University at Buffalo
CSE 574 Introduction To Machine Learning, Spring 2017

Programming Assignment 2
Classification and Regression

Project Report

Submitted by

Group 49

Anurag Devulapalli (5020 8153)
Sandeep Shenoy (5020 5705)
Vipin Kumar (5020 8397)

April 11, 2017

Problem 1: Experiment with Gaussian Discriminators

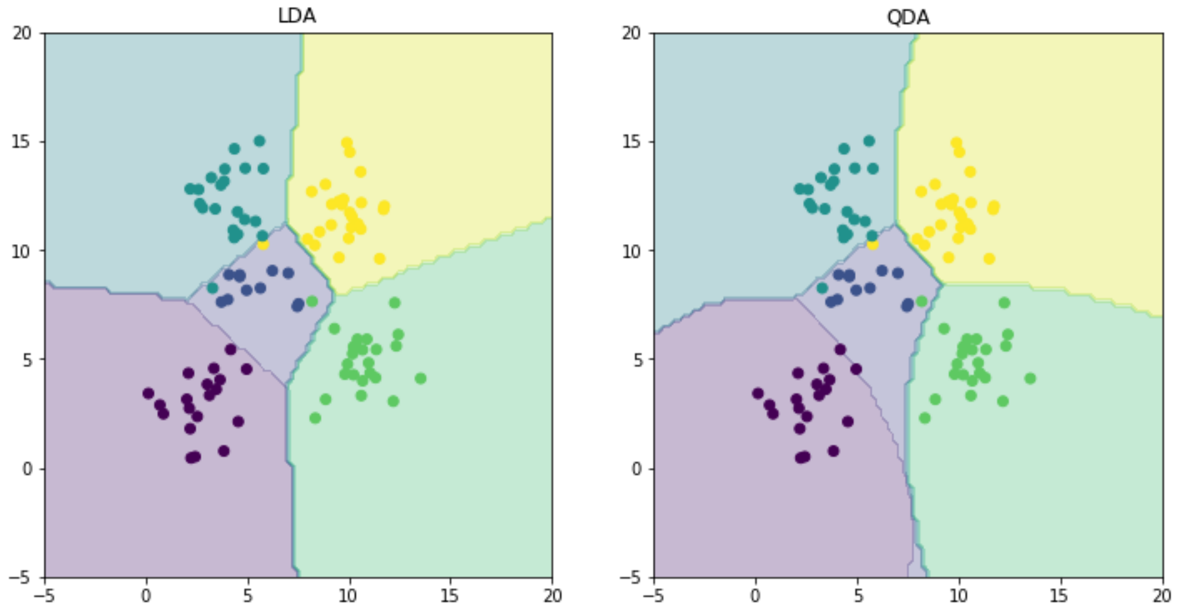
Requirement:

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.

a. Report the accuracy of LDA and QDA on the provided test data set (sample test)

LDA Accuracy = 97
QDA Accuracy = 96

b. Also, plot the discriminating boundary for linear and quadratic discriminators



c. Explain why there is a difference in the two boundaries

For LDA we use a common covariance matrix irrespective of the class. In QDA we use a different covariance matrix for each class. This results in a straight boundary (linear) for LDA whereas the boundaries become slightly curved (quadratic) for QDA.

Problem 2: Experiment with Linear Regression

Requirement:

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

a. Observation:

Data Type	Mean Squared Loss (MSE) without intercept	Mean Squared Loss (MSE) with intercept	% Improvement
Train Data	19099.4468446	2187.16029493	88.54%
Test Data	106775.361558	3707.84018132	96.5%

b. Which one is better?

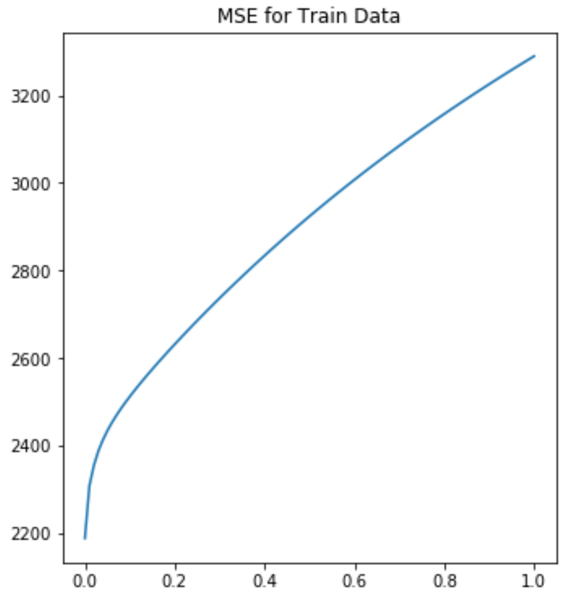
In both the cases (training and test data) we observe that MSE with intercept is smaller. When we do not have an intercept, the line has to pass through the origin and hence it may provide undesirable results which why we observe high loss values. Introduction of an intercept allows the model to pass more accurately through the data and hence gives better results. Thus the model *with intercept* is better.

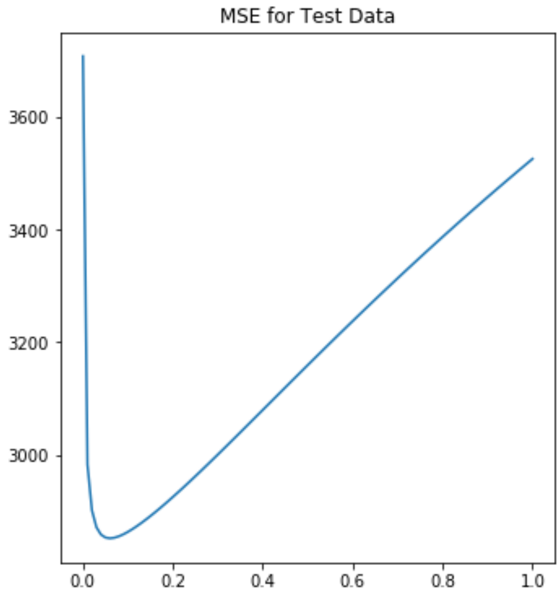
Problem 3: Experiment with Ridge Regression

Requirement:

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?

a. MSE for training and test data using ridge regression parameters



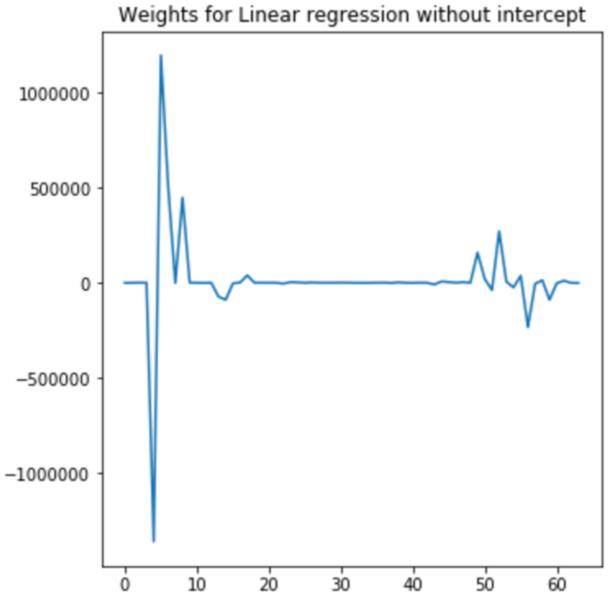


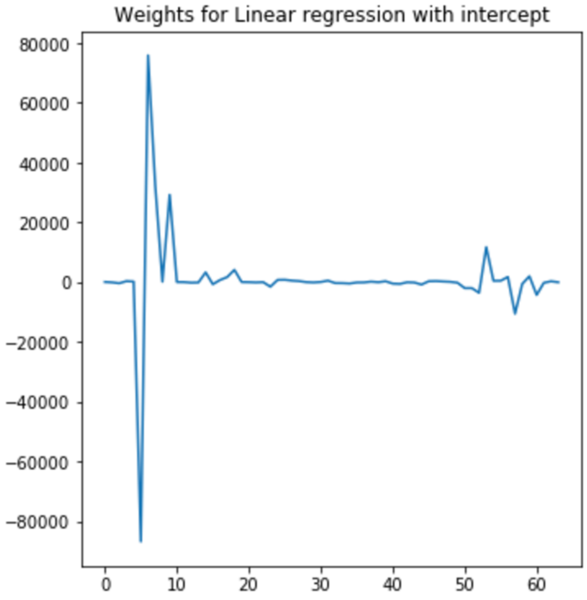
Observation: We observe that the MSE increases as the λ value increases for both training and test data when using ridge regression.

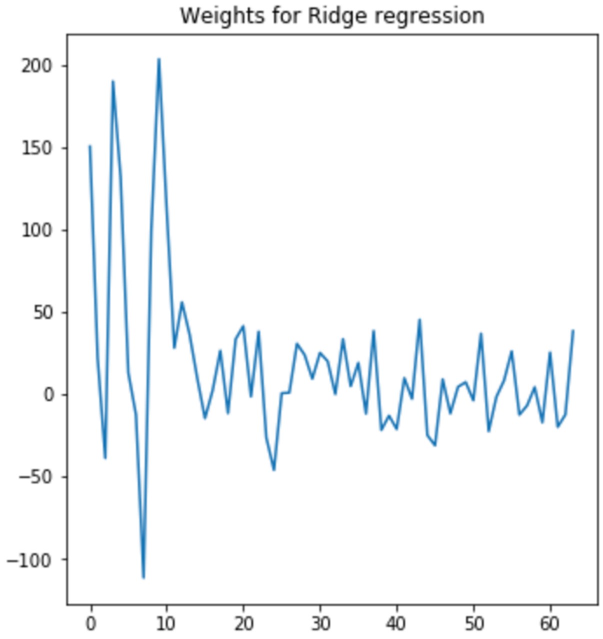
Lambda	MSE for train data	MSE for test data
0.0	**2187.16**	3707.84
0.01	2306.83	2982.45

0.02	2354.07	2900.97
0.03	2386.78	2870.94
0.04	2412.12	2858.0
0.05	2433.17	2852.67
0.06	2451.53	**2851.33**
0.07	2468.08	2852.35
0.08	2483.37	2854.88
0.09	2497.74	2858.44
...	...	...
0.93	3245.35	3478.16
0.94	3251.81	3484.99
0.95	3258.23	3491.8
0.96	3264.61	3498.57
0.97	3270.96	3505.32
0.98	3277.26	3512.04
0.99	3283.53	3518.73
1.0	3289.76	3525.39

b. Relative magnitudes of weights learnt using OLE and weights learnt using ridge regression.


Weights for Linear regression without intercept


Weights for Linear regression with intercept


Weights for Ridge regression

Observation: We observe that the range of weights for ridge regression varies from -100 to +200. Whereas for OLE it varies from -10,00,000 to + 10,00,000 for OLE without intercept and from -80,000 to +80,000 for OLE with intercept.

c. Compare the two approaches in terms of errors on train and test data.

	Lambda	MSE	
		Training data	**Test data**
OLE	Without intercept	19099.44	106775.36
	With intercept	2187.16	3707.84
Ridge Regression	0	2187.16	3707.84
	0.06	2451.52	2851.33

Observation: We observe that OLE with intercept is somewhat similar to ridge regression. If we set λ to 0.06 we get the best results.

d. What is the optimal value for λ and why?

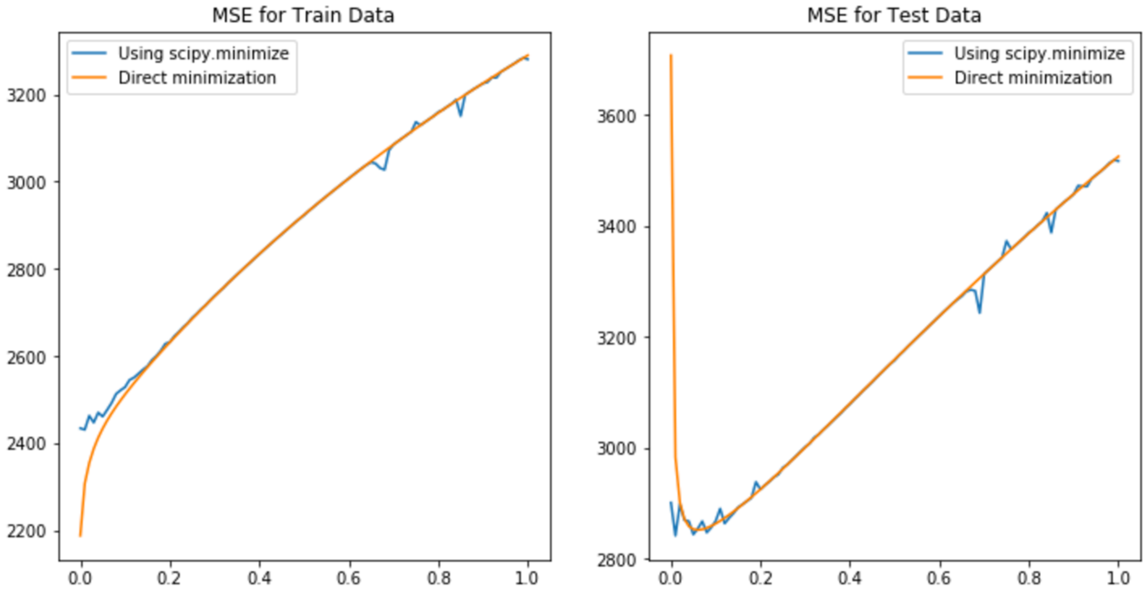
Optimal value for λ is 0.06 in case of test data and 0 in case of training data since at this point the MSE is the least.

Problem 4: Using Gradient Descent for Ridge Regression Learning

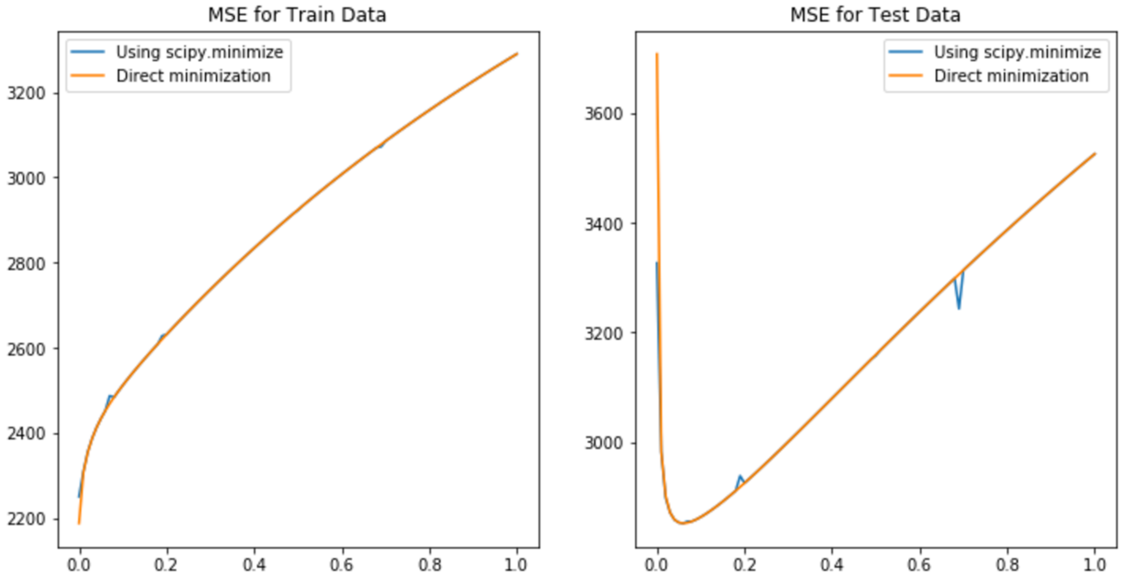
Requirement:

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.

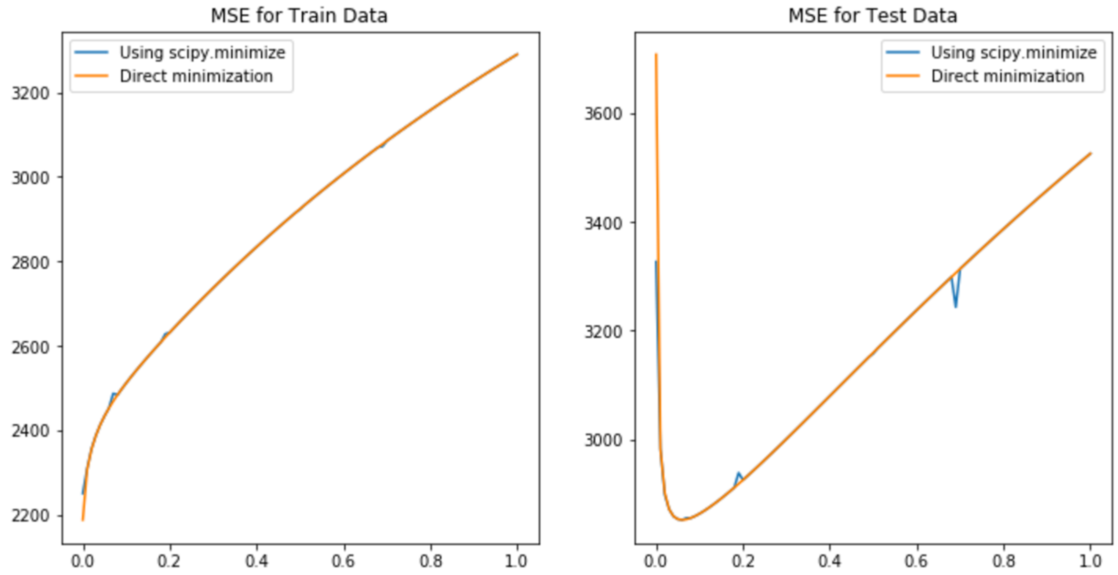
For iteration = 20



For iteration = 100



For iteration = 150



Observation:

By using scipy.minimize at lower iterations such as 20, the errors obtained deviate a little from the results obtained by direct minimization. As we increase the number of iterations, the scipy.minimize learns more and gives results very similar to the ones obtained by direct minimization.

Problem 5: Non-linear Regression

Description:

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.

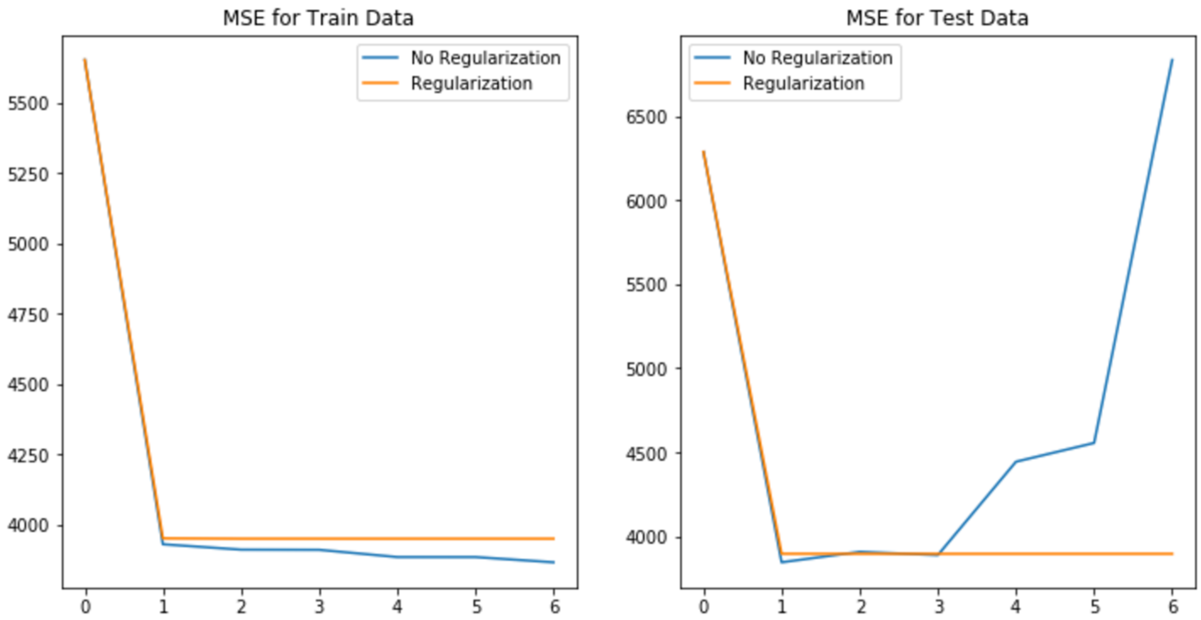
a. Compare the results for both values of λ

p	**MSE (No regularization)**	**MSE (regularization)**
0	6286.4	6286.88
1	**3845.03**	3895.86
2	3907.13	3895.58
3	3887.98	3895.58
4	4443.33	**3895.58**
5	4554.83	3895.58
6	6833.46	3895.58

b. What is the optimal value of p in terms of test error in each setting?

Optimal value of p (no regularization) is **1**
Optimal value of p (with regularization) is **4**

c. Plot the curve for the optimal value of p for both values of λ and compare.



Problem 6: Interpreting Results

Description:

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

Our observations:

Model	**MSE**	
	Training data	**Test data**
Linear OLE Regression with intercept	19099.44	106775.36
Linear OLE Regression without intercept	2187.16	3707.84
Ridge Regression	2451.52	2851.33
Ridge Regression with gradient descent	**2451.52**	**2851.33**
Non-linear Regression with p = 1	3845.03	3895.86
Non-linear Regression with p = 4	4443.33	3895.58

Summary of approaches:

Linear OLE Regression with intercept: In this approach, the model passes through the origin. Thus, the expected results are highly inaccurate as the model does not pass through the data correctly.

Linear OLE Regression without intercept: Here the model passes through most of the data due to introduction of an intercept, thus giving better results.

Ridge Regression: Adds penalties to the magnitude of the coefficients to give better results

Ridge Regression with gradient descent: Maximizes the log likelihood to avoid computation of inverse matrix

Non-linear Regression with p = 1: Higher order polynomials in input

Non-linear Regression with p = 4: Higher order polynomials in input

Our Recommendation:

We recommend the approach of **Ridge Regression with gradient descent** to predict the diabetes level. Our recommendation is based on the following observations

1. Ridge Regression has the lowest MSE value for test data compared to all other models.
2. Low MSE means that the deviation from the true label is the least and hence the accuracy of the prediction will be high.
3. Using gradient descent in weight calculation we can avoid the computation of $(X^{T} X)^{-1}$.