skandiba(50291583) scheekat(50291064) skothapa(50289072)

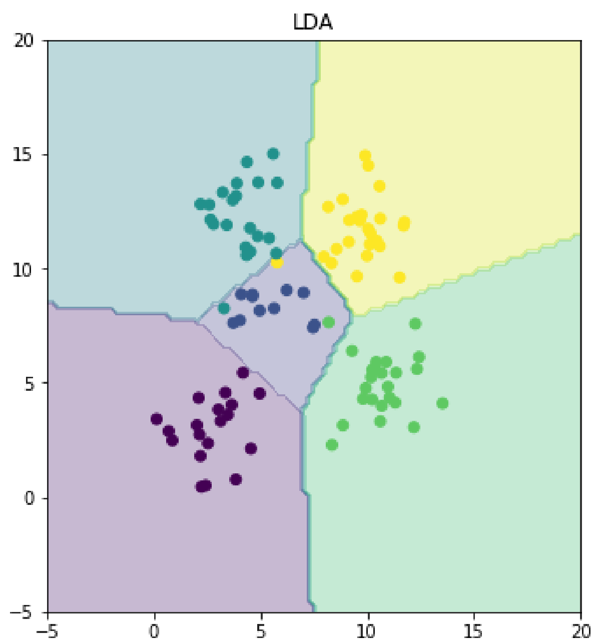
CSE574 Spring 2019 Introduction to Machine Learning Programming Assignment 1

Classification and Regression

We have been given two data sets "sample.pickle" and "diabetes.pickle". We used various Classification and Regression techniques on the data.

Problem 1: Experiment with Gaussian Discriminators

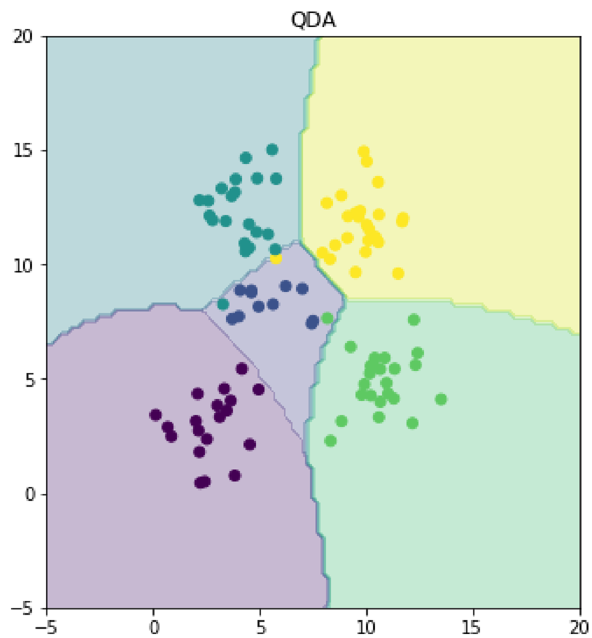
Linear Discriminant Analysis:
LDA Accuracy: 97.0



Observations:

- We observe discriminant boundaries for LDA which differentiate 5 classes.
- In LDA, the covariance for each class is assumed to be the same.
- Boundaries plotted are Linear.

Quadratic Discriminant Analysis:
QDA Accuracy: 97.0



Observations:

- We observe discriminant boundaries for QDA which differentiate 5 classes.
- In LDA, the covariance for each class is calculated separately.
- Boundaries plotted are parabolic (quadratic decision boundaries).

Problem 2 : Experiment with Linear Regression

Linear Regression (Without Intercept):

Train MSE: 19099.4468446
Test MSE: 106775.361451

Linear Regression (With Intercept):

Train MSE: 2187.16029493
Test MSE: 3707.84018096

Observations:
From the above results we can say that Linear Regression (With Intercept) give much better results for the Training and Test data. Because when Linear Regression is modeled without an Intercept, it has to pass through origin, so it is possible that is not the best solution. When Linear Regression is modeled with Intercept, then model gets closely aligned with the actual data and gets better results.

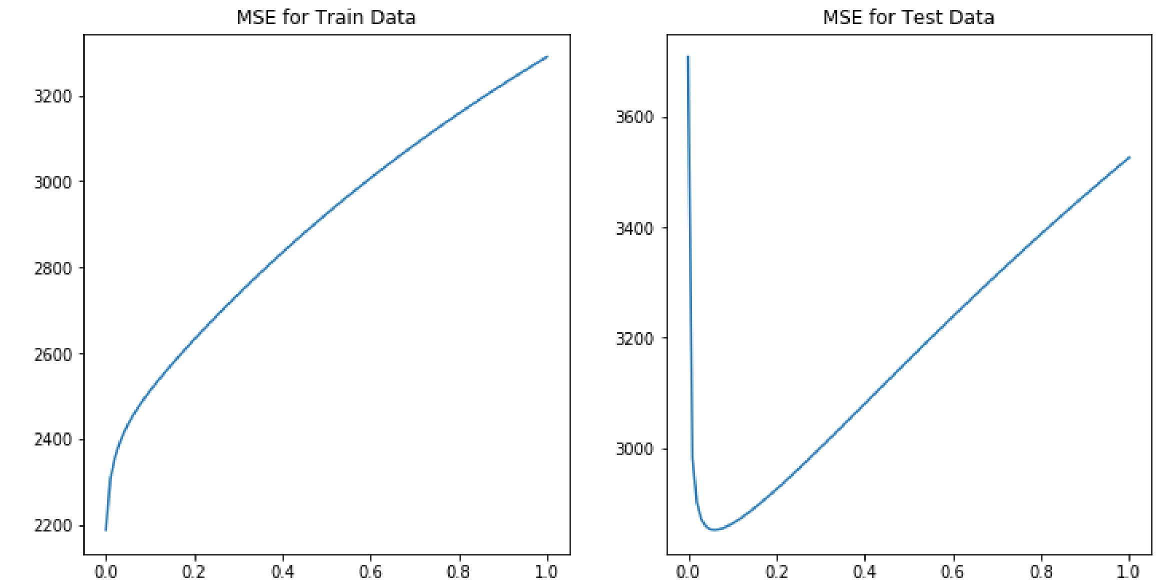
- Improvement in Train error is by 88.549%
- Improvement in Test error is by 96.528%

Problem 3 : Experiment with Ridge Regression

a. MSE for training and test data using ridge regression

Lambda	Trainingdata	Test data
0.0	[2187.16029493]	[3707.84018148]
0.01	[2306.83221793]	[2982.44611971]
0.02	[2354.07134393]	[2900.97358708]
0.03	[2386.7801631]	[2870.94158888]
0.04	[2412.119043]	[2858.00040957]
0.05	[2433.1744367]	[2852.66573517]
0.06	[2451.52849064]	[2851.33021344]
0.07	[2468.07755253]	[2852.34999406]
0.08	[2483.36564653]	[2854.87973918]
0.09	[2497.74025857]	[2858.44442115]
0.1	[2511.43228199]	[2862.75794143]
0.11	[2524.60003852]	[2867.63790917]
0.12	[2537.35489985]	[2872.96228271]
0.13	[2549.77688678]	[2878.64586939]
0.14	[2561.92452773]	[2884.62691417]
0.15	[2573.84128774]	[2890.85910969]
0.16	[2585.55987497]	[2897.30665895]
0.17	[2597.10519217]	[2903.94112629]
0.18	[2608.49640025]	[2910.73937213]
0.19	[2619.74838623]	[2917.68216413]
0.2	[2630.8728232]	[2924.75322165]
0.21	[2641.87894616]	[2931.93854417]
0.22	[2652.77412633]	[2939.22592987]
0.23	[2663.56430077]	[2946.60462378]
0.24	[2674.25429667]	[2954.06505602]
0.25	[2684.84807809]	[2961.59864341]
0.26	[2695.34893502]	[2969.19763677]

0.96	[3264.61386081]	[3498.57090566]
0.97	[3270.95717015]	[3505.3183244]
0.98	[3277.26258207]	[3512.03802854]
0.99	[3283.53048993]	[3518.7300819]
1.0	[3289.7612813]	[3525.39455263]

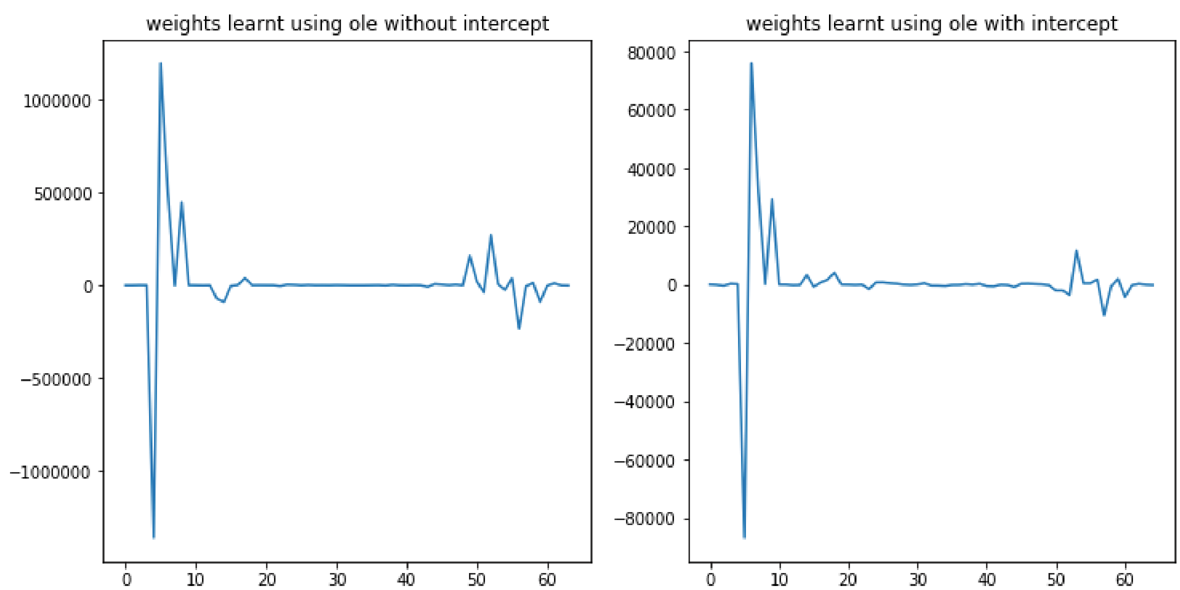


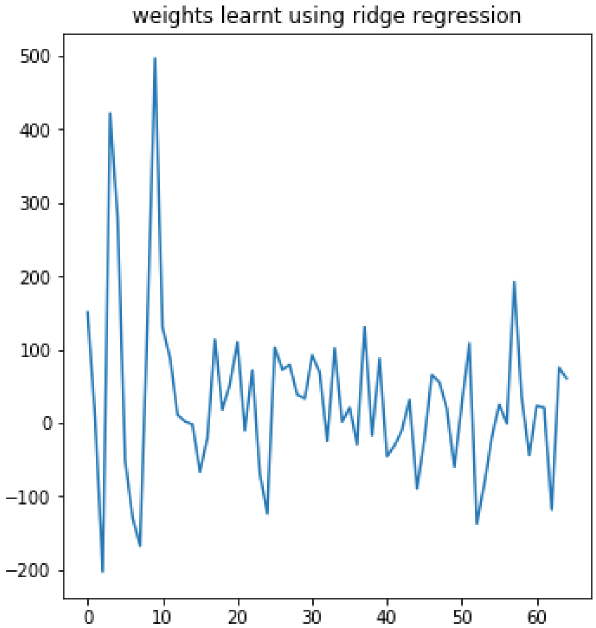
Observations:

From the plots, we observe that the MSE value increases as the λ value increases for both training and test data when using ridge regression.

---

b. Comparison of relative magnitudes of weights learnt using OLE in problem 2 and weights learnt using ridge regression





Observations :
From the plots we see that the range of weights 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. Range of weights for ridge regression varies from -200 to +500.

c. Comparison of MSE for two approaches:

		Training data	Test data
OLE	Without intercept	19099.44	106775.36
	With intercept	2187.16	3707.84

	lambda	Training data	Test data
Ridge Regression	0	2187.16	3707.84
	0.06	2451.52	2851.33

Observations :
From the above tables we notice that MSE values are comparable for ridge regression and OLE with intercept case.
Also the error value is less when we set the lambda value to obtained optimal value (0.06)

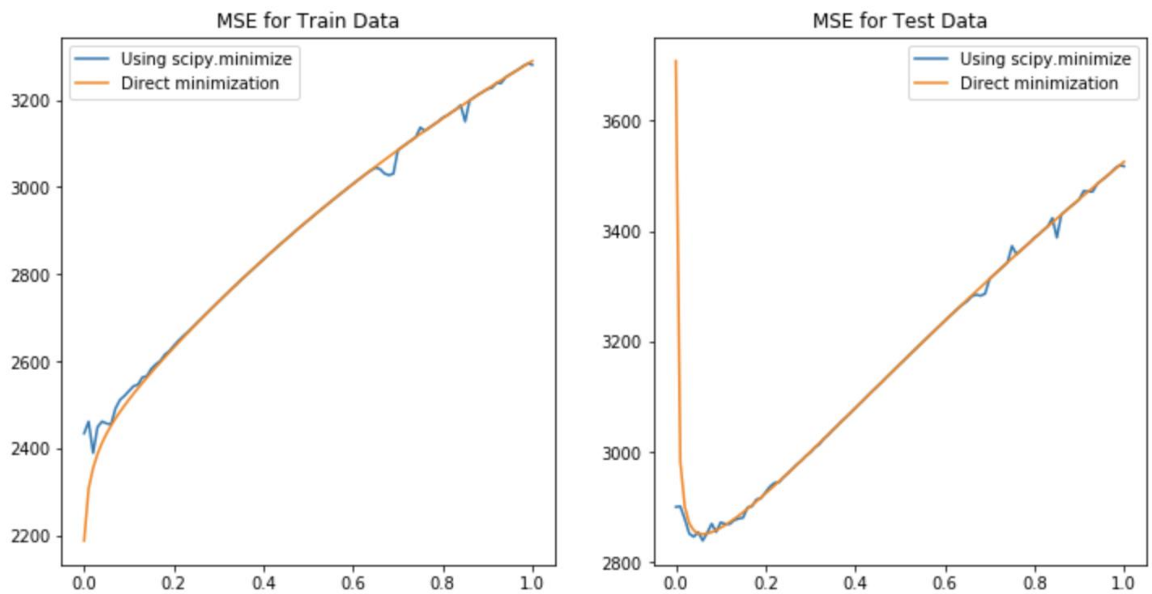
---

d. Optimal value of lambda and why?

Optimal value of Lambda obtained is 0.06. It is optimal because the error values are least for test data when lambda is set to 0.06

Problem 4: Using Gradient Descent for Ridge Regression Learning

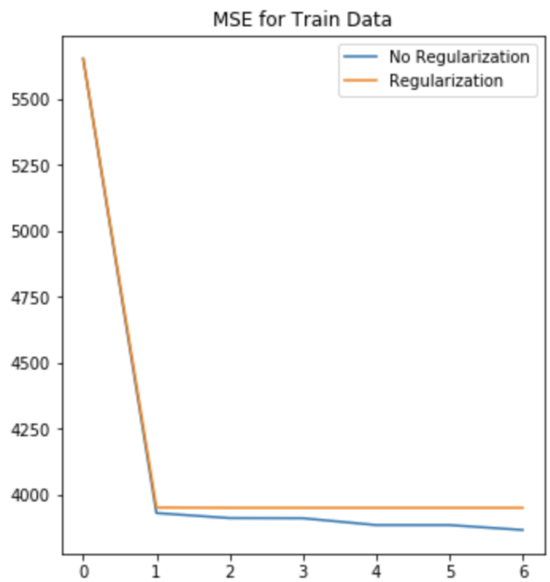
The plots for RMSE for training vs lambda and testing vs lambda look like this:

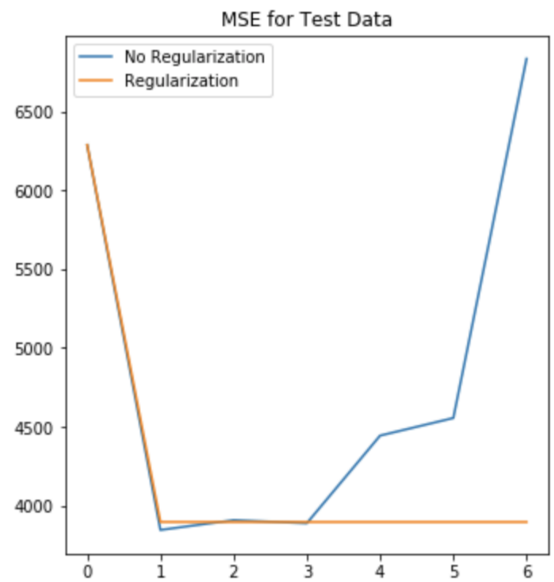


We can see that the value of MSE for testing data using scipy.minimize is 2839.41934832 and the value of MSE for testing data by using direct minimization is 2851.33021344. Both the values are very close.

The optimal value of lambda is 0.06 for both the cases.

Problem 5 : Non Linear regression





The plots for MSE vs p value for training and testing data are given below:

Comparison of p values against testing error with lambda value 0 and optimal lambda value = 0.06

P Value	Test error with zero value of lambda	Test error with optimal value of lambda = 0.06
0	6286.40479168	6286.88196694
1	3845.03473017	3895.85646447
2	3907.12809911	3895.58405594
3	3887.97553824	3895.58271592
4	4443.32789181	3895.58266828
5	4554.83037743	3895.5826687

P Value	Test error with zero value of lambda	Test error with optimal value of lambda = 0.06
6	6833.45914872	3895.58266872

From the plot we can see that the optimal value for p is 1 for lambda value zero and the optimal value for p is 4 for lambda value 0.06.

We can see that when we do not use regularization, when lambda is zero, the test error is optimal at p=1 and as the value of p increases the test error increases eventually and at p = 6 it becomes greater than when p=0. Note that p=0 is the case where we use a horizontal line as the regression line and p = 1 is the same as linear ridge regression. The reason for the test error becoming greater at p=6 is because of overfitting. P = 6 means they are higher order polynomials. As we can see in the plot the training error decreases but the model is heavily bound to the training data and when we use the testing dataset the error steeply increases.

But when we use regularization, when lambda is optimal at 0.06, the test error is lowest at p = 4. The test error doesn't change much in this case as opposed to what we saw in the case of lambda = 0. This is because the regularization term penalizes heavy weights and therefore even with higher order polynomials, the corresponding weights will be low and the curve is smooth and linear.

Problem 6: Interpreting the results.

a. Compare the various approaches in terms of training and testing error.

Model Type	Train MSE	Test MSE
Linear Regression (without intercept)	19099.4468446	106775.361451
Linear Regression (with intercept)	2187.16029493	3707.84018096
Ridge regression	2451.52849064	2851.33021344
Ridge regression with Gradient descent	2388.6579755	2839.41934832

Non- Linear Regression (No regularization)	3911.18866493	3845.03473017
Non- Linear Regression (With regularization)	3950.68233514	3895.58266828

b. What metric should be used to choose the best setting?

By Observing the above data, We should consider the TEST MSE results to choose Model type as it shows the results performed on the Test data .

Recommendation:

We recommend the model type ridge regression with gradient descent to predict the diabetes level.

Our recommendation is based on the following observations:

1. The MSE value is least for ridge regression with gradient descent among all the other models
2. Low MSE means that the deviation from the true label will be less. So accuracy of the prediction will be high
3. If we use gradient descent we can avoid the Computation of $(X^{\top}X)^{-1}$ in the calculations.