Classification and Regression

By

Group 4

Hrishagni De (hrishagn@buffalo.edu)

Sanchari Gautam (sgautam2@buffalo.edu)

Sahana Mukherjee (mukherj3@buffalo.edu)

CSE4/574-Fall 2021: Introduction to Machine Learning

Programming Assignment 1

Problem 1: Experiment with Gaussian Discriminators:

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are trained on sample training data. Both methods are tested on sample test data and accuracy is reported below:

Accuracy for LDA: 97%

Accuracy for QDA: 96%

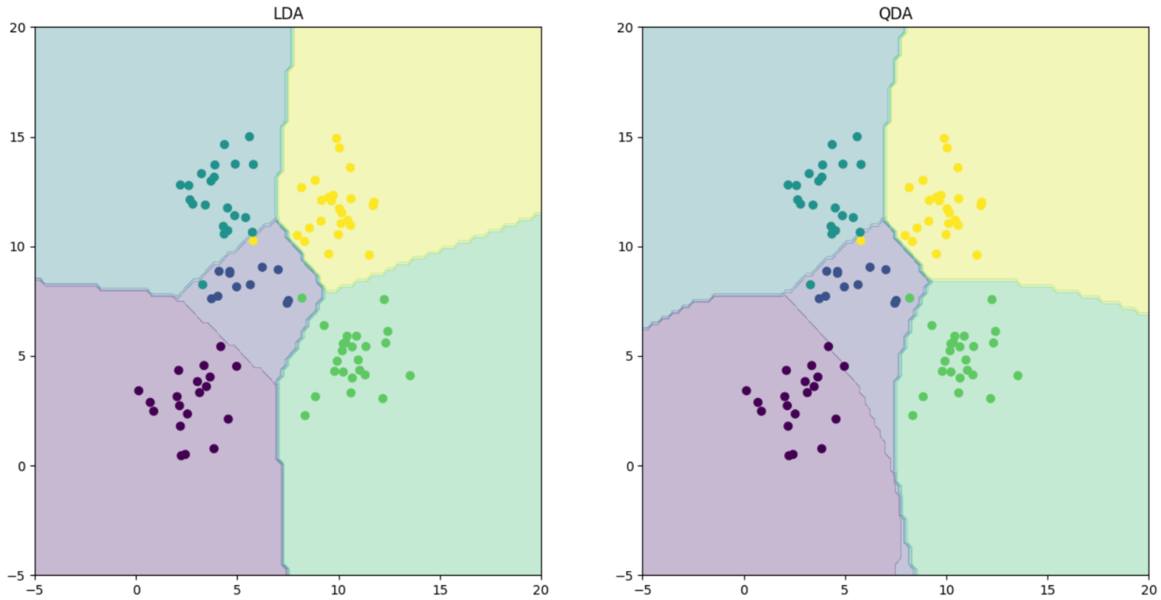


Figure 1: Plot of discriminating boundary for LDA and QDA discriminators

There is a difference between the two boundaries as plotted by LDA and QDA, because QDA computes separate covariance matrix for each class and LDA computes the same covariance for every class. For QDA, if number of classes is less than number of data points, it becomes a problem. Decision boundaries in QDA are quadratic in nature as the discriminant function is quadratic and will contain second order terms.

<u>Problem 2: Experiment with Linear Regression:</u>

Ordinary Least Square is implemented, and the mean squared error (MSE) is calculated and reported for training data and test data, both with and without incorporating the intercept terms.

MSE for Training and Test Data, without and with intercept:

	Without Intercept	**With Intercept**
Training Data	19099.44	2187.16
Test Data	106775.36	3707.84

From the above table, it is evident that MSE is **better** for both training and test data, **with intercept**, than MSE calculated without intercept. This is because, data points might not always pass through the origin and the intercept term folds a bias term in itself, which accounts for such cases when the data points do not pass through the origin.

Problem 3: Experiment with Ridge Regression

Ridge regression is performed on training and test data, and the MSE is calculated and reported below for both cases, using the intercept term.

MSE for Training and Test Data, with intercept:

	With Intercept
Training Data	2306.83
Test Data	2851.33

The errors on the train and test data for different values of **λ are** plotted. Lambda is varied from 0 to 1 in steps of 0.01.

The plot of errors on train and test for different values of λ:

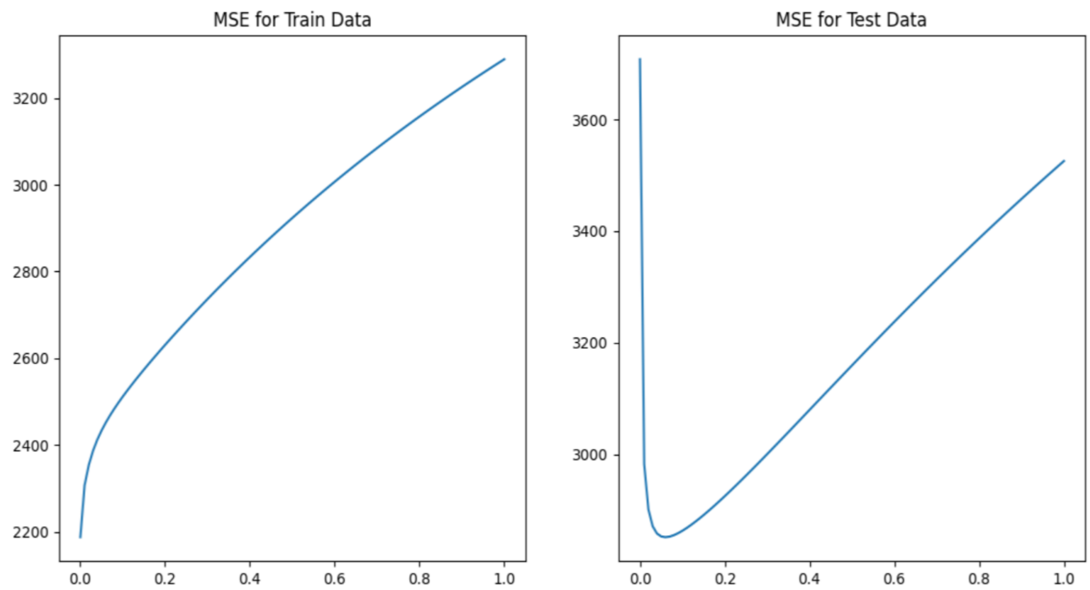


Figure 2: Plot of errors for different values of λ

Comparison of relative magnitudes of weights learned using OLE and Ridge Regression:

- Mean Weight Learnt for OLE: 882.81
- Mean Weight Learnt for Ridge Regression: 17.32

From the above data, it is observed that the mean weight learned during Ridge Regression is less than Mean Weight learned during OLE. This is because, Ridge regression tends to penalize the weights with higher coefficients, thus reducing overfitting.

Comparison of two approaches, OLE and Ridge Regression in terms of errors (MSE) for train and test data:

Model	**Train Data**	**Test Data**
OLE	2187.16	3707.84
Ridge Regression	2306.83	2851.33

From the above comparison of MSE, we can see that Ridge Regression performs better than Ordinary Least Square, in terms of an error on both train and test data.

Finding the optimum value of lambda for Ridge Regression:

Lambda	Train Data	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
0.1	2511.43	2862.76

The optimal value of lambda is 0.06 because from this point the errors again start to increase for test data. The test error is starts to decrease from 3707.84, when lambda = 0, and decreases till 2851.33, at lambda = 0.06. After this point, as lambda increases, test error again starts to increase.

Hence, we take the optimal value of lambda as 0.06, as at this point the test error gets minimum, before again increasing.

In the above table, we have reported the first 11 values of lambda.

Problem 4: Using Gradient Descent for Ridge Regression Learning

The errors on train and test data are plotted, using gradient descent-based learning, by varying the regularization parameter lambda of ridge regression.

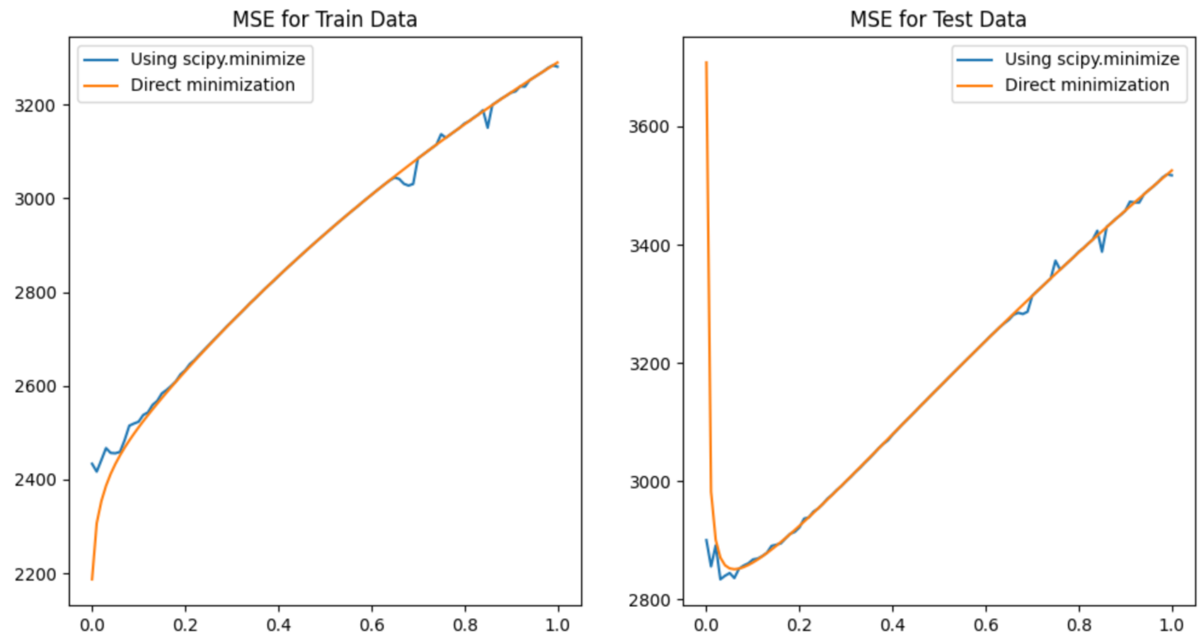


Figure 3: Plot of error using Gradient Descent by varying λ

The MSE obtained for test data, using gradient descent for Ridge Regression, is 2836.38 for λ = 0.06. It is observed that the MSE obtained from gradient descent with Ridge Regression is slightly less than the one obtained in Problem 3, which was 2851.33.

Problem 5: Non-linear Regression

In this problem, we investigate the impact of higher-order polynomials on the input features.

The plot of MSE vs p-value (p-value is varied from 0 to 6) is shown below.

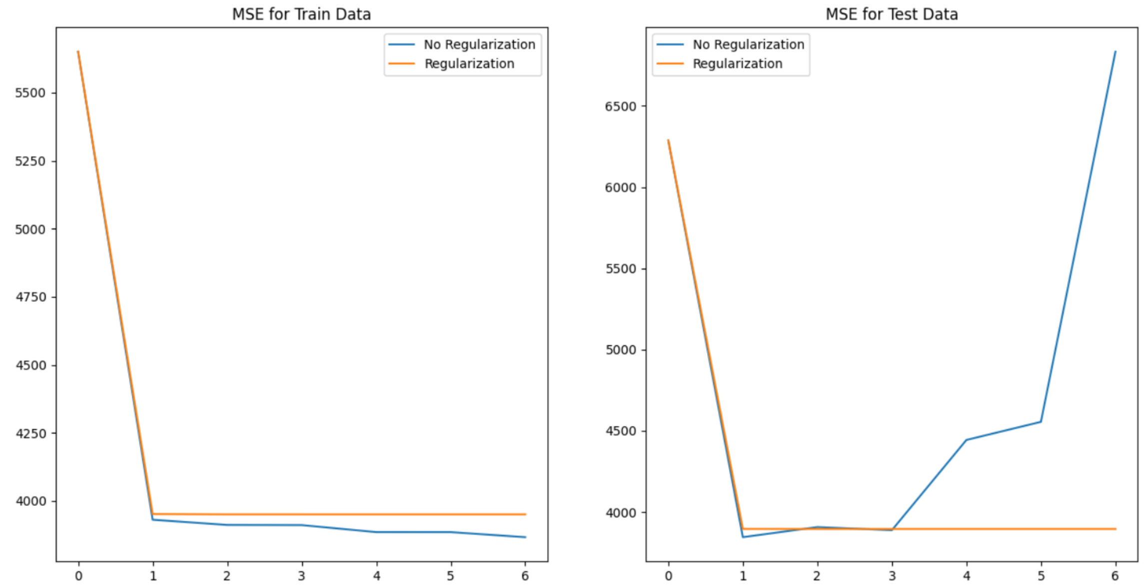


Figure 4: Comparison of p values on train and test data, against λ= 0 and the optimal value of λ

Table showing error on test data, with λ = 0 and λ= 0.06, by varying p from 0 to 6

p-Value	Test Error with Lambda = 0	Test Error with the 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
6	6833.45914872	3895.58266872

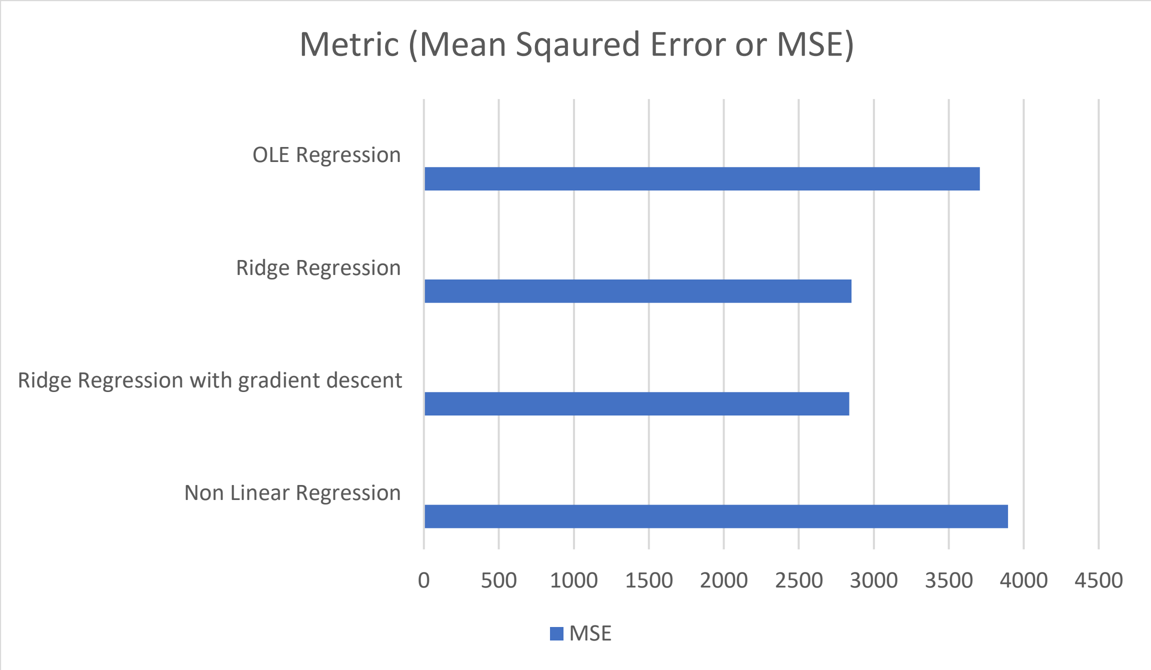
From the plot and the records in the table, it is seen that, **without regularization**, the error on the **test data** is **minimum** at **p=1**. As we increase the value of p, i.e., increase the order of the polynomials, the test error increases, as it leads to overfitting.

This scenario is dealt with by using **ridge regression (λ= 0.06),** when the test error reduces at **p=4**, and remains constant, even when we increase the value of p. This is because regularization tends to penalize the weights with higher coefficients and reduces overfitting.

Problem 6: Interpreting Results

The results obtained from the above techniques are summarized below:

	OLE Regression	Ridge Regression (λ = 0.06)	Ridge Regression with Gradient Descent (λ= 0.06)	Non-Linear Regression
Metric (Mean Squared Error)	3707.84	2851.33	2836.38	3895.85646447



From the above observations, it is evident that Ridge Regression with Gradient Descent is the best approach for performing regression to predict diabetes level using the given input features. It provides the least error on the testing data, (lowest MSE), among all other techniques.

For using this technique, the metric that is most suitable is the Mean Squared Error (MSE). We can also use the Root Mean Squared Error (RMSE) metric, which is simply obtained by computing the root of MSE. The MSE has the units squared and might be difficult to interpret. RMSE is more interpretable and easier to deal with.