**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):

Chart, scatter chart

Description automatically generated

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:

Text

Description automatically generated with medium confidence

Text

Description automatically generated with medium confidence Text

Description automatically generated with medium confidence

Text

Description automatically generated

For Test MSE:

Text

Description automatically generated with medium confidence

Text

Description automatically generated Text

Description automatically generated

Text

Description automatically generated

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

Chart

Description automatically generated

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:

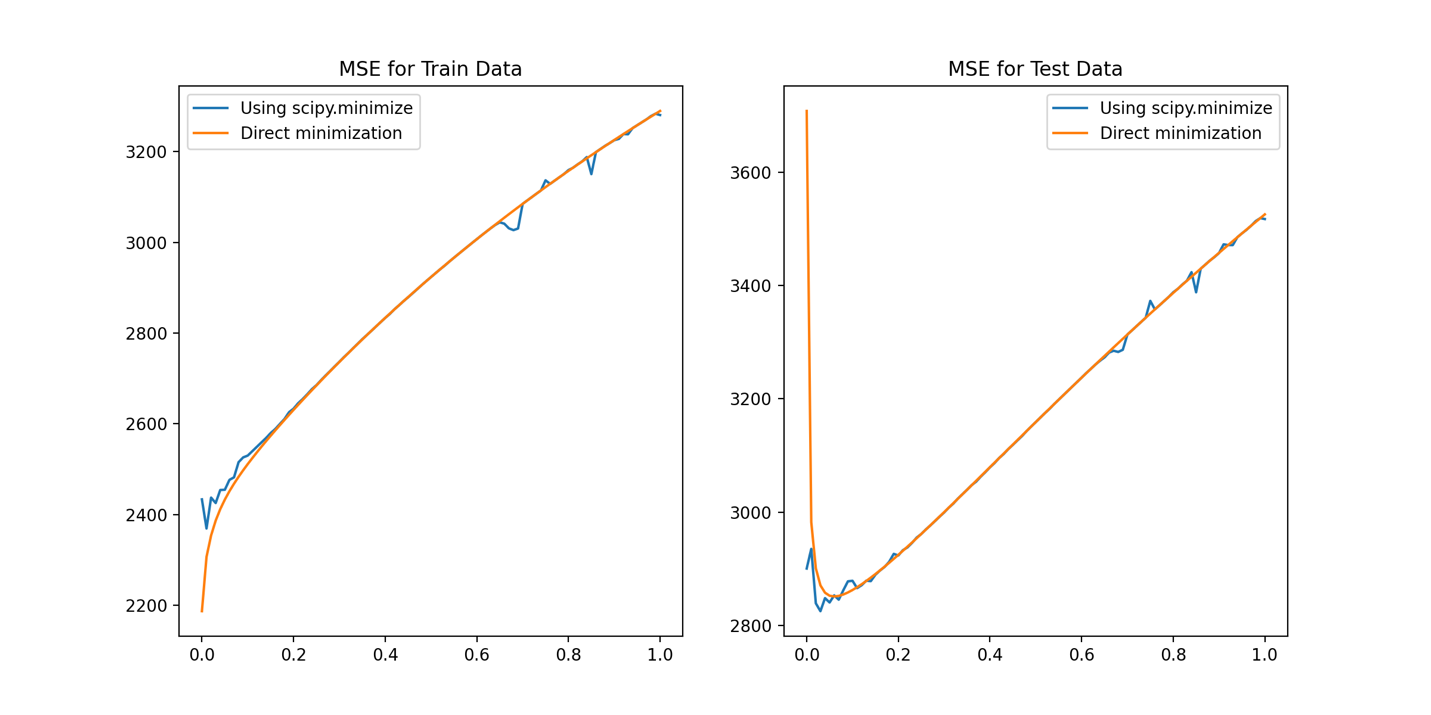
The optimal Value is 0.06 obtained by examining the minimum Test error in Ridge Regression.

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**

**Chart, line chart

Description automatically generated**

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**

**Text

Description automatically generated**

**Test MSE:**

**Left column is for No Regularization and Right is for Regularizatio**

**Text

Description automatically generated**

**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.