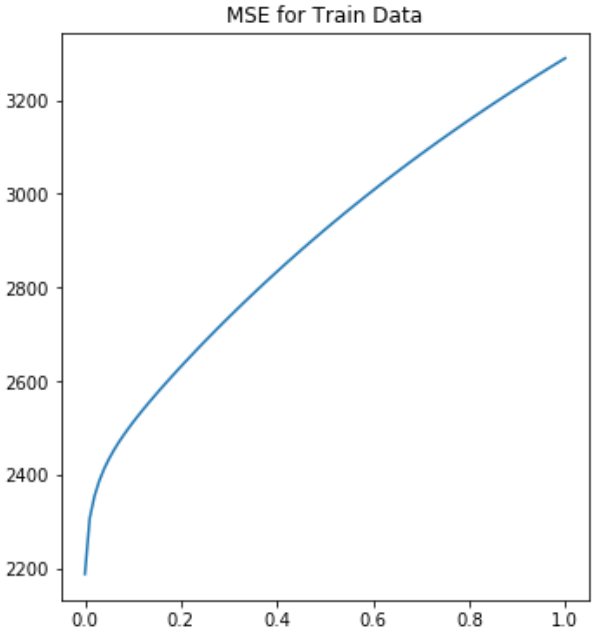
It is clearly evident from the data in the above table that when Intercept is used error decreases drastically in case of any input data (training data or test data). Always lesser MSE is better. So for Linear Regression always, use the intercept to get lesser MSE.

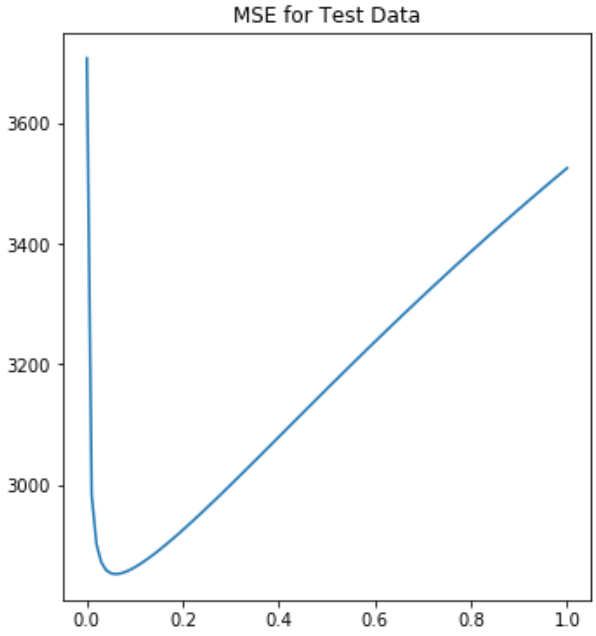
PROBLEM 3: EXPERIMENT WITH RIDGE REGRESSION

Here our aim is to minimize the regularized square loss. This is achieved by estimating the weights for the ridge regression. For a line. Thus, we obtain normal equation: X.T Xw = X.Ty, the corresponding solution for w is: $(X^TX)^{-1}(X^Ty)$

Similarly, for Ridge Regression we solve w to: $(Lambda.I+ X^TX)^{-1}(X^Ty)$

Using this expression, we learn the weights from the train data and then we apply these weights for test data. Thereby we compute the Mean Squared Error for various values of Lambda (0.00-1.00) as shown below for both Train Data and Test Data.





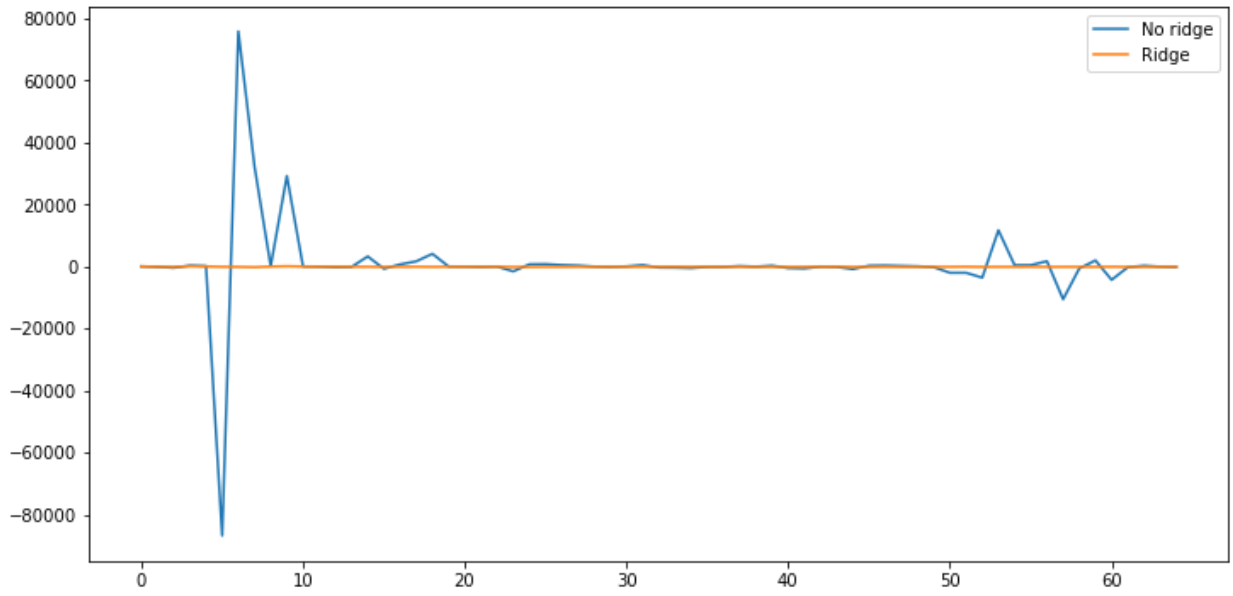
Optimal Lambda:

From the above graph it is evident that as we increase lambda above an optimal value, the MSE is increasing. That is our attempt to **penalize the weights magnitude** is increasing the error for both training data and test data. Therefore, we might consider sticking lambda to an optimal value. This is another way of looking into **Bias -Variance tradeoff.** Also the error in the testing data is higher that the train data.

By differentiating the MSE with respect to lambda the optimal for lambda is found to be at **Lambda = 0.06** where MSE is 2851.3. For this value of lambda, MSE is minimum that means the error for our learned weights is minimum, thereby we can stop penalizing the weight magnitudes above this lambda.

Comparison for weight from problem 2 and problem3:

After applying ridge regression, most of the values are in the orders of 10 for optimal lambda value. Whereas the weights from problem 2 varies from -80000 to 80000. The comparison can be seen from the below figure, since the change in weights for problem 3 is very small compared to weights in problem 2, the weights seem to be in a straight line.



Comparison for MSE from problem 2 and problem3:

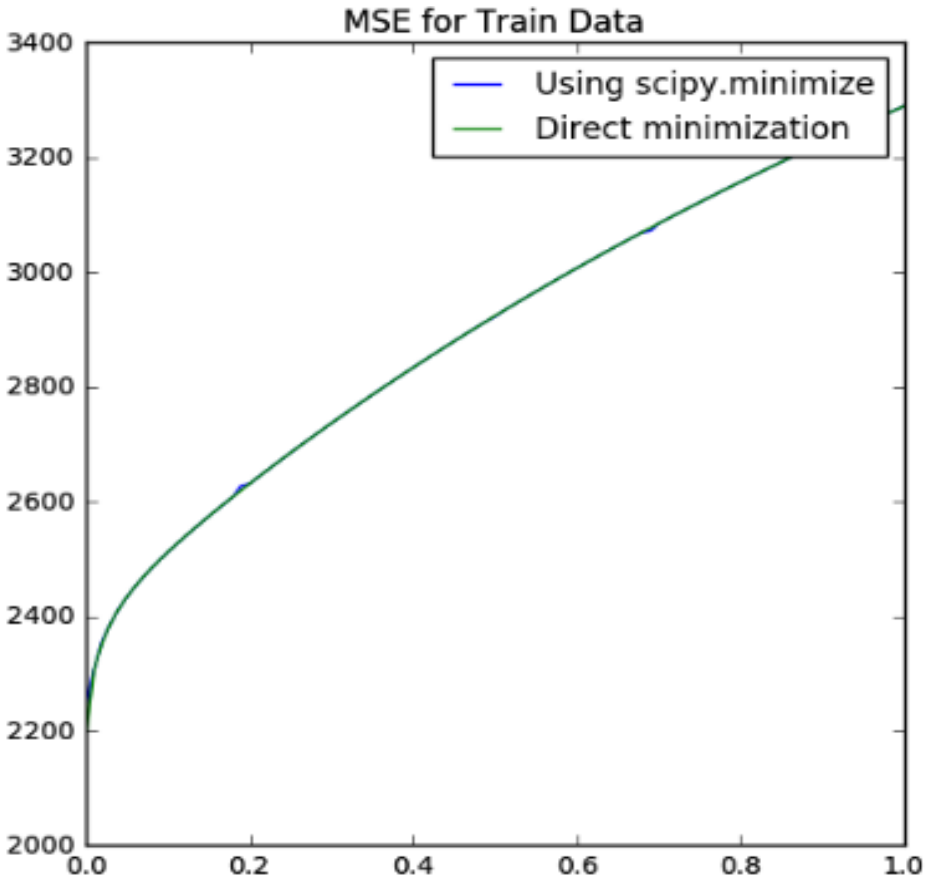
The MSE for problem 2 with intercept is 3707.84018132

Whereas the MSE from problem 3 with intercept using ridge regression for optimal value of lambda is: 2851.3. Therefore, we conclude that Ridge regression is definitely better algorithm for our analysis.

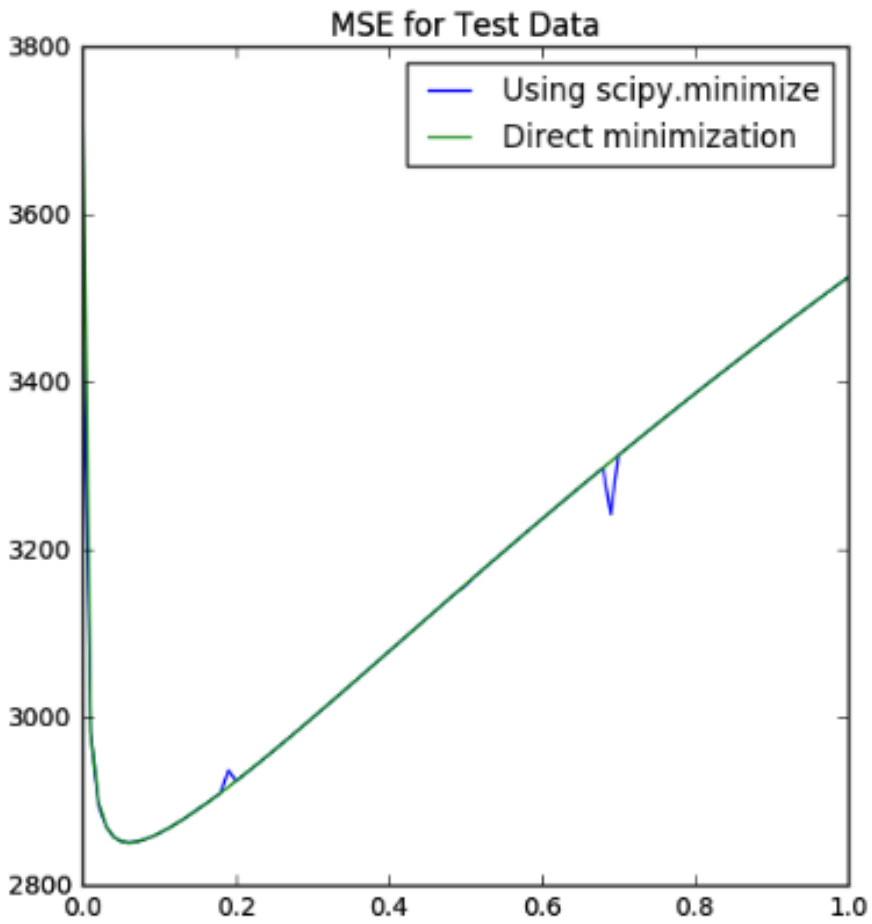
PROBLEM 4: USING GRADIENT DESCENT

The following graphs are observed by comparing Gradient Descent for Ridge Regression Learning using scipy minimize vs direct minimization from problem3:

For train data:



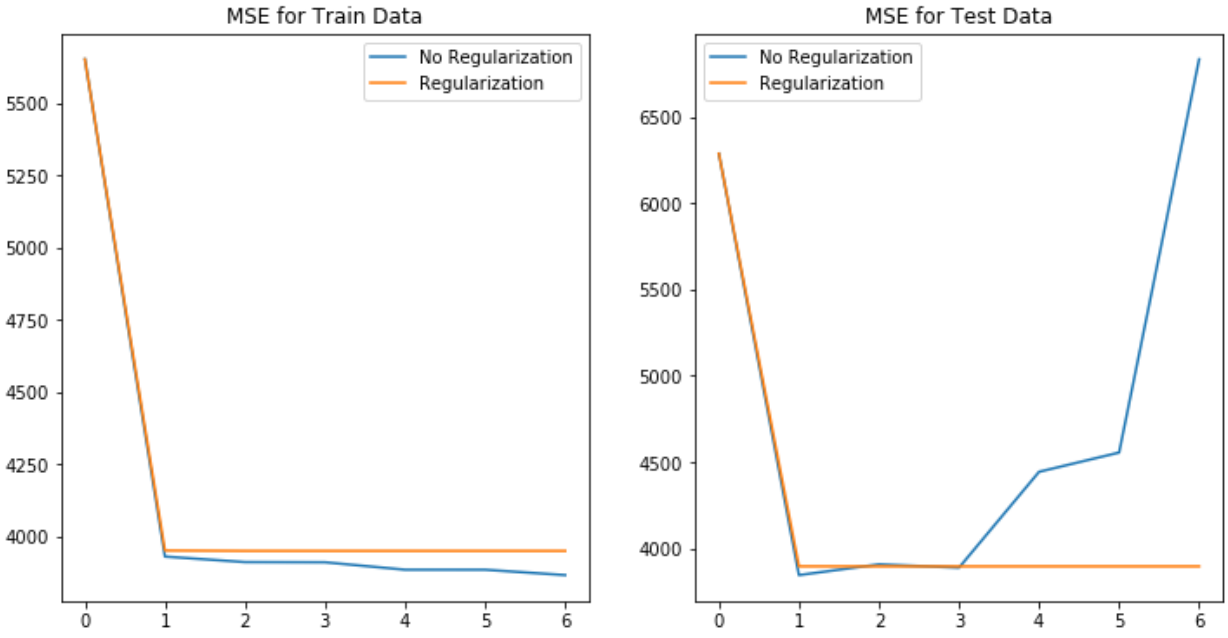
For Test data:



The graphs derived using the gradient descent look similar to the graphs derived in problem 3 by varying the lambda value. Its observed that the error increases with the increased value of lambda for training data. Also, we observed that the error for test data decreased upto a certain value of lambda and then it increased with lambda. The value of lambda where this deviation happened according to problem 3 is 0.06. The plots looked similar even with the use of gradient descent in this problem.

PROBLEM 5: NON-LINEAR REGRESSION

From the below figure, it is very evident that the MSE for train data and test data is reduced or atleast controlled after **Regularization** for higher values of "p" in this non linear regression. This avoids overfitting at higher dimensions(p).



Optimum Value of p in terms of test data for lambda =0

By observing the test data's MSE, we can see that MSE increases after p=2, after which the higher ordered polynomial overfits the train data and causes more test error, which is undesirable. Therefore, we can confine our p with **p =1** as our optimal value.

Optimum Value of p in terms of test data for lambda = lambda_optimal:

However, after regularization the increase in p value didn't increase the MSE for test data, so with regularization we could consider p= [2 to 6] as our optimal values.

Compare the results for both values of lambda:

For lambda = 0	For lambda = Lambda_opt
6286.40479168	6286.88196694
3845.03473017	3895.85646447
3907.12809911	3895.58405594
3887.97553824	3895.58271592
4443.32789181	3895.58266828
4554.83037743	3895.5826687
6833.45914872	3895.58266872

PROBLEM 6: INTERPRETING RESULTS

In Linear Regression model,

MSE without intercept train: 19099.44684457

MSE with intercept train: 2187.16029493

MSE without intercept test: 106775.36155789

MSE with intercept test: 3707.84018132

After applying ridge regression, the weights have been regularized and the magnitude of weights have come down to orders of 10. Thereby we justify the application of ridge regression if we wanted to control the order of weights and avoid overfitting.

This model works very well with minimal weights(magnitude) and with MSE equals to: **2851.3** (for optimal lambda value = 0.06)

For ridge regression models, the training data showed an increase in the error with increased value of the regularization parameter. For test data, the error decreased with the regularization parameter, reached an optimal value and then increased with it. The graphs for both variants of the ridge regression models are similar which infers that this dataset gave similar result for both these models.

In nonlinear regression model, the MSE increases for higher values of p on test data, which is undesirable. But after applying regularization, at optimal lambda value, the minimum MSE was found to be: **3895.58266872** which is more than what we have seen from problem 3. If left unregularized the MSE peaked up to: 6833.45914872, which is also unwanted.

Metric to be used for best setting:

When we check all the results, ridge regression performs better than the linear regression. Performing ridge regression with gradient descent and without gradient descent gave similar result for the given dataset. So, we can use either of these approaches for our ridge regression

model. However, for large datasets, it's recommended to use gradient descent to avoid costly matrix inversions that we need to perform in ridge regression. Also, gradient descent would work for singular matrices where inversion can't be calculated for classic ridge regression models. For non linear regression, the optimal value was reached when p==1, that means we do not really require higher dimensions of p. Just linear regression would suffice for this dataset.