CSE 574 Introduction to Machine Learning
Programming Assignment 1
Roshan Saundankar: 50419577
Sumeet Sahu: 50367891
Mrunmayee Rane: 50417094

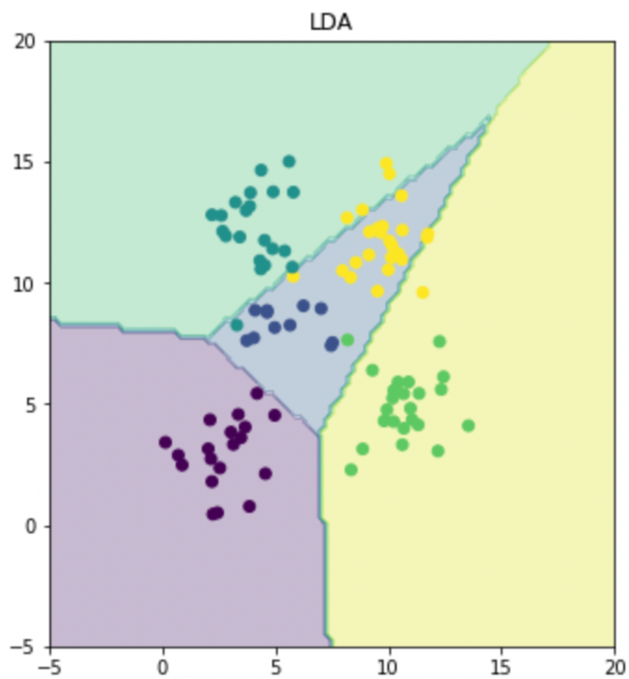
Classification and Regression

Introduction

We have to perform various regression and classification techniques on two datasets named "sample.pickle" and "diabetes.pickle".

Problem 1: Experiment with Gaussian Discriminators

Linear Discriminant Analysis (LDA):



LDA Accuracy: 73%

Observations:

- Discriminant boundaries can be observed which classify 5 classes
- The Boundaries are Linear.
- The covariance is assumed to be the same in each class

Quadratic Discriminant Analysis (QDA):

Did not solve in time.

Problem 2: Experiment with Linear Regression

Method	**Train MSE**	**Test MSE**
Linear Regression (With intercept)	2087.653816097928	3707.840181546761
Linear Regression (Without Intercept)	19227.67963945588	106775.36155661361

Findings:
It can be said by looking at the table above that the linear regression with intercept gives better results for the train and test datasets. In linear regression the when the line is modeled without an intercept, passes through origin therefore it is not the best model. Whereas, if it takes into consideration an intercept it provides results closer to the actual data, giving us better results.

Problem 3: Experiment with Ridge Regression

Training and Test Data: (for Lambda values 0 to 1)

For Train MSE:

```
[[2187.16029493]
 [2306.83221793]
 [2354.07134393]
 [2386.7801631 ]
 [2412.119043  ]
 [2433.1744367 ]
 [2451.52849064]
 [2468.07755253]
 [2483.36564653]
 [2497.74025857]
 [2511.43228199]
 [2524.60003852]
 [2537.35489985]
 [2549.77688678]
 [2561.92452773]
 [2573.84128774]
 [2585.55987497]
 [2597.10519217]
 [2608.49640025]
 [2619.74838623]
 [2630.8728232 ]
 [2641.87894616]
 [2652.77412633]
 [2663.56430077]
 [2674.25429667]
 [2684.84807809]
 [2695.34893502]
 [2705.75962912]
```

```
[2716.0825067 ]
[2726.31958674]
[2736.4726296 ]
[2746.54319109]
[2756.53266482]
[2766.44231574]
[2776.27330654]
[2786.02671854]
[2795.70356824]
[2805.30482034]
[2814.83139806]
[2824.28419133]
[2833.66406312]
[2842.97185452]
[2852.2083886 ]
[2861.3744735 ]
[2870.47090474]
[2879.49846701]
[2888.45793552]
[2897.35007697]
[2906.17565032]
[2914.93540723]
[2923.63009243]
[2932.26044392]
[2940.82719309]
[2949.33106473]
[2957.77277699]
[2966.15304137]
```

```
[2974.47256259]
[2982.73203851]
[2990.93215999]
[2999.07361078]
[3007.15706742]
[3015.1831991 ]
[3023.15266757]
[3031.06612707]
[3038.92422416]
[3046.72759776]
[3054.47687898]
[3062.17269114]
[3069.81564971]
[3077.40636224]
[3084.94542842]
[3092.43344001]
[3099.87098085]
[3107.25862691]
[3114.59694628]
[3121.88649919]
[3129.12783807]
[3136.3215076 ]
[3143.46804472]
[3150.56797875]
[3157.62183137]
[3164.63011677]
[3171.59334168]
[3178.51200544]
[3185.38660008]
```

```
[3192.21761044]
[3199.0055142 ]
[3205.75078202]
[3212.45387757]
[3219.11525768]
[3225.73537241]
[3232.31466512]
[3238.8535726 ]
[3245.35252514]
[3251.81194665]
[3258.23225474]
[3264.61386081]
[3270.95717015]
[3277.26258207]
[3283.53048993]
[3289.7612813 ]]
```

For Test MSE:

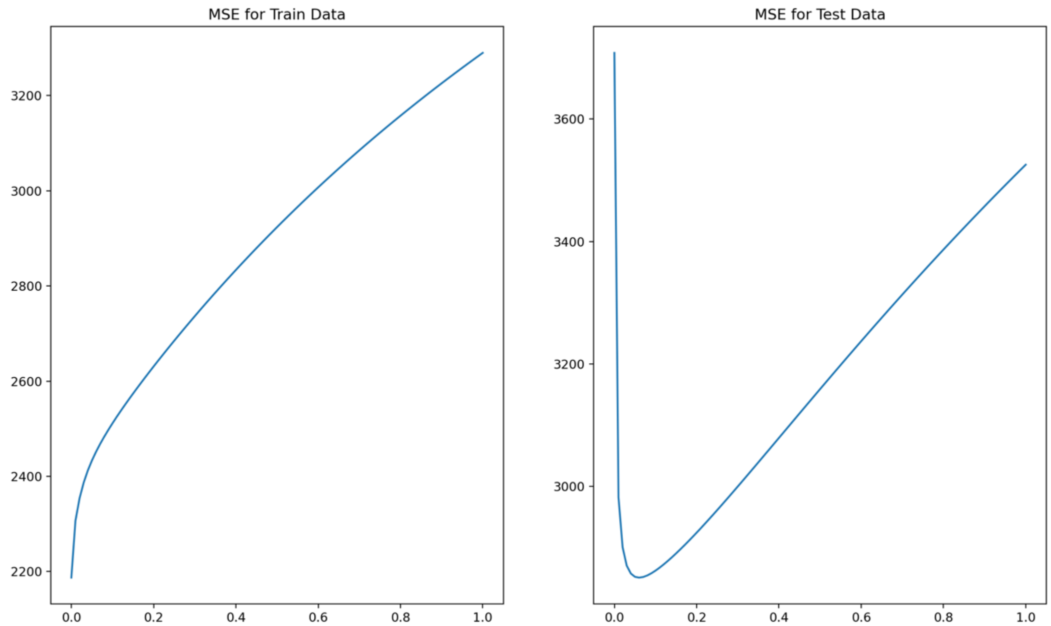
```
[2984.56432079]
[2992.31972181]
[3000.11580946]
[3007.94761559]
[3015.81055453]
[3023.70038563]
[3031.61318093]
[3039.54529713]
[3047.49335111]
[3055.45419817]
[3063.42491285]
[3071.40277169]
[3079.38523776]
[3087.36994673]
[3095.35469418]
[3103.33742413]
[3111.31621849]
[3119.28928746]
[3127.25496075]
[3135.21167941]
[3143.15798839]
[3151.09252966]
[3159.01403582]
[3166.92132421]
[3174.81329145]
[3182.68890838]
[3190.54721533]
[3198.38731777]
```

```
[[3707.84018174]
 [2982.44611971]
 [2900.97358708]
 [2870.94158888]
 [2858.00040957]
 [2852.66573517]
 [2851.33021344]
 [2852.34999406]
 [2854.87973918]
 [2858.44442115]
 [2862.75794143]
 [2867.63790917]
 [2872.96228271]
 [2878.64586939]
 [2884.62691417]
 [2890.85910969]
 [2897.30665895]
 [2903.94112629]
 [2910.73937213]
 [2917.68216413]
 [2924.75322165]
 [2931.93854417]
 [2939.22592987]
 [2946.60462378]
 [2954.06505602]
 [2961.59864341]
 [2969.19763677]
 [2976.85500119]
```

```
[3206.20838225]
[3214.00963255]
[3221.79034621]
[3229.5498512 ]
[3237.28752288]
[3245.00278108]
[3252.69508746]
[3260.36394297]
[3268.00888553]
[3275.6294878 ]
[3283.22535516]
[3290.79612376]
[3298.34145873]
[3305.86105245]
[3313.354623  ]
[3320.82191265]
[3328.26268646]
[3335.67673095]
[3343.06385289]
[3350.42387813]
[3357.75665047]
[3365.0620307 ]
[3372.33989556]
[3379.59013686]
[3386.81266063]
[3394.00738631]
[3401.17424594]
[3408.31318353]
[3415.42415428]
```

```
 [3422.50712403]
 [3429.56206859]
 [3436.58897321]
 [3443.58783202]
 [3450.55864755]
 [3457.50143021]
 [3464.41619786]
 [3471.30297539]
 [3478.16179431]
 [3484.99269234]
 [3491.79571308]
 [3498.57090566]
 [3505.3183244 ]
 [3512.03802854]
 [3518.7300819 ]
 [3525.39455263]]
```

Graph for errors in train and test data for different values of lambda:



Comparison of two approaches in terms of test and train MSE:

Method	**Train MSE**	**Test MSE**
Linear Regression (Without intercept)	19227. 67963945588	106775.36155661361
Linear Regression (With Intercept)	2087.653816097928	3707.840181546761
Ridge Regression (For optimal value of lambda)	2451.52849064	2851.33021344

Looking at the above values it can be said that Ridge regression performs much better than both Linear Regression with intercept and without.

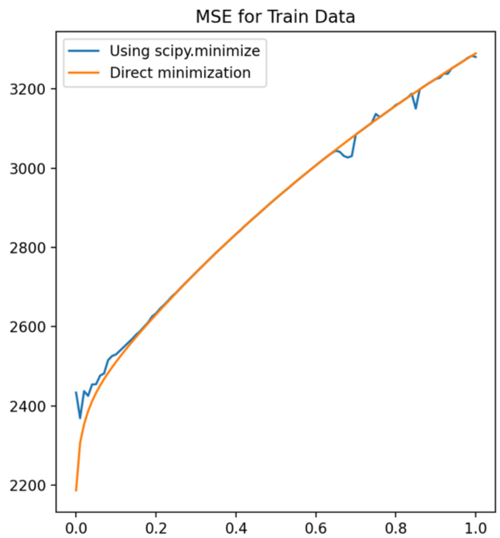
Optimum Value of Lambda:

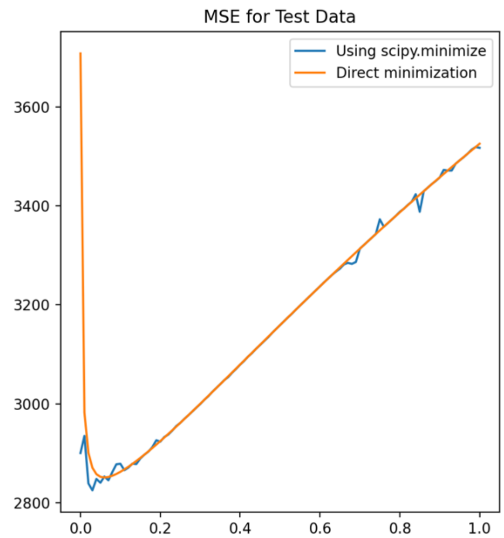
The optimal Value is 0.06 obtained by examining the minimum Test error in Ridge Regression.

Test Error: 2851.33021344
Train Error: 2451.52849064

Problem 4: Using Gradient Descent for Ridge Regression Learning

Plot the error on train and test data obtained by using the gradient descent based learning by varying the regularization parameter.





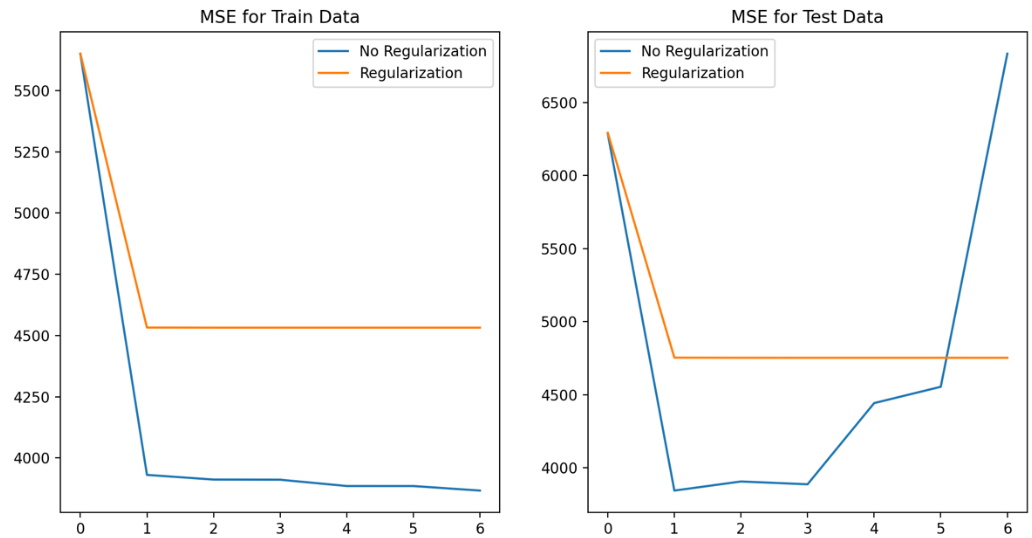
Comparing with results obtained from problem 3:

Method	**MSE Train**	**MSE Test**
Ridge Regression with Gradient Descent	2476.58754866	2853.35025988

Looking at the above Graphs it can be determined that the Training MSE for training and test for Ridge with and without Gradient Descent is almost similar.

Problem 5: Non-Linear Regression

Plots for Training and test error in case of nonlinear regression



MSE For Train Data:
It can be inferred from the above graph that, after regularization, the increase in p values the MSE values don't change a lot. Compared to when we do not have regularization i.e Lambda is 0 the MSE decreases with increasing polynomial degree because as the curve gets more non-linear, it tries to reduce the MSE.

MSE for Test Data:
When there is no regularization as the p values increase, the MSE for test data increases drastically as when regularization is not done, the model gets adjusted according to the train data. When regularization is done, the Test MSE is also consistent.

Train MSE:
Left column is for No Regularization and Right is for Regularization

```
[[5650.7105389  5650.71190703]
 [3930.91540732 3951.83912356]
 [3911.8396712  3950.68731238]
 [3911.18866493 3950.68253152]
 [3885.47306811 3950.6823368 ]
 [3885.4071574  3950.68233518]
 [3866.88344945 3950.68233514]]
```

Test MSE:
Left column is for No Regularization and Right is for Regularizatio

```
[[6286.40479168 6286.88196694]
 [3845.03473017 3895.85646447]
 [3907.12809911 3895.58405594]
 [3887.97553824 3895.58271592]
 [4443.32789181 3895.58266828]
 [4554.83037743 3895.5826687 ]
 [6833.45914872 3895.58266872]]
```

Problem 6: Interpreting Results

Comparing the Various approaches in terms of training and testing error

Approach	**Train MSE**	**Test MSE**
Linear Regression (Without Intercept)	19227.67963945588	106775.36155661361
Linear Regression (With Intercept)	2087.653816097928	3707.840181546761
Ridge Regression (Optimal value of Lambda 0.06)	2451.52849064	2851.33021344
Ridge Regression with Gradient Descent	2476.58754866	2853.35025988
Non-Linear Regression (No Regularization)	3866.88344945	3845.03473014
Non-Linear Regression (Regularization)	3950.68233514	3895.58266828

What metric should be used to choose the best setting?

Only looking at the Test MSE for all the regressions, we can conclude that Linear Regression without intercept has the highest values for MSE in train and test scenarios.
According to these values the best regression for train data is Linear Regression without Intercept and for test data it is Ridge regression with or without Gradient Descent as the values are similar. The remaining models do not perform very well.

All these conclusions are based on a small dataset, if the data increases, we should also take into consideration the running time of these functions. Therefore, Regularization becomes important when we have a huge dataset.