TEAM 21
PROGRAMMING ASSIGNMENT-1

SRAVANTHI ADIBHATLA(sadibhat)
ANUPRIYA GOYAL(anupriya)
SAGAR POKALE(sagarpok)

Problem 1: Experiment with Gaussian Discriminators

What is LDA and QDA?

Linear discriminant analysis (LDA), normal discriminant analysis (NDA), or discriminant function analysis is a generalization of Fisher's linear discriminant, a method to find a linear combination of features that characterizes or separates two or more classes of objects or events. LDA works when the measurements made on independent variables for each observation are continuous quantities. When dealing with categorical independent variables, the equivalent technique is discriminant correspondence analysis.

A quadratic classifier is used in machine learning and statistical classification to separate measurements of two or more classes of objects or events by a quadric surface. It is a more general version of the linear classifier.

LDA Accuracy = 0.97
QDA Accuracy = 0.96

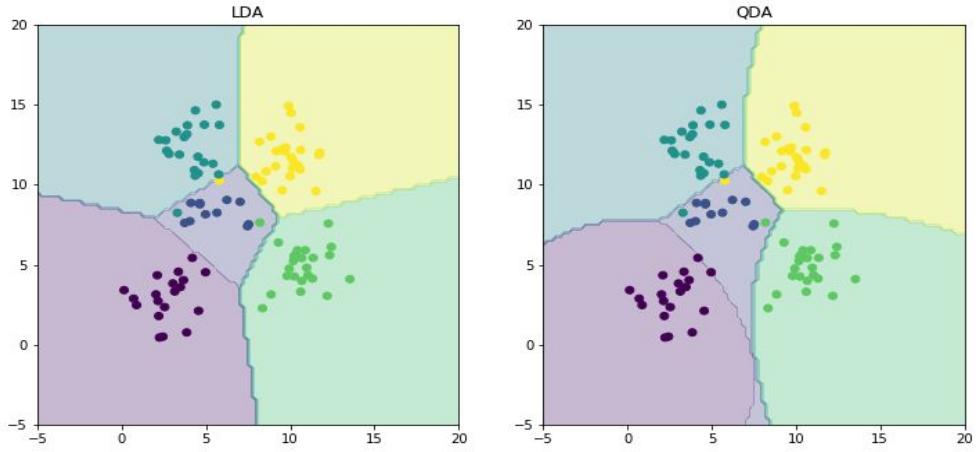

Fig 1: LDA and QDA decision boundaries

Comparing the difference in boundaries for LDA and QDA on the test data:
In LDA, the quadratic terms get cancelled and we have only **linear terms,** therefore, the **boundary is linear** because it has equal variances in all classes. In QDA, there are quadratic terms which will give us **quadratic boundaries** as they are defined by quadratic equations. It is not as straight as in LDA because we compute different covariance matrices for each one in classes, and the variability in each is not same.

Problem 2: Experiment with Linear Regression:

Report 2 Results:
Output:

```
MSE without intercept [[106775.36155355]]
MSE with intercept [[3707.84018128]]
```

MSE with intercept is better.
Reason:
Dropping the intercept in a regression model forces the regression line to go through the origin–the y intercept must be 0.
The problem with dropping the intercept is if the slope is steeper *just* because you're forcing the line through the origin, not because it fits the data better. If the intercept really should be something else, you're creating that steepness artificially. A more significant model isn't better if it's inaccurate.

If we observe, both the MSE values with an without intercept, the MSE with intercept is reduced approximately by factor of 30. This adding of intercept is analogous to adding a **bias term(i.e. adding dimension to the data).**

Therefore, using linear regression with intercept is better and helps in predicting classes in better way.

Problem 3: Experiment with Ridge Regression:

Report 3 results:
Why do we use ridge regression?
To avoid overfitting, we can do L2 regularization with linear regression. This is called ridge regression.

$$\widehat{w} = arg_w \ \ \textbf{\textit{min}}\ \mathbf{J(w)} + \boldsymbol{\lambda} ||w||_2^2$$

Here, we add a square of magnitudes of coefficients and multiply with lambda in the result.

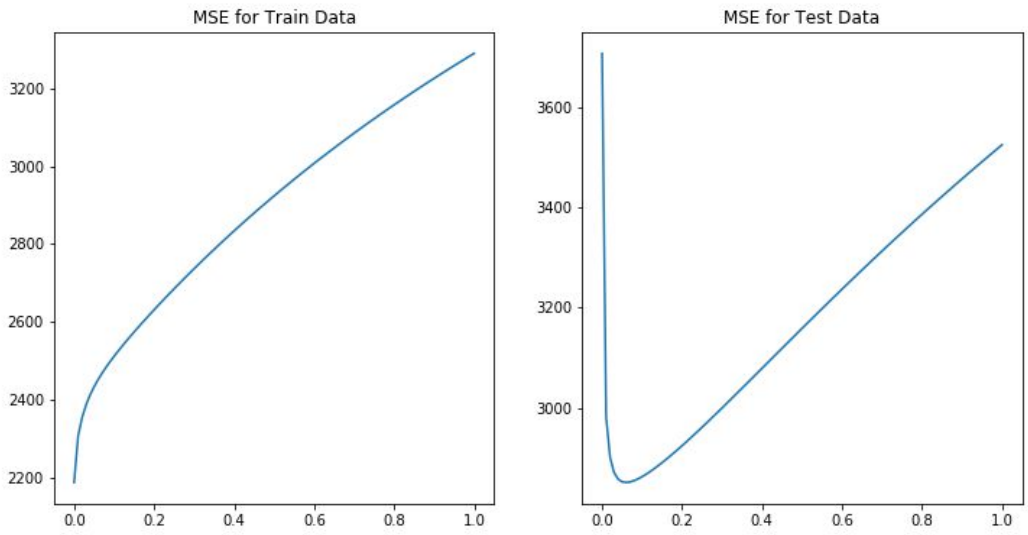

Fig 2: Ridge regression

Comparing w values of linear and ridge regression:

When we try to compare w values of both linear and ridge regression, we observe that, output is not biased in both cases. But, in linear regression, the variation is more. Bias- Variation trade off is balanced in ridge regression.

Comparing train and test data for ridge regression:

In train data, there is no regularization as $\lambda = 0$, so the error is the lowest here. But when you apply to test data, your error is higher, this is the notion of poor regularizability.

However, as you regularize it more, there comes a point at which incurring a higher training error, we are getting in lower tester error. This is the point weight vector is regularized just enough. This is the optimal value for λ.

Optimal Value of λ:

The optimal value of λ is 0.06 , which is the lowest point in the curve of MSE for Test Data i.e. where the test error is minimum. Here, the curve or weight vector is regularized just enough. Beyond, this point if you try to regularize too much, again the error will increase.

Problem 4: Using Gradient Descent for Ridge Regression Learning
Report 4 results:
Ridge regression with Gradient descent:

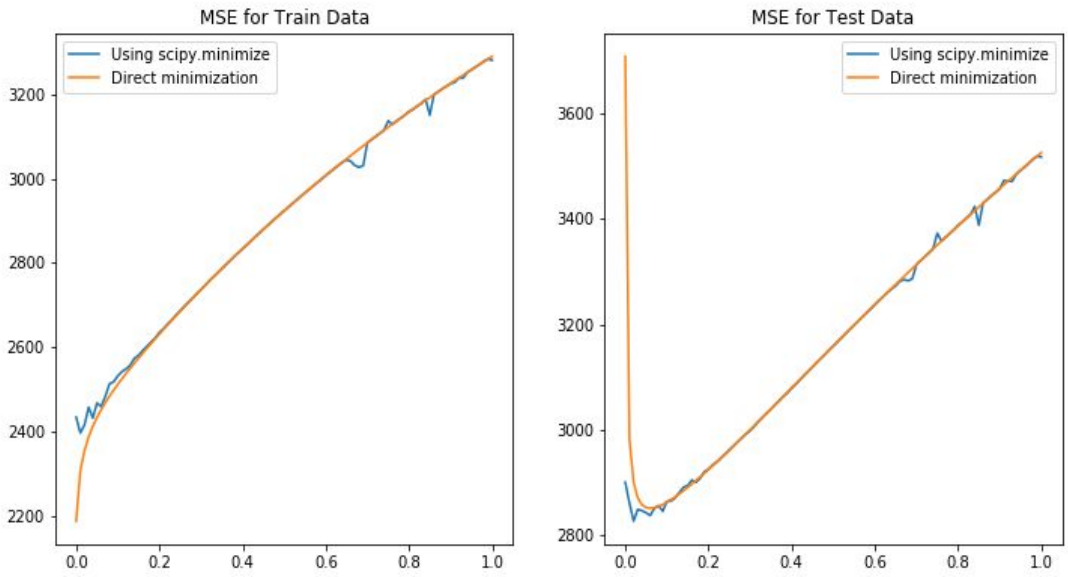

Fig: 3 Ridge with Gradient Descent

Comparing ridge regression with and without gradient descent:

If we observe the figure 2 and 3, the graphs are not very dissimilar. But the curve with gradient descent is not as smooth as in Problem 3. The spikes in the curves are due to outliers and the weights do not converge as quick as normal ridge regression based learning. As the data set size is small in terms of numbers and features, so will be the value of W, thus reducing the overhead and time for learning. So, in this scenario, since the data set in small, it's better to use normal ridge regression than gradient descent based training , because we are getting smoother curves.
In gradient descent, we are fixating the iterations because of which we might not reach the optimal value.This is another reason to go for normal ridge regression.

Problem 5: Non-linear Regression:

Report 5 results:

Nonlinear regression is a form of regression analysis in which observational data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables. The data are fitted by a method of successive approximations. In this problem, we are using higher order polynomials and are checking the impact of it on input.

Variation in mean squared error with p values from 0 to 6 :

lambda=0

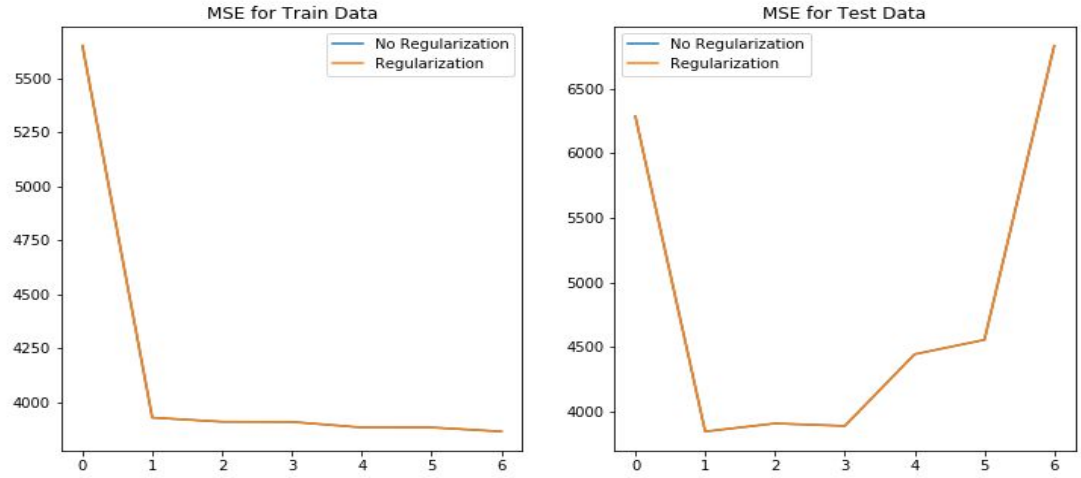

Fig 4.1 No regularization in non-linear regression

lambdopt=0.06

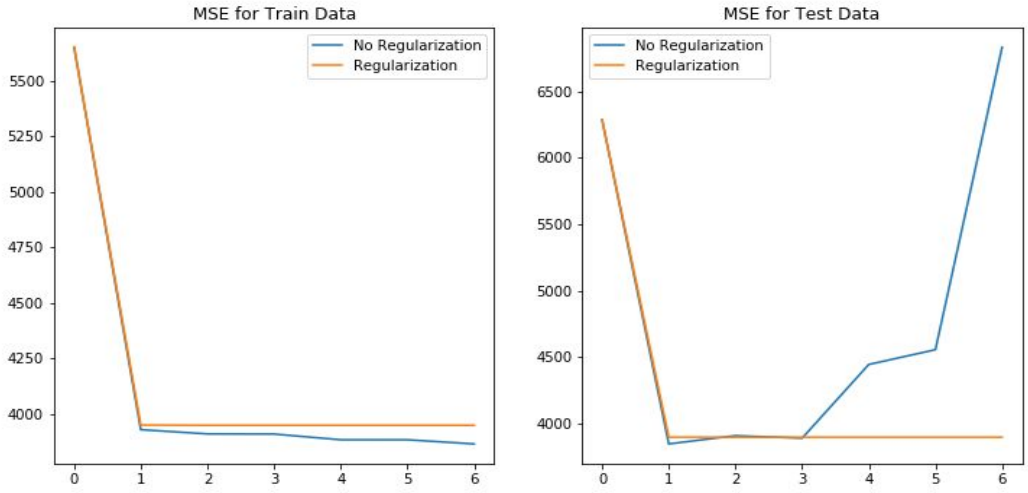

Fig 4.2 Regularization in non -linear regression

Analysis on training data:

In *Fig 4.1*, we did not perform any regularization, and the optimal value of p = 1, in both cases. As we increase the complexity we are making the curve to be more non-linear, but our test error actually goes up. This is the case of poor generalisability. We made our solution too complex, so even though training error was small, test error shot up. It happens when p=3 and it is due to the overfitting of model when higher degree of polynomial is used.

However, in *Fig 4.2* when do regularization, even though the training error is slightly worse than the no-regularization one, the test error is good. The solution is still non-linear but not too complex. As the order of polynomial increases, the squared error of ridge regression learning decreases. The same holds true for lambda as well as lambdopt. So, p=6 is the optimal value of p in this case.

Problem 6: Interpreting Results:

Report 6 Analysis:

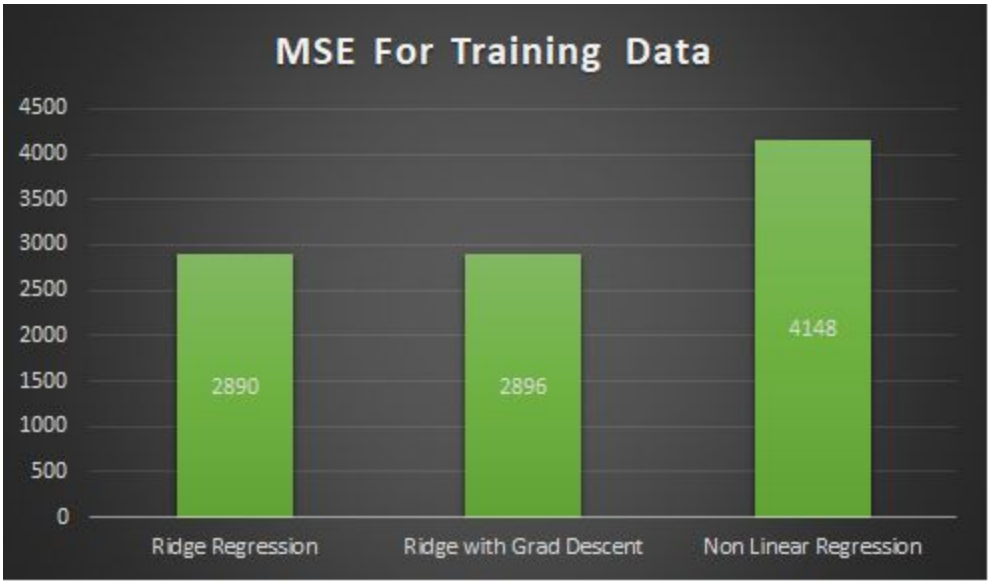

Fig 5.1: Comparison of training errors for Ridge Regression with optimal value and with gradient descent, Non-Linear Regression

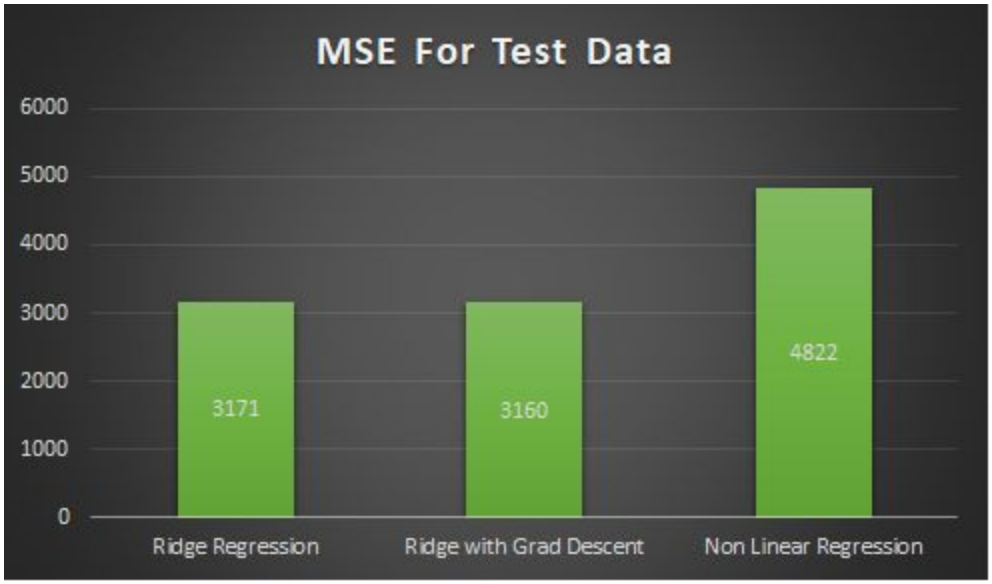

Fig 5.1: Comparison of testing errors for Ridge Regression with optimal value and with gradient descent, Non-Linear Regression

Analysis:

From the above graph, we can see that the train error increases, and test error decreases, in the order of Linear regression, ridge regression with optimal value to gradient descent.
Test error is measure of accuracy and efficiency of the method.

For polynomial of higher order, non-linear regression causes overfitting problem on the test data. That is the reason why test error increases suddenly after p =3.

So, finally we can conclude that even though gradient descent graphs are not smooth, the test error is least in ridge regression with gradient descent.

References:

https://en.wikipedia.org/wiki/Linear_discriminant_analysis
https://www.theanalysisfactor.com/regression-models-without-intercepts/
https://codingstartups.com/practical-machine-learning-ridge-regression-vs-lasso/