Report for Assignment 1

Vishesh Singh Thakur
50322513, vthakur@buffalo.edu

Ashwin Nair
50321006, anair3@buffalo.edu

Divya Gawande
50340326, divyagaw@buffalo.edu

Group 2
CSE 574 Introduction to Machine Learning
October 2020

Abstract

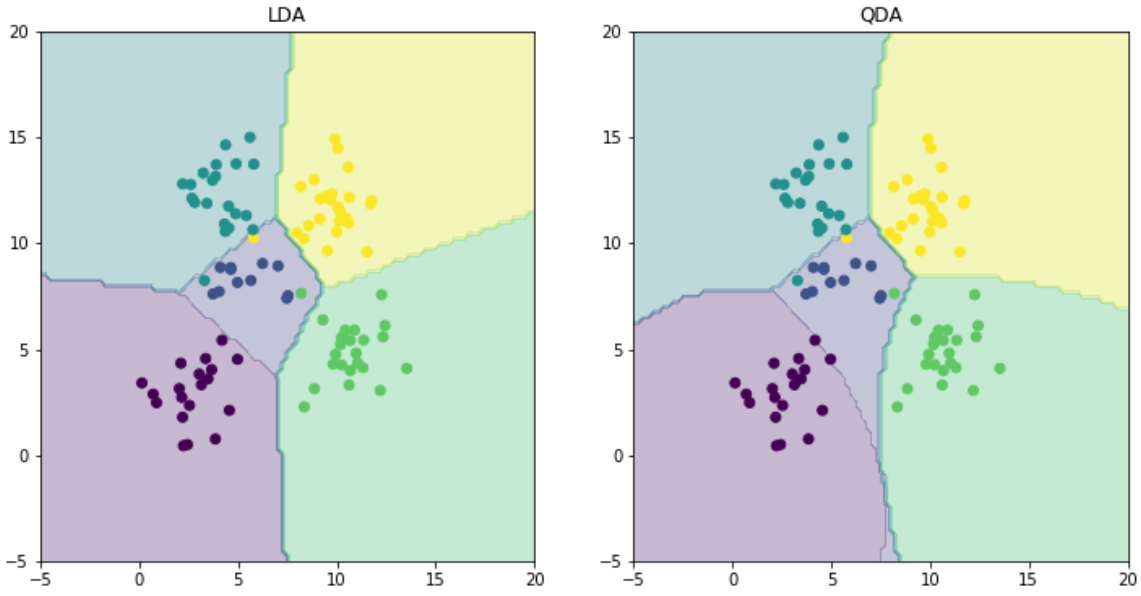
This assignment deals with the comparison of different predictive approaches, namely Linear Discriminant Analysis, Quadratic Discriminant Analysis, Linear Regression, Ridge Regression or Linear Regression with L2 regularization, Gradient Descent method, and Non-linear with regularization and without regularization. The comparison is done with respect to the MSE values obtained for the test data. The graphs show how each approach varies in terms of MSE for both training and testing data

Problem 1.

After training both the methods using the sample training data, we measured the accuracy of Linear Discriminant Analysis and Quadratic Discriminant Analysis on the test data set. The accuracy we achieved for **LDA is 0.97**, and for **QDA is 0.96.**

LDA assumes that the observations within each class are drawn from a multivariate Gaussian distribution and the covariance of the predictor variables are common across all *k* levels of the response variable *Y*. Quadratic discriminant analysis (QDA) provides an alternative approach. Like LDA, the QDA classifier assumes that the observations from each class of *Y* are drawn from a Gaussian distribution. However, unlike LDA, QDA assumes that each class has its own covariance matrix. In other words, the predictor variables are not assumed to have common variance across each of the *k* levels in *Y*.

Consider the image below. It is trying to classify the observations into the five(color-coded) classes, LDA (left plot) provides linear decision boundaries that are based on the assumption that the observations vary consistently across all classes, and hence is able to capture very accurate linear decision boundaries. While, QDA (right plot) is able to capture the differing covariances and provide more accurate non-linear classification decision boundaries. (Github)



The following formula is used for the LDA and QDA, where the mean remains the same, and while the covariance is fixed for Linear approach, it changes with every iteration for Quadratic (Reference: Class notes)

$$p(y|\mathbf{x}) \propto p(y)\prod_j p(x_j|y) = p(y)\prod_j \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x_j-\mu_i)^2}{2\sigma_j^2}}$$

$$= p(y)\frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}}$$

Where $\mathbf{\Sigma}$ is a diagonal matrix with $\sigma_1^2, \sigma_1^2, \ldots, \sigma_D^2$ as the diagonal entries

$\boldsymbol{\mu}$ is a vector of means

Problem 2.

Both the training and testing MSE calculated with and without using an intercept can be observed the table below

MSE	With intercept	Without intercept
Train data	2187.16029493	19099.44684457
Test data	3707.84018172	106775.3615526

In both cases, training and testing, the MSE with intercept is lower than the MSE without intercept. This is because using an intercept is better, as it provides more flexibility for the graph and allows it to begin from a non-origin position. So, calculating **MSE with an intercept is better**.

The following method is used for calculating the MSE values in this approach (Reference: Class notes)

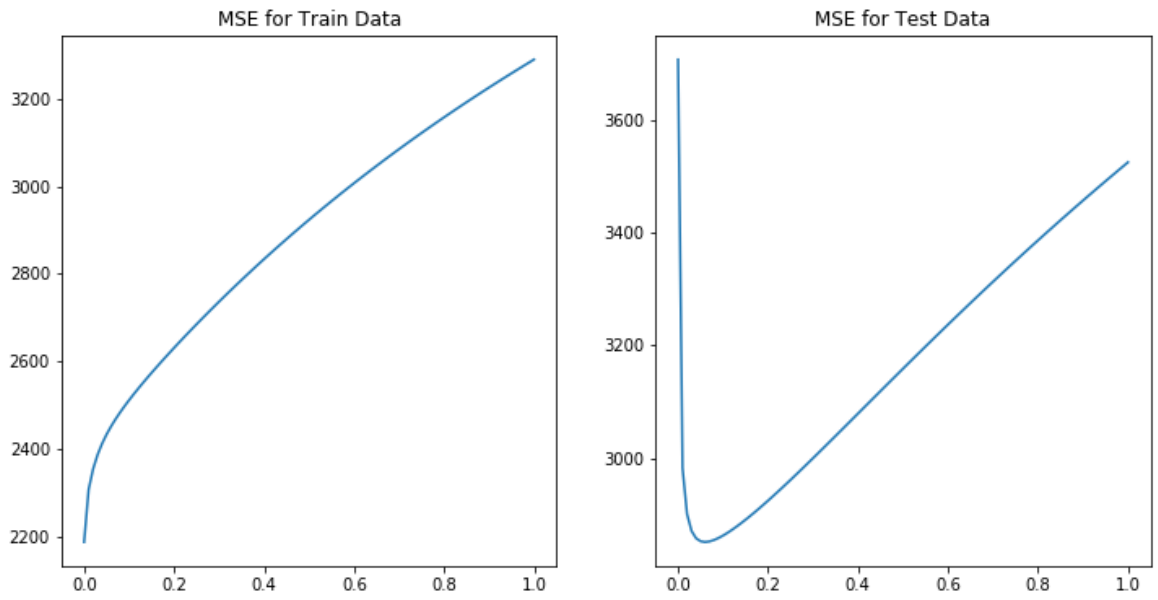
$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2$$

Problem 3.

The following method is used to calculate the MSE or the loss function (Reference: Class notes)

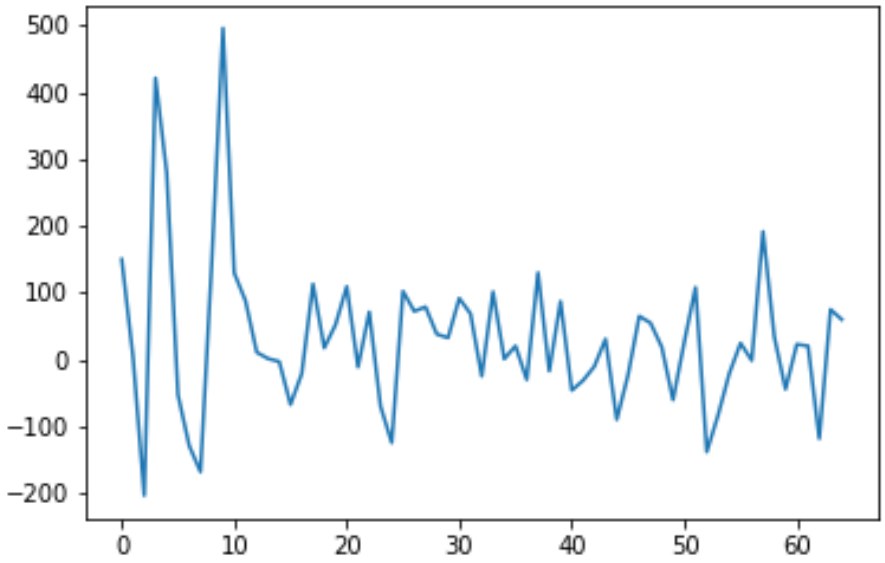
$$J(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \frac{1}{2}\lambda \mathbf{w}^\top \mathbf{w}$$

The following plot shows the MSE for training and testing data for various values for λ

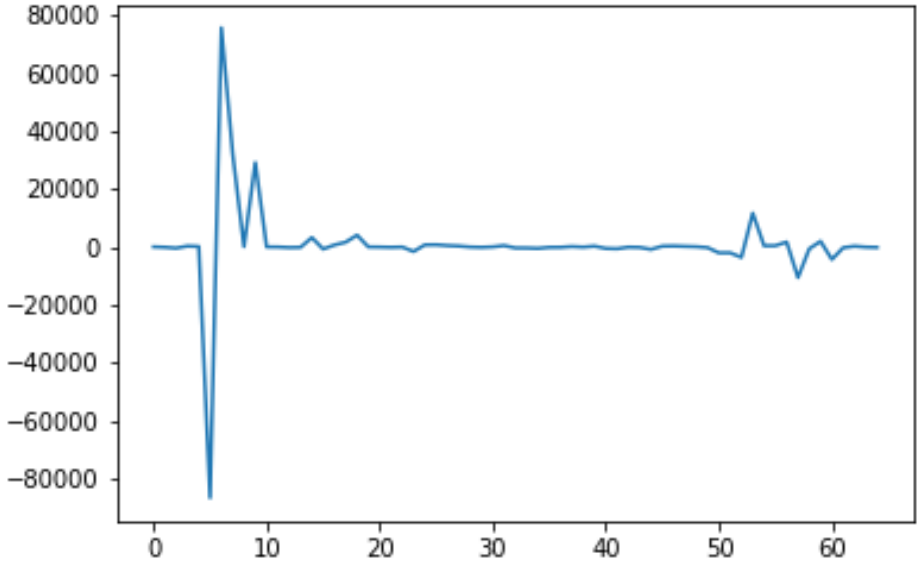


The following plots show the comparison between the weights for both Ordinary Linear Regression and Ridge Regression

For Ridge Regression



For Ordinary Linear Regression



From the graph we can see that the weights for Ridge Regression does not fluctuate as much as the weights for Ordinary Linear Regression

The following snippet gives us the minimum values for both training and testing data along with the corresponding λ values

```
Minimum MSE for Test data is 2851.3302134438463
Minimum MSE for Train data is 2187.160294930388
Min Test λ is 0.06
Min Train λ is 0.0
```

	Test	Train
0.00	3707.840182	2187.160295
0.01	2982.446120	2306.832218
0.02	2900.973587	2354.071344
0.03	2870.941589	2386.780163
0.04	2858.000410	2412.119043
...	...	...
0.96	3498.570906	3264.613861
0.97	3505.318324	3270.957170
0.98	3512.038029	3277.262582
0.99	3518.730082	3283.530490
1.00	3525.394553	3289.761281

101 rows × 2 columns

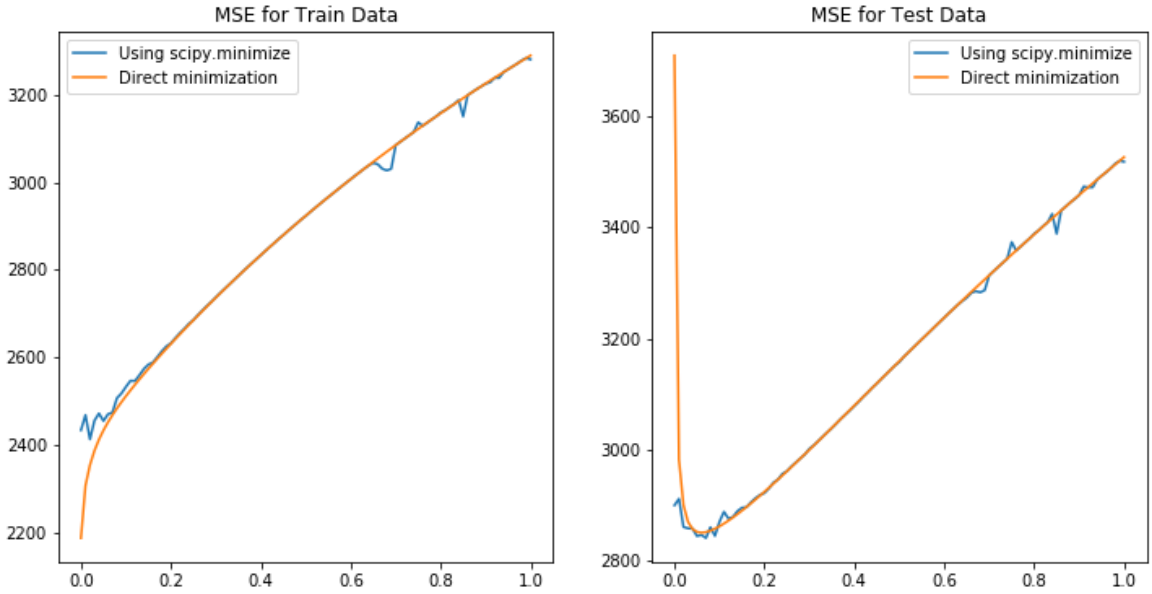
On comparison with the results from Problem 2, we see that the training errors for **both Ordinary Linear Regression and Ridge Regression is the same, which is 2187.16029493**. This is because the **λ value** for the train data during Ridge Regression was calculated **to be 0.0**. In the absence of any penalty, λ, Ridge regression gives the same results as Ordinary Linear Regression.

Meanwhile the testing errors are very different for Ridge Regression and Ordinary Linear Regression. The MSE for the **Ordinary was calculated to be 3707.84018172**, while the MSE for **Ridge was 2851.33021344**. This is due the penalty term **λ, the value** of which is calculated **to be 0.06** for the test data

The calculation for the optimal value for the λ can be seen from the snippet, as the value for the penalty term λ for the Ridge Regression is varied in its range, which is 0 to 1, and the λ value that gives the minimum value for the MSE is selected as the optimal value. For the given data, the λ for the train data is 0.0 and for the test data is 0.06. The **optimal value of λ** would be taken as **0.06**, as it gives the least MSE for the testing data, whereas the λ value of 0.0 which gives the least train MSE will cause overfitting, which can be observed by the large MSE for the test data at the same λ value.

Problem 4.

The following plot shows the comparison between the MSE values for Train and Test data for varying values for λ, calculated using Ridge Regression, and using Gradient Descent



From the graph we can see the difference between Ridge Regression Gradient Descent approach. We can see that for the most part the MSEs for Ridge Regression and Gradient Descent approach are similar but the difference is very large for very small λ values, and also for large values of λ. This is because in Ridge Regression, there is an addition of a penalty term to the loss function, where as in Gradient Descent approach, the loss function is minimized at every iteration. The difference arises when the values of the penalty term λ are either too small or too large, which causes the difference between the MSE values

The following snippet shows the calculation of MSE values for different values of λ. The least MSE and its corresponding values of λ are calculated in the snippet

	Train	Test
0.00	[2433.663715748937]	[2900.5465270951986]
0.01	[2468.0694188476455]	[2912.085061530347]
0.02	[2412.7288041591605]	[2861.573039827326]
0.03	[2455.453764295272]	[2859.1090394481666]
0.04	[2471.8986299413605]	[2858.7410590738023]
...	...	...
0.96	[3264.6151092608543]	[3498.088106993073]
0.97	[3270.9571902549274]	[3505.3083246133733]
0.98	[3278.3997222001444]	[3513.5109124395412]
0.99	[3283.6998786551235]	[3518.7835132583896]
1.00	[3280.783436426949]	[3517.1916829444212]

101 rows × 2 columns

```
The MSE and the corresponding λ value for the Train data are [2412.72880416] and 0.02
The MSE and the corresponding λ value for the Test data are [2841.78722097] and 0.07
```

From the snippet, we can see that the minimum MSE for training data is 2412.728804 at the λ value of 0.02, and the minimum MSE for test data is 2841.787220 at the λ value of 0.07. On comparing the MSE with ones obtained in Problem 3, we see that calculated the testing error to be 2851.33021344 at the λ value of 0.06, and training error to be 2187.16029493 at the λ value of 0.0. It can be observed that the optimal value of λ is different for both Ridge and Gradient Descent approach, with the least training MSE obtained with Ridge Regression, while the least testing MSE obtained by using Gradient Descent approach

Problem 5.

The following snippet shows the minimum MSE values for both train and test data for all the values of p, for $\lambda = 0$

```
          Train         Test
0  5650.710539  6286.404792
1  3930.915407  3845.034730
2  3911.839671  3907.128099
3  3911.188665  3887.975538
4  3885.473068  4443.327892
5  3885.407157  4554.830377
6  3866.883449  6833.459149
The minimum MSE for Training data is 3866.8834494460493 for p value of 6
The minimum MSE for Testing data is 3845.034730173414 for p value of 1
```

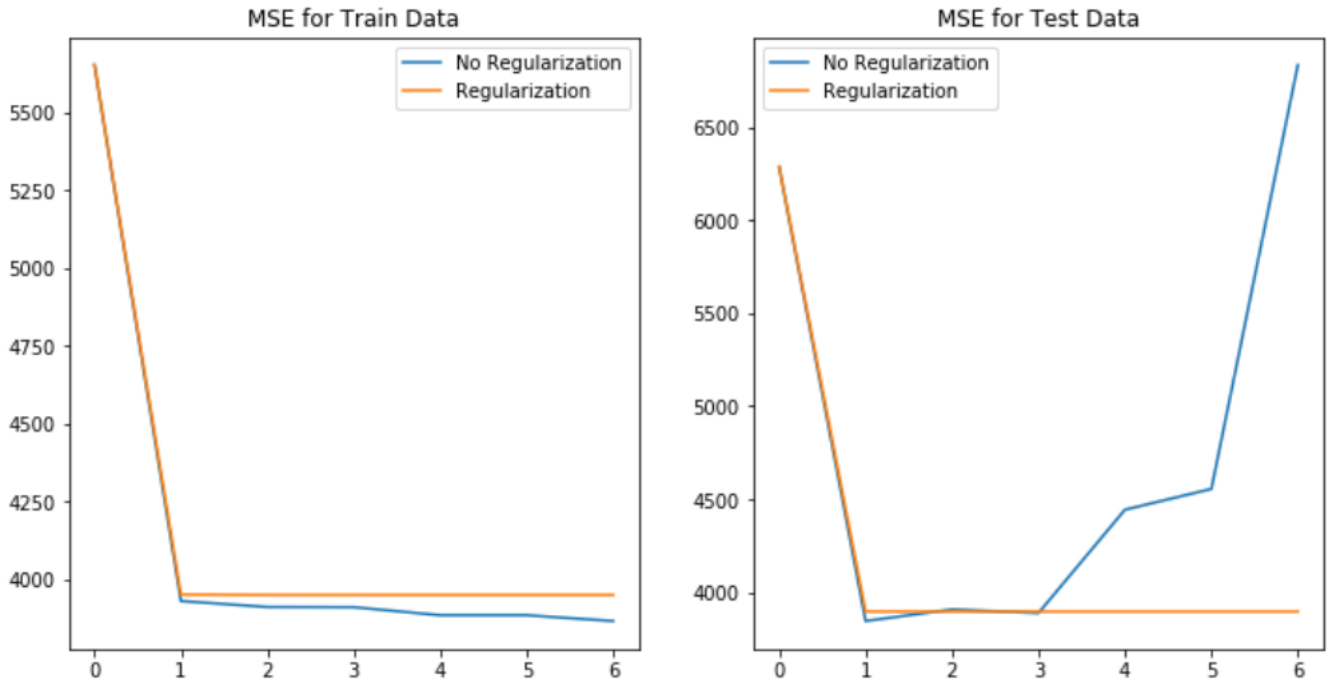
From the snippet, we can conclude that the test error is the least for p = 1

The following snippet shows the minimum MSE values for both train and test data for all values of p, for $\lambda = 0.06$

```
          Train         Test
0  5650.711907  6286.881967
1  3951.839124  3895.856464
2  3950.687312  3895.584056
3  3950.682532  3895.582716
4  3950.682337  3895.582668
5  3950.682335  3895.582669
6  3950.682335  3895.582669
The minimum MSE for Training data is 3950.682335142783 for p value of 6
The minimum MSE for Testing data is 3895.582668283526 for p value of 4
```

From the snippet, we can conclude that the test error is the least for p = 4

The following graph gives us the comparison between the train and test data for all values of p, using regularization, and without regularization

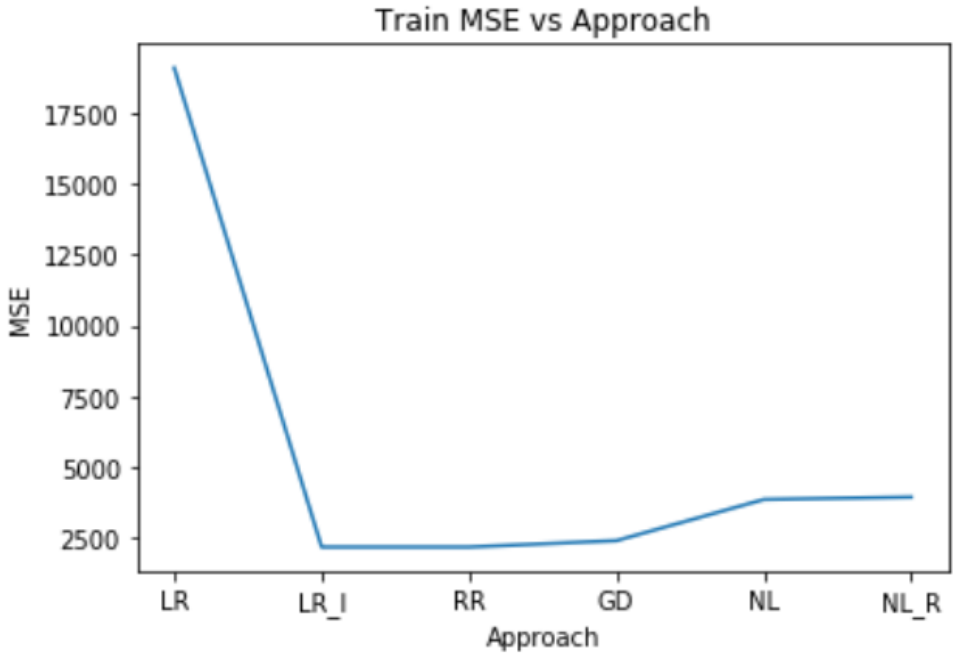


From the comparison graphs, we see that as p value increases, the more the training error decreases for λ value set as 0, whereas for the same λ value, the testing error first decreases for p value as 1, and then drastically increases as p value increases. This can be seen as the overfitting of training data with the increase in p value, which increases the testing error by a large margin. The optimal value of p can be taken as p = 1, where the testing error is the least.

On the other hand, in the absence any regularization, or λ value as 0, the MSE for both train and test data decreases from p value 0 to 1, but then becomes constant. This is because there is no penalty term for the increase in p value.

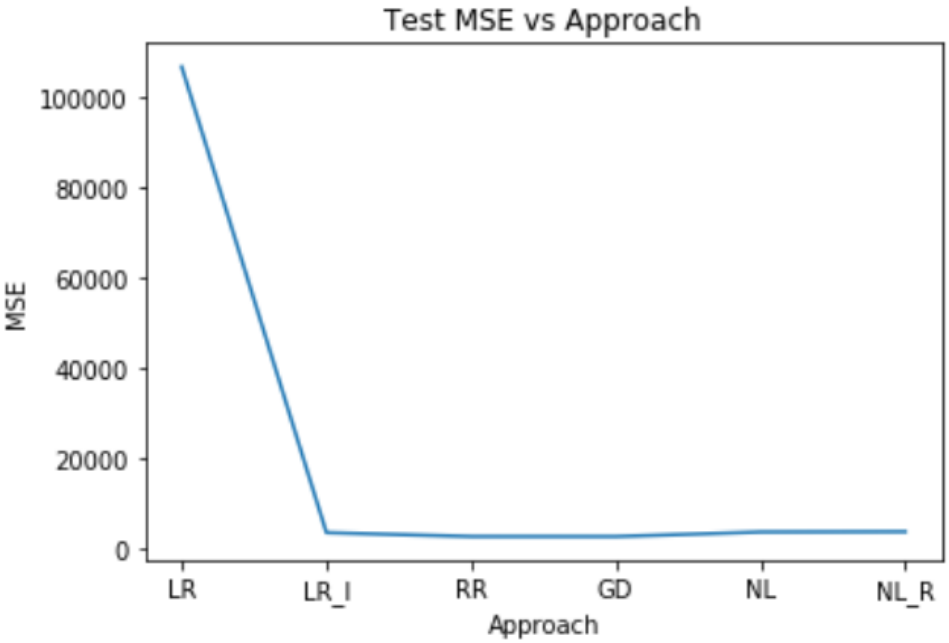
Problem 6.

The following plot shows the comparison for minimum MSE values for various approaches for training data



From the graph we can conclude that both Linear regression with intercept and Ridge Regression gives out same results with the least training error, which is 2187.160294

The following plot shows the comparison for minimum MSE values for various approaches for testing data



From the graph we can conclude that both Ridge Regression and Gradient Descent approaches give us similar results, with Gradient Descent giving out the least MSE value of 2841.787220

After comparing the errors calculated from various approaches, the metric for choosing the best setting is the MSE calculated from the testing data. The MSE or Mean Squared Error defines the variance of a prediction from the actual value. The smaller the MSE is, the more the data values are dispersed closely around its central movement, and the better it is able to predict future values. (Wikipedia, n.d.)

References

1) Github (n.d.). Retrieved from http://uc-r.github.io/discriminant_analysis

2) Wikipedia. (n.d.). Retrieved from Wikipedia, The Free Encyclopedia.: https://en.wikipedia.org/w/index.php?title=Mean_squared_error&oldid=981406211