CSE 574 PA2- Classification & Regression

By

Ishan Bhatt-50134076
Pratik Chavan-50134114
Daniel Nazareth-50134215

PROJECT SUMMARY

- The project aims to implement two supervised machine learning techniques namely classification (where the output variable is assigned class labels, essentially a form of “group membership”) and regression (where the output is a continuous variable).
- We use the SciPy Python toolkit to practically examine a variety of classification and regression techniques on two pre-provided sets of training and testing data.

PROBLEM 1-LINEAR/QUADRATIC DISCRIMINANT ANALYSIS

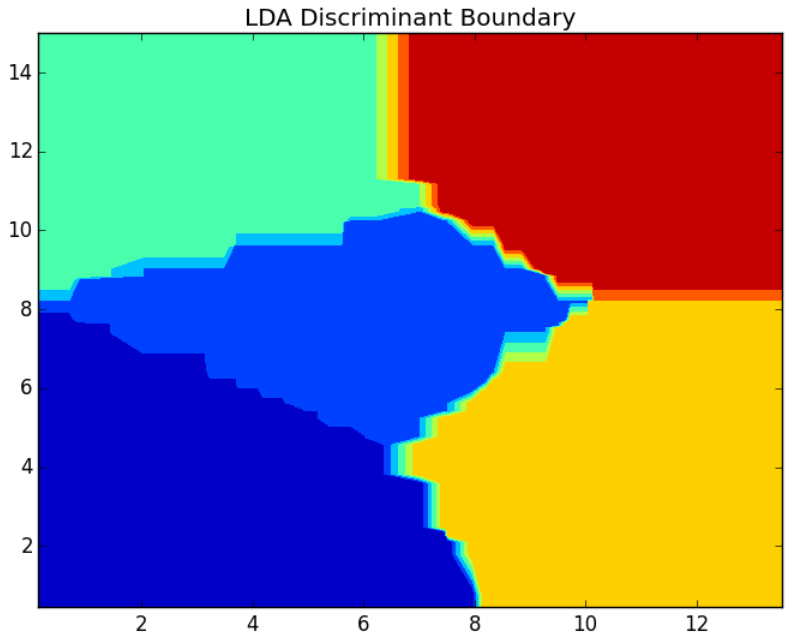
- Discriminant function analysis is used to examine a set of data examples and the groups they belong to in order to establish the discriminating factors that determine group membership.
- Given sufficient training data examples, a linear or quadratic discriminant function (specifically ldaLearn and qdaLearn in our code) is able to then accurately predict the group that a new data example should belong to(implemented in the ldaTest and qdaTest testing functions.)

TESTING RESULTS:

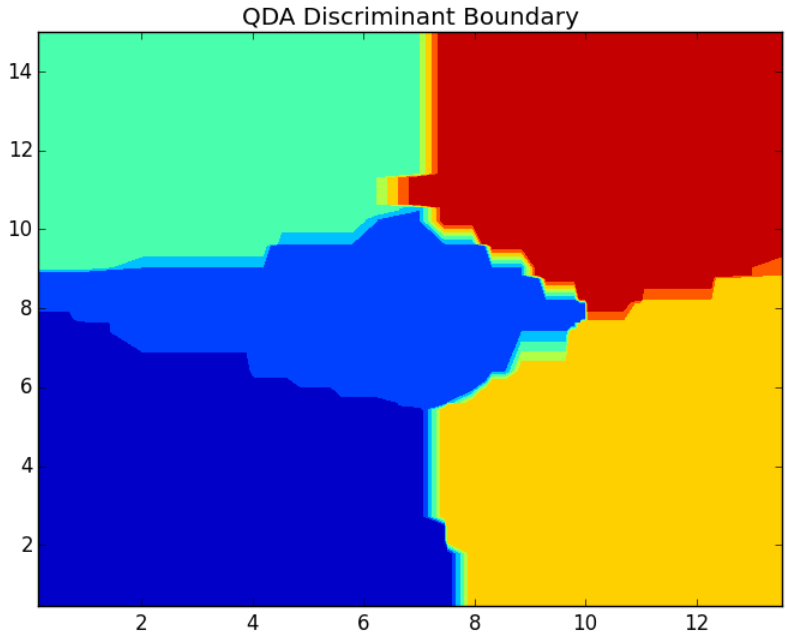
TRAINING FUNCTION	ACCURACY
Linear Discriminant	97.0 %
Quadratic Discriminant	97.0%

DISCRIMINANT BOUNDARIES:

- ***LDA BOUNDARY***



QDA BOUNDARY:



The clear observed difference in the discriminant boundaries is due to the fact that we assume a single, identical covariance matrix 'covmat' for LDA whereas we use different covariance matrices for each label.

PROBLEM 2-LINEAR REGRESSION

- Here we estimate linear regression parameters using the least squares method(code function learnOLERegression) and then use this to calculate the RMSE(Root Mean Squared Error) in training and test data prediction. We calculate the RMSE both with and without prior bias factored in.

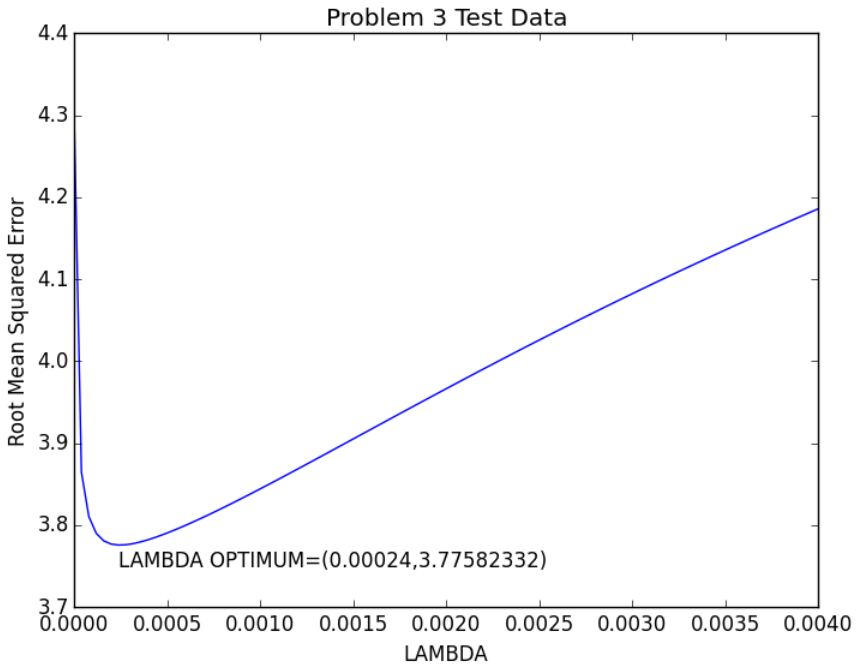
RESULTS:

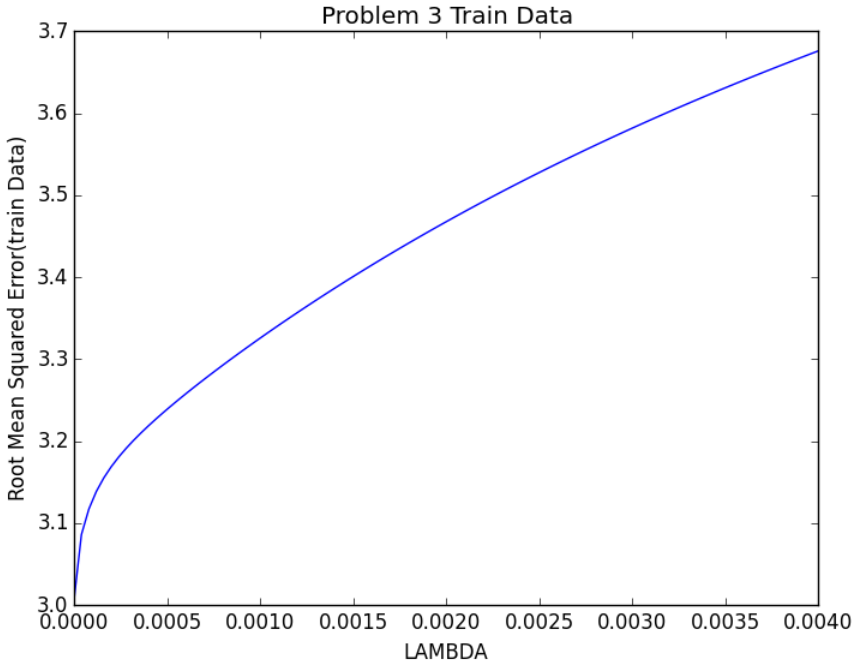
RMSE without intercept	23.10577
RMSE with intercept	4.30571

INTERPRETATION: We observe that the RMSE with intercept is considerably lower and therefore better than without intercept. This is because linear regression in general is highly susceptible to outliers. Using an intercept reduces the impact of outliers on the fitted line.

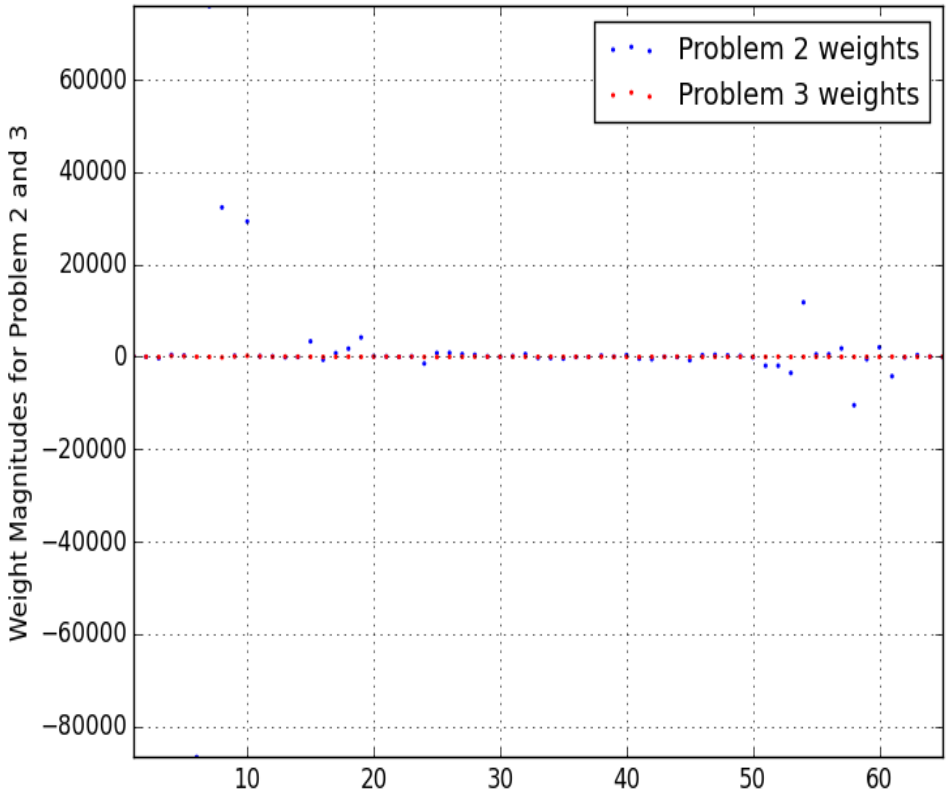
PROBLEM 3-RIDGE REGRESSION

- This problem implements a simple enhancement to linear regression wherein we calculate the RMSE in training and test data prediction using a better fitting set of parameters called *ridge regression parameters.* These parameters are calibrated in the function learnRidgeRegression and then, as in Problem 2, used to estimate the RMSE.
- The graph of lambda versus the RMSE is indicated below. As indicated in the graph, the optimum value of lambda is 0.00024 and corresponding minimum RMSE is 3.77582332.





WEIGHT MAGNITUDE COMPARISON : A comparison of the weight magnitudes for Problem 2 and 3 is shown below in the form of a scatter chart. **Clearly we can see that using ridge regression parameters as in problem 3 leads to a better fitting and more regularized set of weights.**

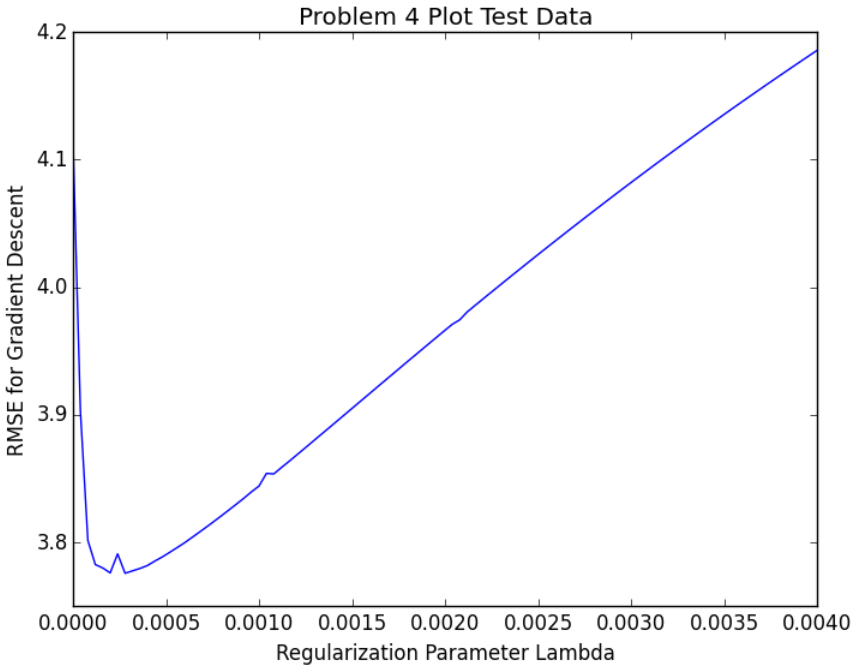


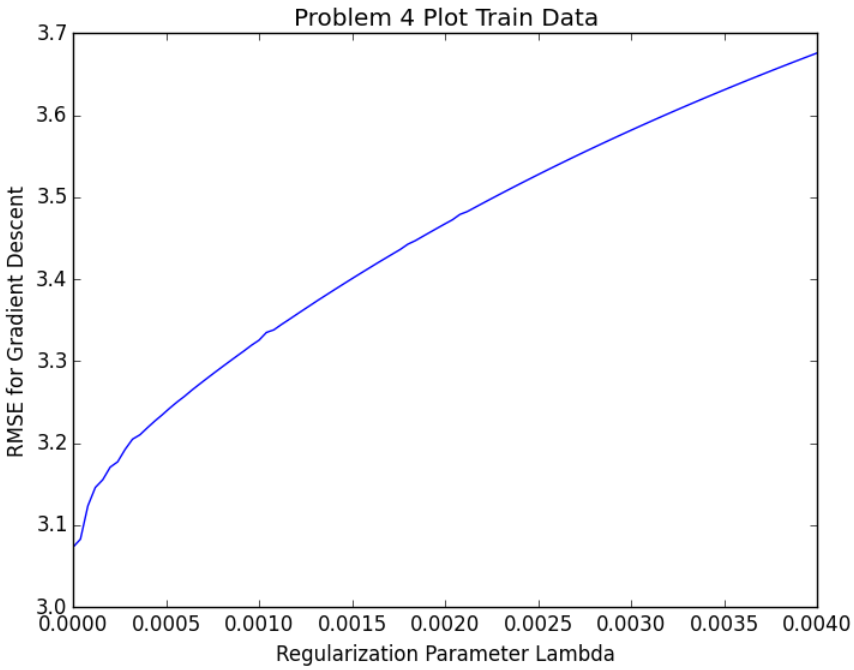
ERROR COMPARISON: Clearly we see that ridge regression leads to a lower RMSE as compared to linear. In fact if we select Lambda=Lambda_Opt=0.00024, as in Problem 3, we can obtain a significantly lower RMSE.

PROBLEM 4: RIDGE REGRESSION W/ GRADIENT DESCENT

- In problems 2 and 3, regression parameters were calculated using analytical expressions. In Problem 4 we minimize the loss using the gradient descent procedure (with the help of the **minimize** function from the SciPy library).

- The variation in regularization parameter versus the RMSE using this approach is shown below graphically.

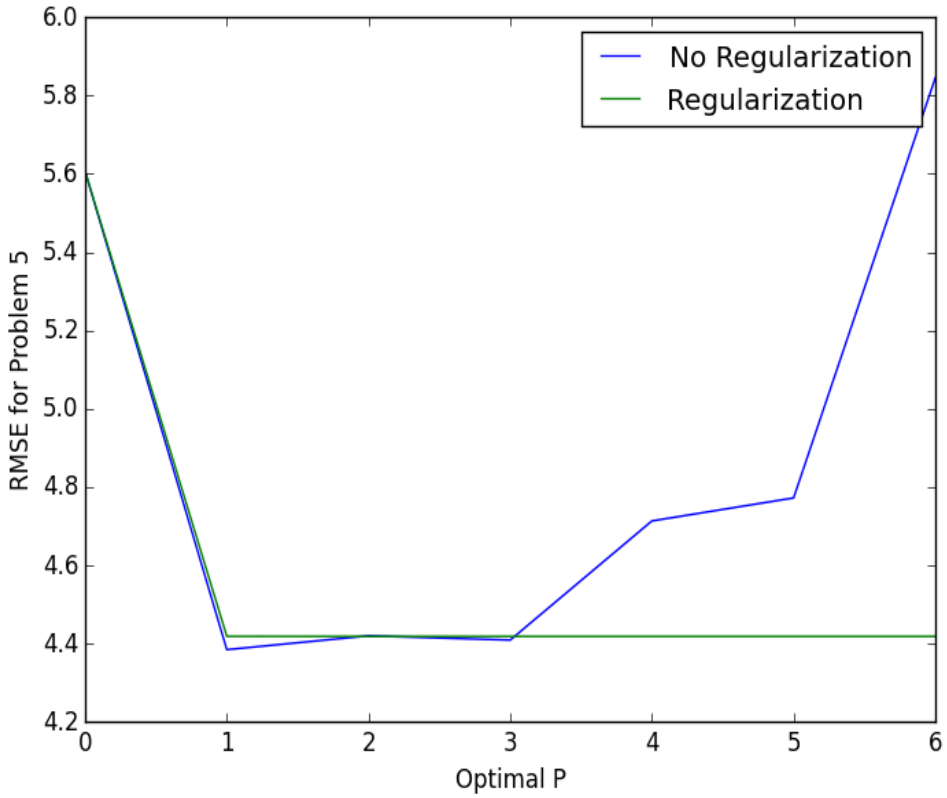




COMPARISON WITH PROBLEM 3: We see that the minimum RMSE obtained is nearly identical to that of Problem 3, in fact it is exactly equal to RMSE 3 upto 3 decimal places in the case of the test data. **For our run, we observed $RMSE4_{minimum}$=3.77584533 .** In the case of the training data, the difference is a bit more apparent, minimum occurring at zero for Problem 3 and at roughly 3.08 for Problem 4.

PROBLEM 5: NON LINEAR REGRESSION

- Problem 2 described the input features using a linear function for fitting. In Problem 5, we use higher order polynomials to describe input features and calculate RMSE. We also utilize the optimum value of lambda from Problem 3 to train our ridge regression weights.
- We plot two graphs for this problem, the RMSE with and without the regularization parameter. The graph is indicated below. As can be seen, the value of RMSE cannot be brought as low as in Problems 3 and 4.



Optimal Value of P: As can be seen, value of RMSE is nearly constant and takes its lowest value between P=2 and P=3. Therefore we can choose P=2 as the optimal value.

PROBLEM 6: INTERPRETING RESULTS

We summarize the results for all five problems in tabular form as below:

CLASSIFICATION/REGRESSION METHOD	MEASURE OF QUALITY	OPTIMUM MEASURE VALUE
Linear Discriminant Analysis	Prediction Accuracy	97.0%
Quadratic Discriminant Analysis	Prediction Accuracy	97.0%
Linear Regression	RMSE	4.30571
Ridge Regression	RMSE	3.77582332

Ridge Regression W/ Gradient Descent	RMSE	3.77584533
Non Linear Regression	RMSE	4.41854531

The following factors and metrics should be taken into account while choosing a method:

- **RMSE :** Lower the RMSE, the better it is for us in general, although it is possible for a low RMSE to be due to underfitting.
- **Nature Of Fit:** Certain models lend themselves to fitting the given data much better than others. For example a linear fit may be inferior to a ridge fit because it is unlikely that random data will have a very good linear relationship.
- **Computational Expense:** Method chosen should be as computationally inexpensive as possible.

Based on the above factors, Ridge Regression with gradient descent would be the preferred choice for learning and classifying the diabtes data set. This is because it offers a solid fit, pretty much the lowest RMSE and does not involve any matrix inversion which is computationally expensive to calculate.