CSE 574 Introduction to Machine Learning

Programming Assignment 2

Classification and Regression

Group 39

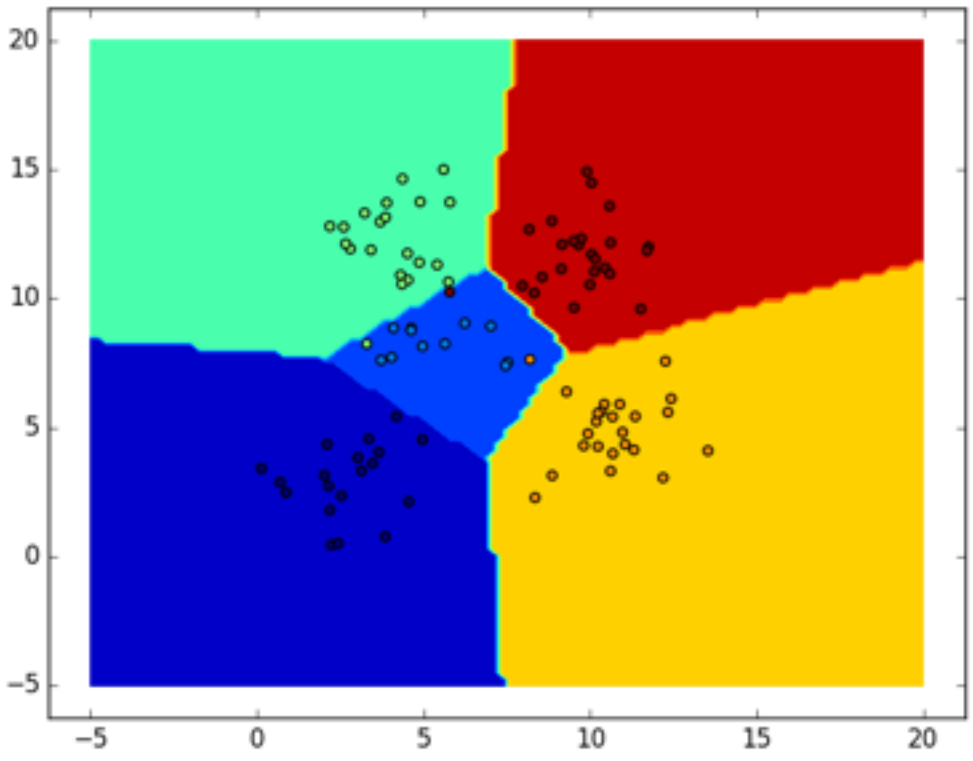
Rohith Doraiswamy Konoor (50169785)
Falguni Bharadwaj (50163471)
Malavika Tappeta Reddy (50169248)

Problem 1:

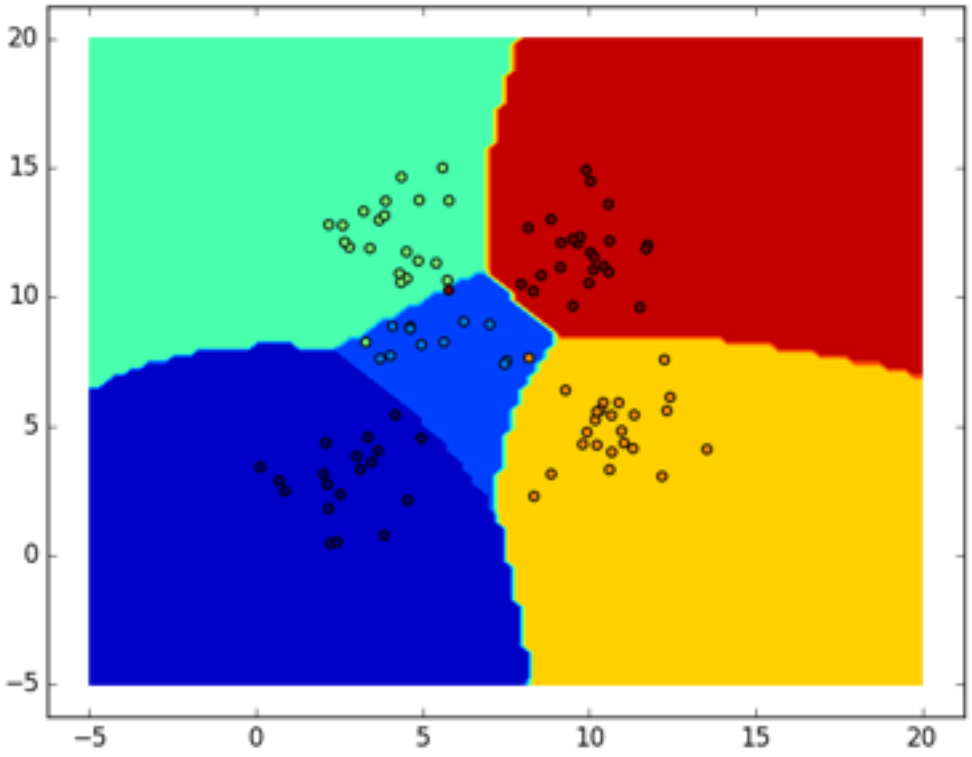
We implemented Linear Discriminant Analysis(LDA) and Quadratic Discriminant Analysis(QDA) by training it using training data set and tested it using the testing data set. The accuracy we obtained for LDA is 97% and QDA is 97% too.

From the following graphs we can see that the boundary lined in LDA are less curved compared to QDA. One of the reasons for this is the calculation of co-variance. In LDA co-variance is calculate for the entire data set where as in QDA co-variance is calculated for each class. Even though in this particular problem the accuracy is the same for LDA and QDA when applied to bigger and more complex data sets QDA proves to be more accurate than LDA because of its flexible boundary.

The discriminating boundary for LDA is shown below.



The discriminating boundary for QDA is shown below.



Problem 2:

We calculated the RMSE of training and test data sets in two was. Firstly without an intercept and then with an intercept(bias) term added to the set. Following are our observations:

RMSE without intercept for test data: 326.764994389
RMSE with intercept for test data: 60.8920370937

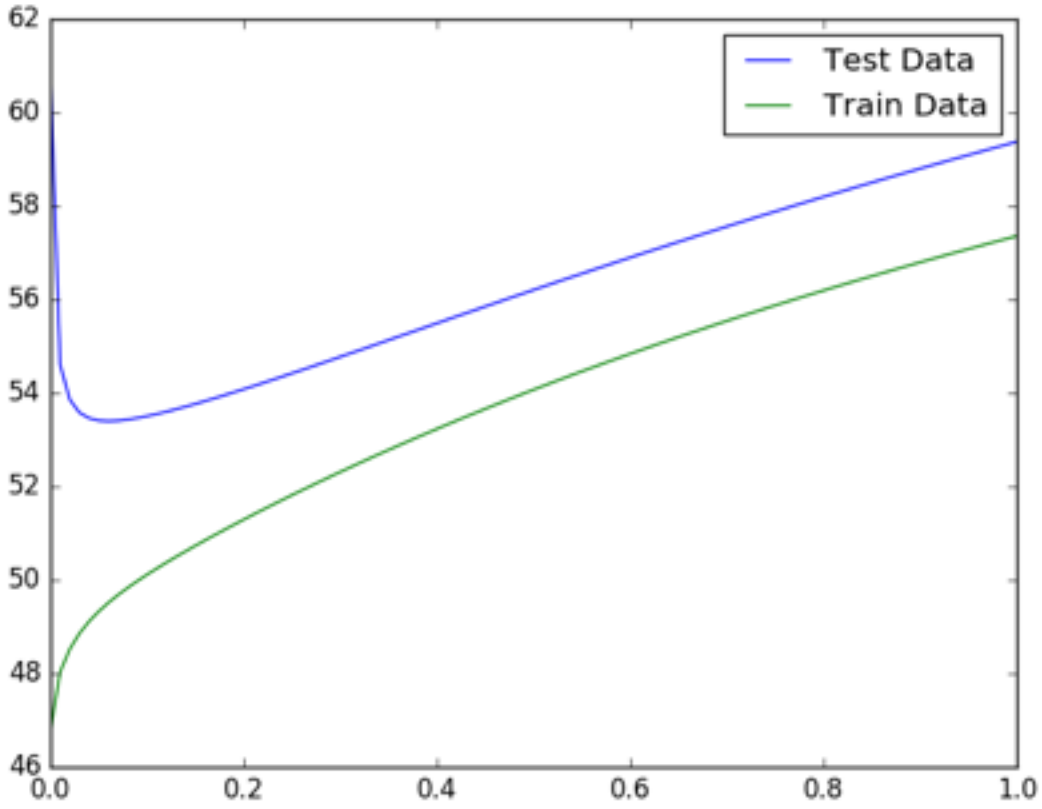
RMSE without intercept for training data: 138.20074835
RMSE with intercept for training data: 46.7670855937

As we can see from above, in both training and test data sets, the RMSE with intercept is lesser than RMSE without intercept. Thus, RMSE with intercept is better because the error is lesser with an intercept.

Problem 3:

Following is the graph of errors on test and training data for ridge regression for different values of lambda. According to it, optimal lambda value is approximately 0.06. This is the value for which we have observed the minimum RMSE. Further increasing values of lambda, errors continue increasing thus hindering effective performance.

Minimum RMSE value of ridge regression for test data: 53.3978483971
Minimum RMSE value of ridge regress for training data: 46.7670855937



In terms of errors for training and test data, we are getting less error for training data and more for test data as seen from the graph.
On comparing the weights of OLE Regression and Ridge Regression we see that values for Ridge Regression are higher than OLE Regression.

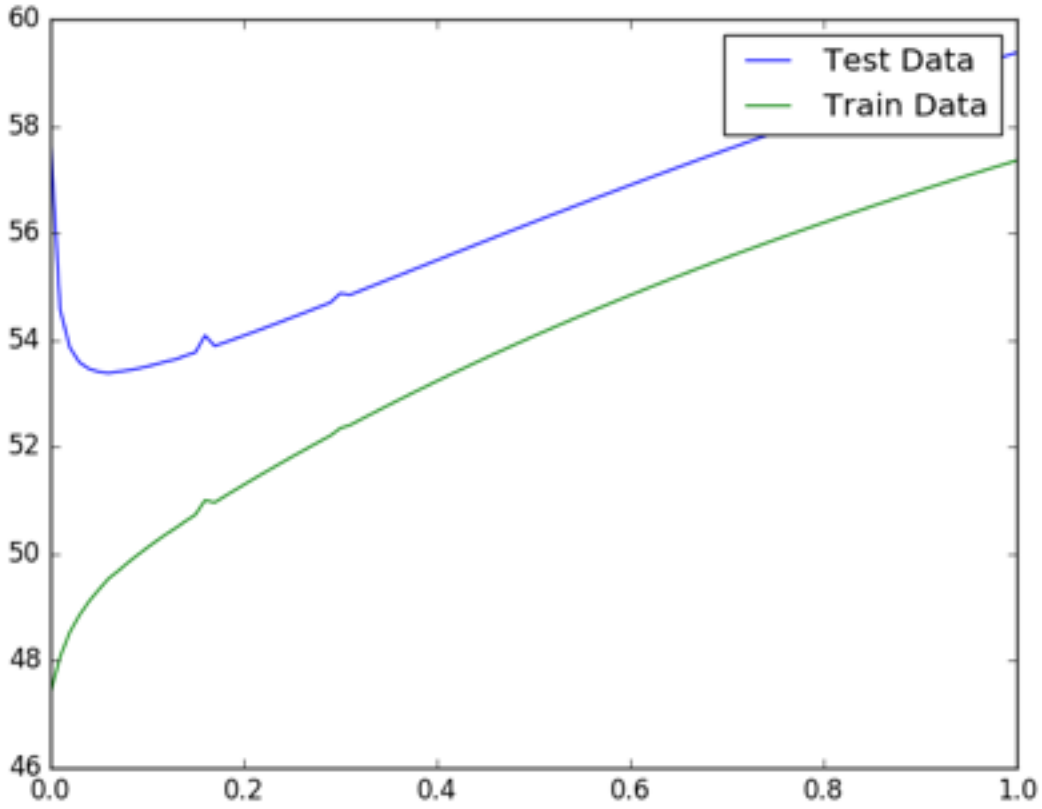
Weights of Ridge Regression	Weights of OLE Regression	Difference in both weights
[7.46131102e+01]	[1.48154876e+02]	[7.35417658e+01]

Weights of Ridge Regression	Weights of OLE Regression	Difference in both weights
[2.99197971e-01]	[1.27485221e+00]	[9.75654234e-01]
[9.80791747e-02]	[-2.93383522e+02]	[-2.93481602e+02]
[1.66425859e+00]	[4.14725449e+02]	[4.13061190e+02]
[1.26377850e+00]	[2.72089134e+02]	[2.70825356e+02]
[3.05923991e-01]	[-8.66394570e+04]	[-8.66397630e+04]
[2.78648760e-01]	[7.59144679e+04]	[7.59141893e+04]
[-1.25301657e+00]	[3.23416228e+04]	[3.23428758e+04]
[1.17596025e+00]	[2.21101216e+02]	[2.19925255e+02]
[1.69178145e+00]	[2.92995512e+04]	[2.92978594e+04]
[1.33733260e+00]	[1.25230360e+02]	[1.23893028e+02]
[3.03101761e-02]	[9.44110833e+01]	[9.43807731e+01]
[2.49345592e-01]	[-9.38628632e+01]	[-9.41122088e+01]
[2.63805425e-01]	[-3.37282800e+01]	[-3.39920854e+01]
[-9.89267283e-02]	[3.35319772e+03]	[3.35329665e+03]
[-1.03026319e-01]	[-6.21096300e+02]	[-6.20993273e+02]
[-9.53648418e-02]	[7.91736534e+02]	[7.91831899e+02]
[3.27109110e-01]	[1.76776039e+03]	[1.76743328e+03]
[-3.05116694e-01]	[4.19167405e+03]	[4.19197917e+03]
[6.89719464e-01]	[1.19438121e+02]	[1.18748401e+02]
[4.36159719e-01]	[7.66103401e+01]	[7.61741803e+01]
[4.91525755e-02]	[-1.52001293e+01]	[-1.52492819e+01]
[3.14509155e-01]	[8.22424594e+01]	[8.19279502e+01]
[-4.20788242e-01]	[-1.45666208e+03]	[-1.45624130e+03]
[-5.97629059e-01]	[8.27386703e+02]	[8.27984332e+02]
[1.19592393e-01]	[8.69290952e+02]	[8.69171360e+02]
[-2.81234006e-01]	[5.86234495e+02]	[5.86515729e+02]
[7.34522978e-02]	[4.27026727e+02]	[4.26953274e+02]
[1.81080060e-01]	[9.02467690e+01]	[9.00656890e+01]
[3.55542219e-01]	[-1.78876224e+01]	[-1.82431646e+01]
[2.15196262e-01]	[1.41696774e+02]	[1.41481578e+02]
[3.65296331e-01]	[5.82819385e+02]	[5.82454089e+02]
[3.12318027e-01]	[-2.34037511e+02]	[-2.34349829e+02]
[1.02934009e-01]	[-2.56071453e+02]	[-2.56174387e+02]
[3.43667073e-01]	[-3.85177401e+02]	[-3.85521068e+02]
[1.89456829e-01]	[-3.34176740e+01]	[-3.36071309e+01]

Weights of Ridge Regression	Weights of OLE Regression	Difference in both weights
[2.23174971e-01]	[-1.07350066e+01]	[-1.09581816e+01]
[2.69446900e-01]	[2.57107189e+02]	[2.56837742e+02]
[-3.49799395e-01]	[5.99554593e+01]	[6.03052587e+01]
[-3.05637032e-01]	[3.83728042e+02]	[3.84033679e+02]
[-1.52149755e-01]	[-4.04158390e+02]	[-4.04006240e+02]
[1.09819882e-01]	[-5.14286434e+02]	[-5.14396254e+02]
[-7.18151056e-02]	[3.83636642e+01]	[3.84354793e+01]
[5.84103580e-01]	[-4.46102889e+01]	[-4.51943925e+01]
[-2.95218676e-01]	[-7.29643531e+02]	[-7.29348312e+02]
[-3.18856715e-01]	[3.77408336e+02]	[3.77727193e+02]
[1.10301700e-02]	[4.39794290e+02]	[4.39783260e+02]
[-1.16614485e-01]	[3.08514373e+02]	[3.08630988e+02]
[1.98283755e-02]	[1.89859679e+02]	[1.89839850e+02]
[3.33512788e-01]	[-1.09773797e+02]	[-1.10107310e+02]
[-1.18262935e-01]	[-1.91965699e+03]	[-1.91953873e+03]
[3.21734551e-01]	[-1.92463378e+03]	[-1.92495551e+03]
[-2.44030388e-01]	[-3.48979528e+03]	[-3.48955125e+03]
[-4.60907260e-01]	[1.17969687e+04]	[1.17974296e+04]
[8.32253571e-02]	[5.30674415e+02]	[5.30591189e+02]
[3.05867095e-01]	[5.43305906e+02]	[5.43000039e+02]
[-7.59713153e-02]	[1.82107518e+03]	[1.82115115e+03]
[-5.28346599e-01]	[-1.04639807e+04]	[-1.04634523e+04]
[9.99524871e-02]	[-5.16627611e+02]	[-5.16727563e+02]
[-1.82863265e-01]	[2.06435917e+03]	[2.06454204e+03]
[2.13189235e-01]	[-4.19941335e+03]	[-4.19962654e+03]
[-6.07451351e-01]	[-1.40495705e+02]	[-1.39888254e+02]
[-2.88586416e-01]	[3.74157090e+02]	[3.74445676e+02]
[6.84430343e-01]	[5.14757492e+01]	[5.07913188e+01]
[4.00670103e-01]]	[-4.64492730e+01]]	[-4.68499431e+01]]

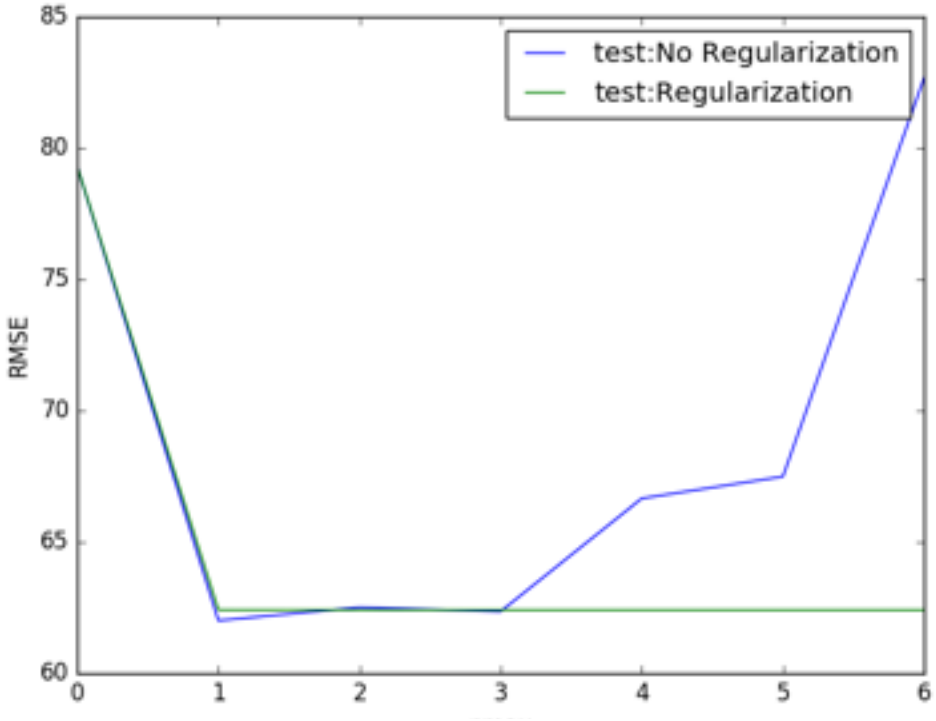
Problem 4:

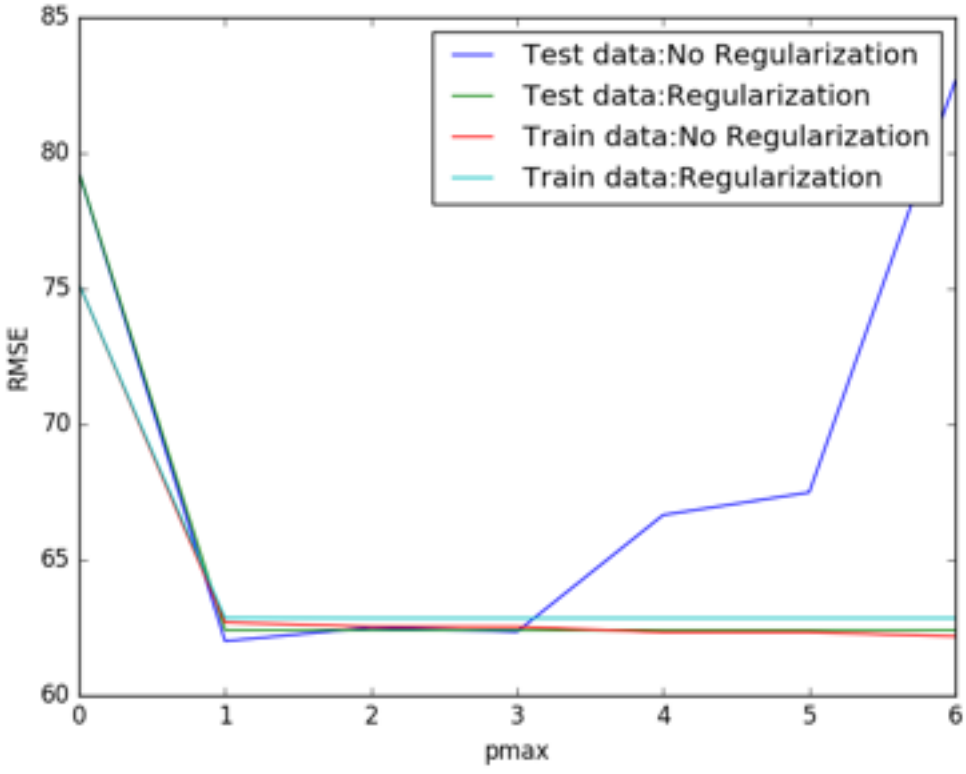
The following graph shows error on training and test data based on ridge regression with gradient descent by varying lambda values. As we can see the graph obtained in Problem 4 is very similar to that obtained in Problem 3. Also, the difference in lambda values is negligible. We can observe in the resultant graph that the default regularization parameter is inline with the optimal value for lambda. Different values of the parameter result in graphs that are not as optimal as the graph below.



Problem 5:

Following graph shows errors using Non-linear regression with and without regularization. For no regularization, optimal values of p are upto 3, then the error increases for higher values of p. Whereas after using regularization parameter, error remains same after p = 3 too (i.e. it's same for p=1,2,3,4,5,6). Thus optimal value of p is 6.





Problem 6:

The selection of the best regression model to implement on a given dataset can never be considered as a static methodology. The sheer possibilities in the behavior and nature of the provided datasets and their characteristics lead us to conclude that a model which may work perfectly for one problem will not be so impressive when applied to another dataset. The ultimate goal should be concentrated on finding the best fit model for a given dataset. However complicated the problem, the solution sometimes may lie in a very simple model that captures the most essential features or metrics of the dataset.

For the given dataset of the diabetes data sample, we try our hands at different regression models as specified by the problems above. Though we can cannot say with conclusive evidence that one particular method is superior to the others, we can notice the difference in the performance of the models experimented with. We notice that the results provided by Ridge Regression for the diabetes problem shows a slightly higher accuracy level as compared to the other models as the error is not that high. The application of gradient descent in ridge regression is close in terms of performance with ridge regression but for this problem, ridge regression has a lesser error. Non linear regression method generates a higher error for the given data set but linear regression models take a higher time to process the data with respect to lambda value and other metrics. Hence, we can safely assume that non linear regression model is not an appropriate fit for the diabetes dataset as it is directly affecting the lives of people in the real world. This model may be used to develop an overall general idea and behavior about the dataset in a time constrained environment as compared to the other models, making it a good fit for a different category of datasets which concentrate on a particular aspect of the data.

The metric that can be considered as the best setting for this project is the RMSE component which gives us a simplified view of the overall performance of the models. We can compare the RMSE value generated by each of the model to determine the deviation of the results from the ideal results.