CSE 574 Introduction to Machine Learning

Programming Assignment 2

Classification and Regression

Members of GROUP 43

1) Arundhati Ravikumar Rao
2) Cuurie Athiyaman
3) Prithvi Ganesan

Instructor: Prof.Varun Chandola

Problem 1: Experiment with Gaussian discriminators

Linear discriminant analysis (LDA) and Quadratic discriminant analysis (QDA) are linear and quadratic decision surface classifiers, discriminating boundaries between multiple classes.

We have trained the sample_train dataset by both LDA and QDA and then tested them with the given sample_test.
The accuracies observed were:

LDA - 97%
QDA-94%

From the figure for LDA below, we can see that almost all the data points are linearly separable for the given test dataset. In LDA, we assume the same covariance matrix for Gaussian distributors with multiple classes. LDA uses straight lines to separate the various classes and hence a higher accuracy is expected for linear separable dataset, which is observed in the problem.

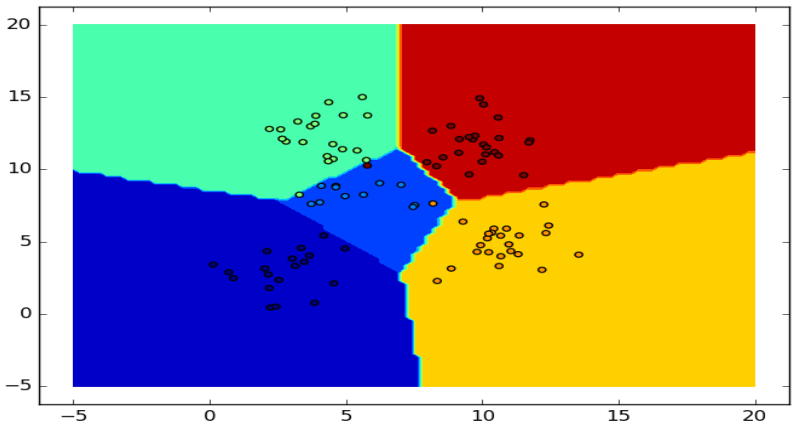

Figure 1

As opposed to as in LDA, for QDA we do not assume that the Gaussian distributions for all classes share the same covariance. And since each class has its own covariance matrix, the classifier is observed to be quadratic. QDA makes use of curved lines for separating or discriminating data points in a quadratic fashion.

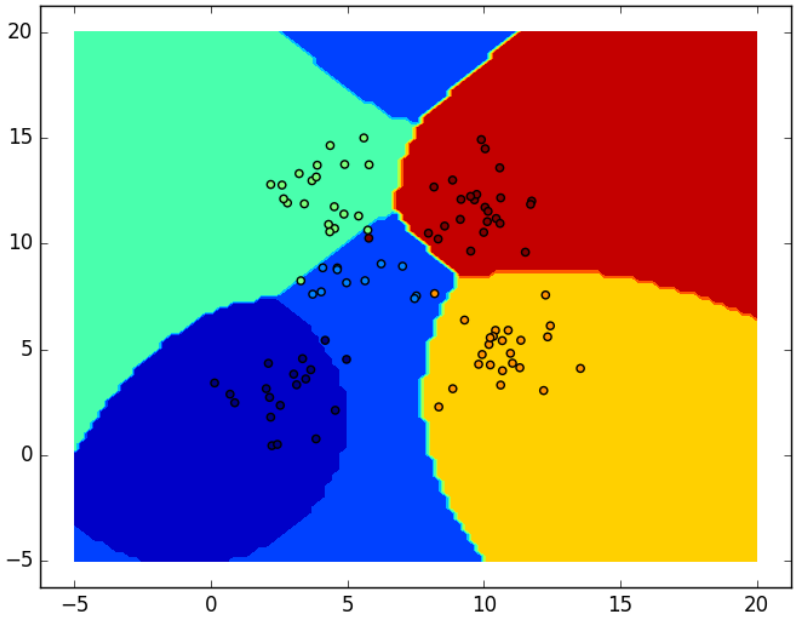

Figure 2

The Figure above shows QDA's discriminating boundaries for the same dataset as for LDA. The boundary's shape is determined in a quadratic fashion, resulting in curved lines. This is more suitable for real world situations where the data points do not fit into rigid linearly separable boundaries.

As seen from our observations, QDA offers lesser accuracy when compared to LDA.

Problem 2: Experiment with Linear Regression

Linear Regression is an approach which is used to model the relationship between a scalar dependent variable (y) and one or more independent variable(s) (X).

In this problem, we implemented the Linear Regression function, in which we learned the weight vector (w) using the training data matrix (X) and the corresponding labels (y).

The below squared loss function was minimized and used for learning the weights:

$$J = \frac{1}{2}(y - w^T X)^T (y - w^T X)$$

The weights were obtained by:

$$w = (X^T X)^{-1} X^T y$$

Data Type (Training / Testing)	Intercept Present?	RMSE Values
Training	Yes	46.76708559
Training	No	138.20074835
Testing	Yes	60.89203717
Testing	No	326.76499438

Table 1

The Root Mean Squared Error (RMSE) calculated for the two cases are tabulated as above (Table 1).

RMSE is similar to Standard Deviation .In regression, it calculates the deviation of the actual y-values from the regression line.

From the above table, we can see that that the RMSE calculated with the intercept, for both training and testing, is lesser than without the intercept. Hence, the OLE with the intercept is better than without intercept, for both training and testing data. This is as expected since if without intercept, the hypothesis line must pass through the origin point, allowing the line to only rotate. Only when an intercept is introduced does it give the line freedom to rotate as well as translate. Hence, the learnt hypothesis is closer to the actual concept, which leads to lesser errors when using an intercept.

Hence, errors are lesser when we use intercepts.

Problem 3: Experiment with Ridge Regression

Issues associated with Linear Regression are as follows:

- Linear Regression is susceptible to outliers.
- Linear Regression is too simplistic- which leads to the problem of under fitting.

On the other hand, Ridge Regression or Tikhonov Regression helps in reducing the impact of correlated inputs.

The RMSE values for training and test data using Ridge Regression parameters using the *testOLERegression* function are as follows:

SR. NO.	TEST	TRAIN	SR.NO.	TEST	TRAIN	SR.NO.	TEST	TRAIN
1	60.8920371	46.7670856	35	55.060087	52.6903531	69	57.4311889	55.4059171
2	54.6117764	48.0294932	36	55.1320714	52.782826	70	57.4966178	55.4743757
3	53.8606868	48.5187731	37	55.2041063	52.8744132	71	57.5617462	55.5422851
4	53.5811682	48.8546842	38	55.276163	52.9651283	72	57.626573	55.6096524
5	53.4602695	49.1133286	39	55.3482151	53.0549847	73	57.6910971	55.676485
6	53.4103523	49.327218	40	55.4202379	53.1439949	74	57.7553178	55.7427899
7	53.3978484	49.5129124	41	55.4922088	53.2321713	75	57.8192343	55.8085741
8	53.4073964	49.6797499	42	55.5641066	53.319526	76	57.8828462	55.8738445
9	53.4310747	49.8333788	43	55.6359119	53.4060707	77	57.946153	55.9386078
10	53.464422	49.9773975	44	55.7076065	53.4918169	78	58.0091547	56.0028705
11	53.5047469	50.1141924	45	55.7791737	53.5767758	79	58.0718512	56.0666393
12	53.5503306	50.2453982	46	55.8505979	53.6609585	80	58.1342424	56.1299205
13	53.6000213	50.3721639	47	55.9218648	53.7443759	81	58.1963286	56.1927205
14	53.6530136	50.4953155	48	55.992961	53.8270385	82	58.2581101	56.2550453
15	53.7087229	50.6154574	49	56.0638742	53.9089571	83	58.3195872	56.316901
16	53.7667101	50.7330394	50	56.134593	53.9901418	84	58.3807604	56.3782937
17	53.8266352	50.8484009	51	56.2051069	54.0706029	85	58.4416303	56.4392293
18	53.8882281	50.9618013	52	56.275406	54.1503504	86	58.5021976	56.4997134
19	53.9512685	51.0734412	53	56.3454816	54.2293942	87	58.562463	56.5597517
20	54.0155734	51.1834777	54	56.4153251	54.3077441	88	58.6224272	56.6193499
21	54.0809876	51.2920347	55	56.4849291	54.3854096	89	58.6820912	56.6785134
22	54.147378	51.3992115	56	56.5542865	54.4624003	90	58.741456	56.7372475
23	54.2146284	51.5050884	57	56.6233908	54.5387254	91	58.8005224	56.7955577
24	54.2826365	51.6097307	58	56.6922361	54.6143941	92	58.8592915	56.853449
25	54.3513115	51.7131927	59	56.760817	54.6894154	93	58.9177645	56.9109267
26	54.4205719	51.8155197	60	56.8291286	54.7637984	94	58.9759425	56.9679956
27	54.4903444	51.91675	61	56.8971662	54.8375516	95	59.0338267	57.0246609
28	54.5605627	52.0169168	62	56.9649259	54.9106838	96	59.0914183	57.0809272
29	54.6311662	52.1160485	63	57.0324038	54.9832035	97	59.1487186	57.1367995
30	54.7020998	52.2141704	64	57.0995967	55.055119	98	59.2057288	57.1922824
31	54.7733129	52.311305	65	57.1665014	55.1264385	99	59.2624504	57.2473806
32	54.8447592	52.4074727	66	57.2331153	55.1971702	100	59.3188847	57.3020985
33	54.916396	52.502692	67	57.2994359	55.267322	101	59.3750331	57.3564406
34	54.9881841	52.5969801	68	57.3654611	55.3369017			

For this problem, the weight vector (w) is learnt from the given dataset (training data matrix – x & corresponding labels – y) by implementing the *learnRidgeRegression* function.

The impact of outliers on the weights is controlled by the regularization factor – λ. This ultimately helps in avoiding overfitting.

The regularized square loss function:

$$J = \frac{1}{2N}(y - Xw)^T(y - Xw) + \frac{1}{2}\lambda w^T w$$

And the weight w is obtained by:

$$w = (X^T X + \lambda N I)^{-1} + X^T Y$$

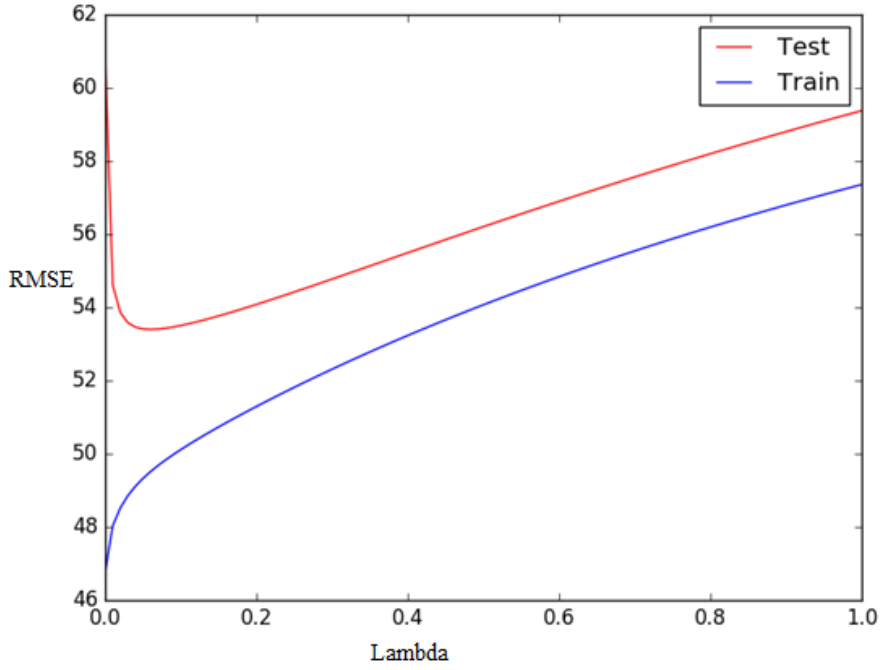

Figure 3

Observations: -

From the graph above, it's clear that by varying the λ value, the RMSE value changes too. RMSE decreases steeply with increasing λ value. Initially, the λ value being zero, causes the second term containing λ become zero in the objective function above. But as the λ value increases from 0, the term plays a part in the minimization and emphasizes how complex the model can get. λ value controls the weight values leading to lower RMSE. But as λ value goes beyond 0.0599, the RMSE begins increasing which shows that the λ puts a tighter restriction on the weight vectors from growing leading to increased errors.

Comparison of Relative Magnitudes of Weights:-

Its observed that OLE tends to vary the weights more sharply than ridge regression. Relative to weights learnt using Ridge Regression, the weights learnt using OLE are higher.

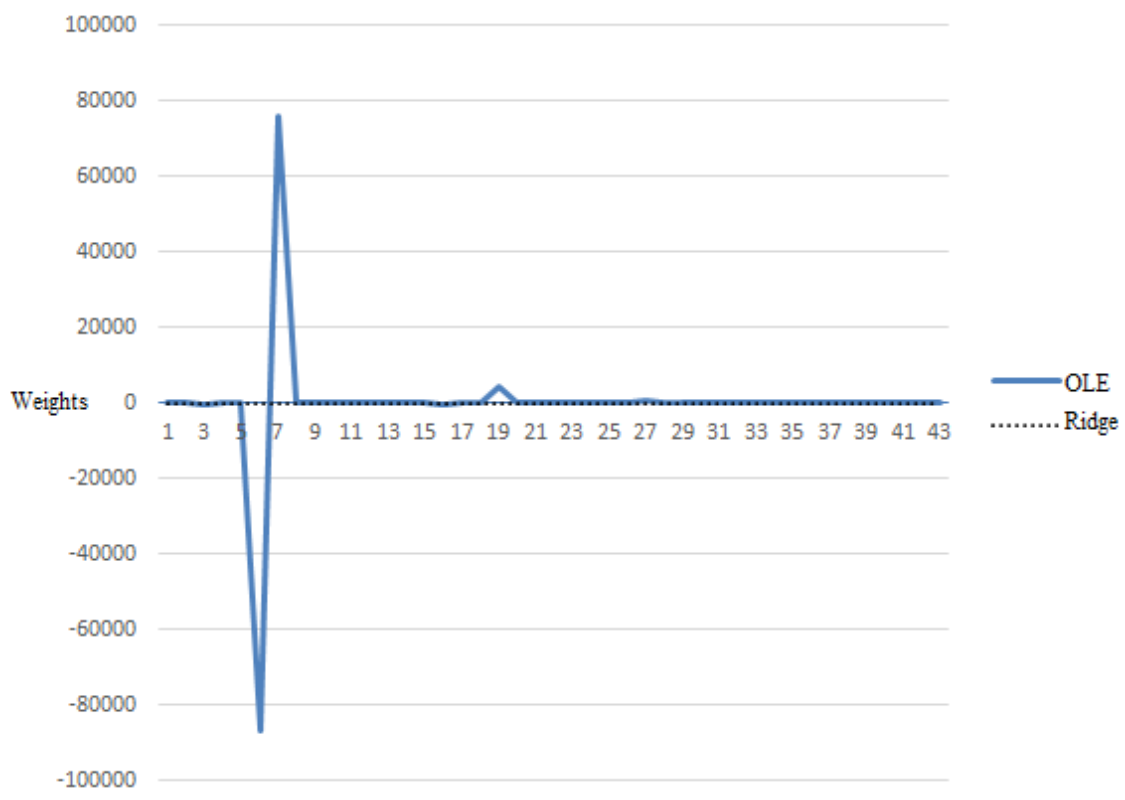

Figure 4

Calculating the mean error from Linear Regression and Ridge Regression, we get the following table:

Training/Testing Data	**Mean RMSE from OLE**	**Mean RMSE from Ridge Regression**
Training Data	46.76708559	56.2805719779
Testing Data	60.89203717	53.7038339189

As seen from the table, Linear Regression is seen to perform better in the learning phase. However, the mean error for the testing data. This is an example of the previously mentioned shortcomings of Linear Regression. Ridge Regression, on the other hand, performs better than Linear Regression with a mean error of 53.7038339189 on test data.

Finally, the Optimal value of λ calculated from this problem is as follows:-

Optimum λ value was observed at **0.0599** with the lowest **RMSE** error of **53.3979**.

As mentioned before, the first term in the equation for error is for error minimization and the second term controls the emphasize on weights. The reason why error is the lowest when λ is 0.0599 is because at that value both the terms in the equation for error are balanced.

$$J = \frac{1}{2N}(y - Xw)^T(y - Xw) + \frac{1}{2}\lambda w^T w$$

Problem 4: Using Gradient Descent for Ridge Regression Learning

In *regressionObjVal* function, given data (X) and corresponding labels (y), and regularization parameter (λ), we calculated squared error value using the equation for regularized squared loss function as below:

$$J = \frac{1}{2N}(y - Xw)^T(y - Xw) + \frac{1}{2}\lambda w^T w$$

Then, the gradient of squared error with respect to w is given by the formula:

$$\frac{dJ}{dw} = \frac{1}{N}(w^T(X^T X) - y^T X) + \lambda w^T$$

The error value and its gradient are then used to minimize the weights for Ridge Regression.

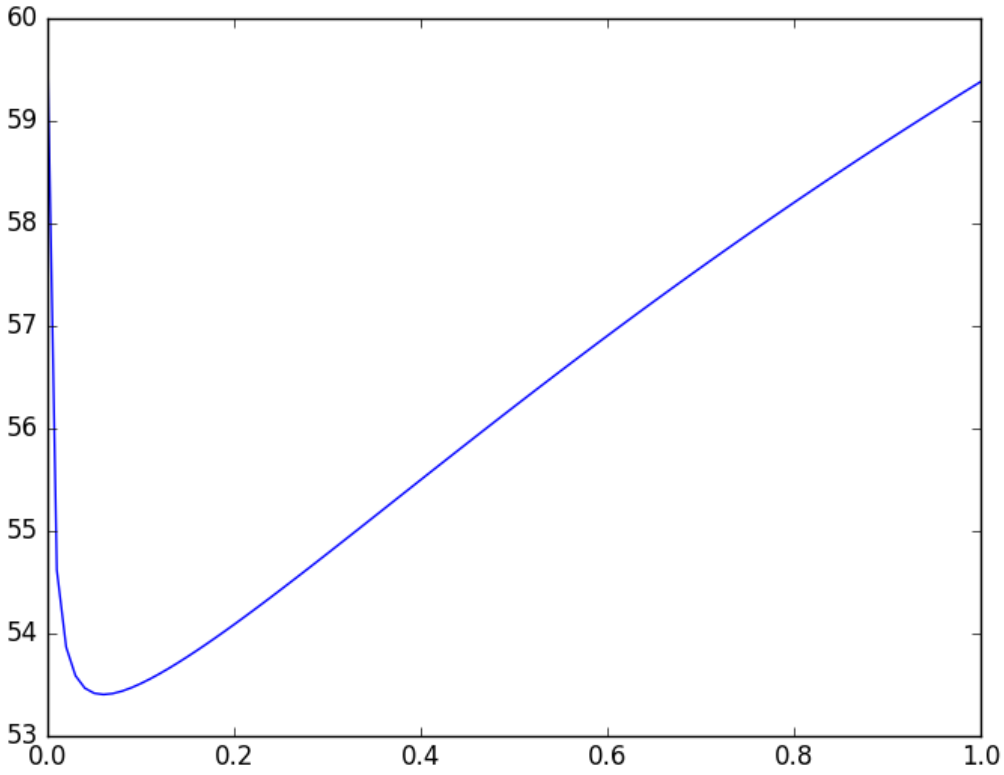

Figure 5 – RMSE versus λ for Test data

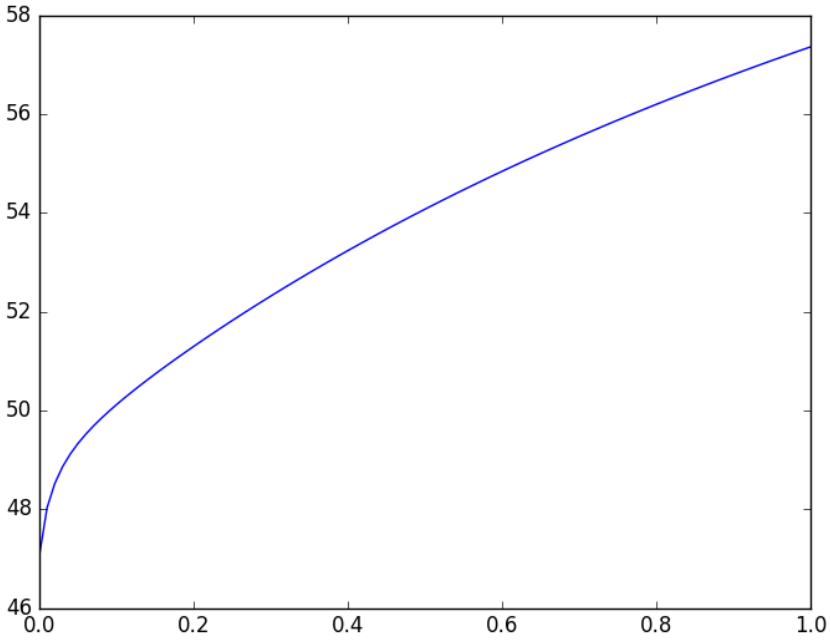

Figure 6 – RMSE versus λ for Training data

Similar to problem 3, RMSE initially decreases steeply with increase in λ value. Due to under fitting, when the λ value increases, error also increases.

Using the gradient descent method provides satisfactory results when the number of features becomes large or in cases where finding the inverse of the covariance matrix is not possible.

On comparing the graph above for training and testing data for Ridge Regression using Gradient descent with that of simple Ridge Regression, we can infer that the error value (RMSE) doesn't change much between both methods. The conclusion is that by observing both the graphs, the optimal lambda values are approximately same for both the methods. i.e.

λ = 0.0599 for simple Ridge Regression and λ = 0.0499 for Ridge Regression using gradient descent.

Problem 5: Non-linear Regression

Non-linear mapping, in this case polynomial expansion, is done to get non-linear curves.
In this problem, using non-linear mapping we have trained the weights for ridge regression in the *non-linear regression* function. We have used optimal λ value (0.0599) and λ as zero. The function was implemented so that we can convert a single attribute vector (N*1) into a p attributed (N*p+1) matrix.
The learning phase is similar to that used in problem 3, except that here we will use the (N*p+1) matrix. Then the prediction for each data point is calculated with the weights obtained from learning function. The graph below shows the RMSE values for training and testing data with and without regularization.

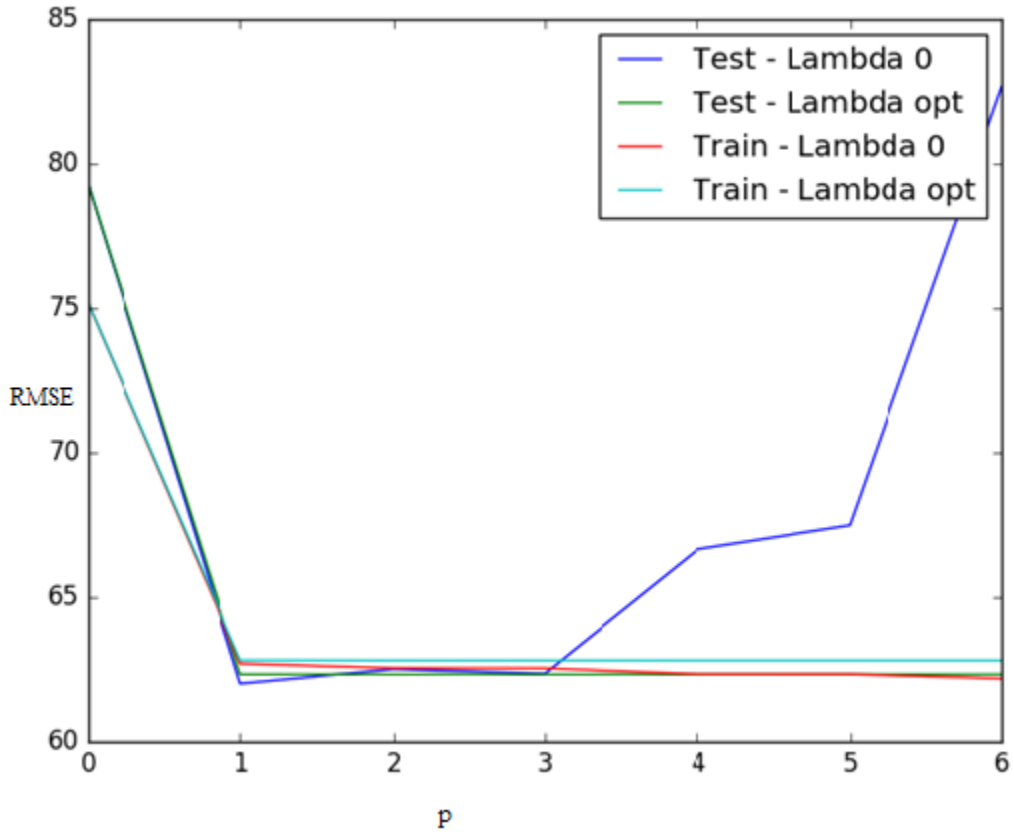

Figure 7

Observations: -

When p=0, RMSE value is highest as no learning will be done on the training data. Further, the RMSE value begins decreasing after p=0, as is expected because of being trained by ridge regression.

When p=1 and λ as 0, RMSE increases steeply because there is no regularization factor, causing overfitting of data. This overfitting can be proven because, when p=1 and λ is optimal, RMSE

value doesn't increase. This is because regularization parameter (λ) takes care of the overfitting problem. RMSE for training data continues to be low, as the training was done on the data.

The RMSE values for both training and testing data with λ as zero and optimal value with different p values are tabulated in the table given below:

P	**Train when λ = 0**	**Test when λ = 0**	**p**	**Train when λ is opt**	**Test when λ is opt**
0	75.17120818	79.28685132	0	75.1712145	79.28935782
1	62.69701275	62.00834404	1	62.81715398	62.32732645
2	62.54470138	62.5070244	2	62.80666908	62.32612357
3	**62.53949684**	**62.35363292**	3	62.80663223	62.32613129
4	62.33356293	66.658292	**4**	**62.80663049**	**62.32613124**
5	62.33303424	67.48948346	5	62.80663047	62.32613125
6	62.18427011	82.66473945	6	62.80663047	62.32613125

Optimum values of p for testing are observed as:

p=3 for RMSE = 62.35363292 with no regularization, (i.e.) λ value zero
p=4 for RMSE = 62.32613124 with regularization, (i.e.) λ is optimal value

Problem 6: Interpreting Results

The factors to be considered while trying to decide the best method would be

- Nature of the data set itself,
- Accuracy
- Stability and
- Performance.

Nature of the data set refers to the volume and variety of the dataset and the number of features in it. Accuracy is the measure of correctness of the predictions made by the learning model. Stability and Performance is the behaviour of the model when new data points are introduced.

For this project, the diabetes dataset has only 242 training examples and 200 examples in test data which is not very large. So we consider Accuracy as an important factor to be taken into consideration when trying to compare the different methods. For this purpose, we can consider the RMSE values (Error Values) which is the mean squared error.

In the table below, the RMSE values have been calculated using optimal λ value.

Method Used	RMSE - Training	RMSE - Testing
OLE	46.76708559	60.89203717
Ridge Regression	53.7038339189	56.2805719779
Ridge Regression using Gradient	53.646550611	56.4036097492
Non Linear Regression	62.80663049	62.32613124

On comparing the accuracies by considering the RMSE values for the above methods, **Ridge Regression** and **Ridge Regression using Gradient Descent** give minimal error for testing data than Non-linear and OLE methods.

But it could happen that the inverse of covariance matrix becomes unstable for large number of features, when we use derivative optimization technique. Also finding inverse of single covariance matrix is not possible. For these situations, **Gradient Descent Ridge Regression** method should be chosen over Ridge Regression method.