COMP6245 Foundations of Machine Learning – Lab 4

Supritha Konaje
Email: ssk1n21@soton.ac.uk
Student ID: 32864477

1. Linear Least Squares Regression

Linear regression analysis is a statistical technique for predicting the value of one variable based on the value of another. We used samples from the diabetes dataset, which had 442 samples. The sklearn package is used to load the data. The histogram of targets is shown in Figure 1. In Figure 2, we created a scatter plot using two features from the dataset.

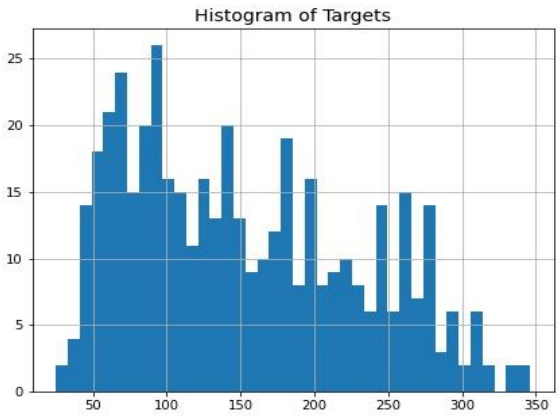

Figure 1: Histogram of Targets

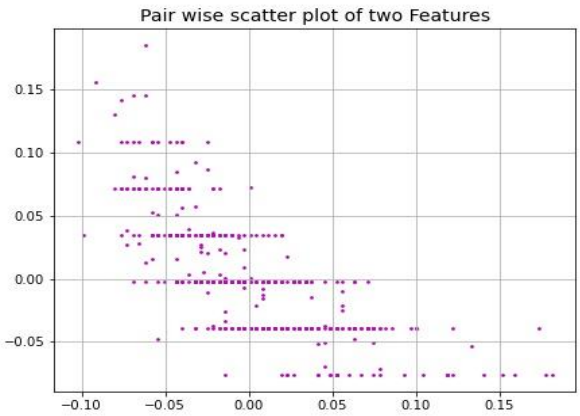

Figure 2: 7th and 8th input Feature

We will be predicting the results using sklearn and Pseudo Inverse Solution. Pseudo Inverse is used to compute the best fit solution to a linear equation that lacks a solution. The formula for Pseudo Inverse is $w = (X^t X)^{-1} X^t t$. In Figure 3, scatter plot for linear model is plotted using sklearn Linear Regression function. In Figure 4, scatter plot for Pseudo Inverse is plotted. The parameters used to plot the scatter are the inputs of two features and the target. Comparing the two plots, the scatter plot is the same. There is not much of a difference in both the plots. Hence, it can be concluded that both the predictors have almost the same accuracy.

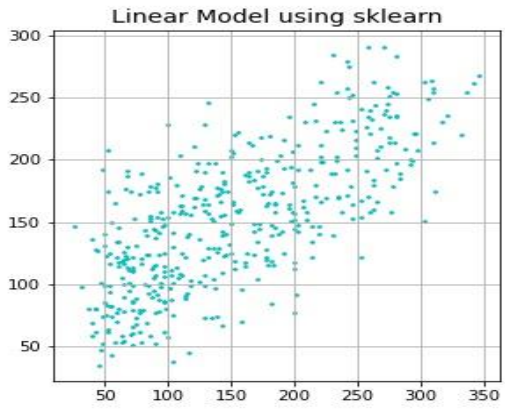

Figure 3: Prediction using sklearn

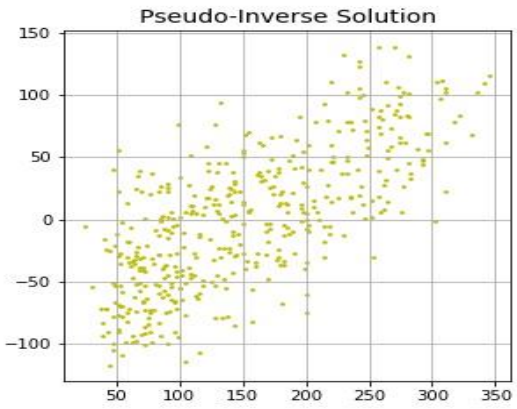

Figure 4: Prediction using Pseudo Inverse

2. Regularization

Another approach of regularising a problem with several solutions is Tikhonov regularisation. It aids in preventing the predictor variable in a multiple regression model from being linearly predicted from the others with a high degree of accuracy, which is common in models with many parameters. In Figure 5, the parameters for Pseudo Inverse Predictor are plotted. Figure 6 shows the parameters for Tikhonov regularization. The parameters for Tikhonov are smaller as compared to Pseudo Inverse regularization. Hence, the prediction will be also be quicker.

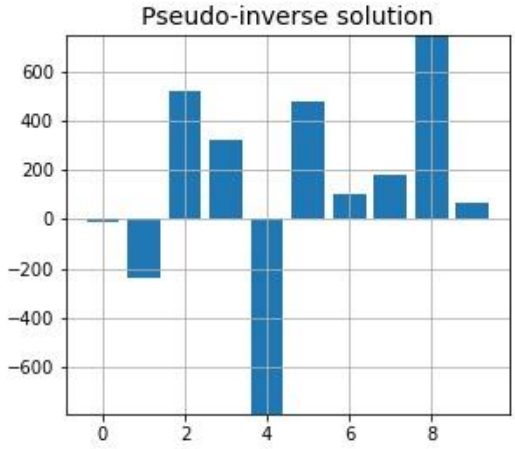

Figure 5: Pseudo Inverse Solution

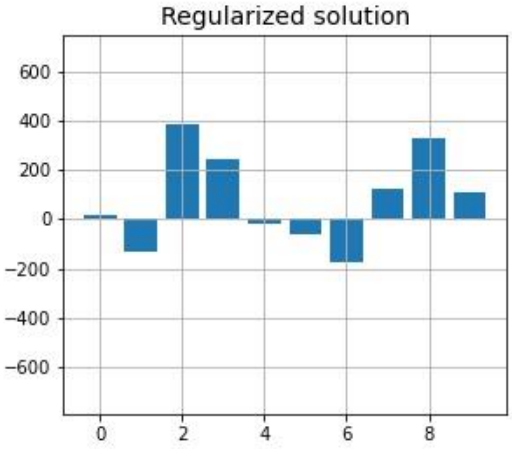

Figure 6: Tikhonov Regularization

3. Sparse Regression

Sparse Regression helps reduce the dimensions of the parameters. In Figure 7, graphs for each value of alpha are plotted with alpha as 0, 0.1, 1, 2 respectively. As per observation, as and when we increase the value of alpha the number of non-zero weights decreased. But their error increases.

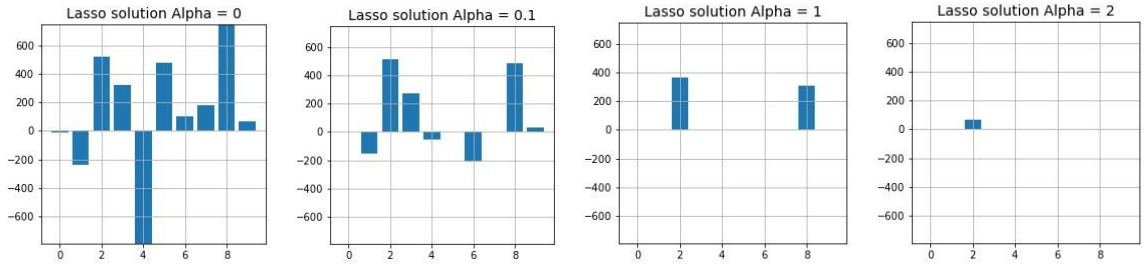

Figure 7: Parameters of Solutions for Lasso Solution for Alpha = 0, 0.1, 1, 2 respectively

For better understanding, we shall plot the regularization path for 6 variables. In Figure 8, we can see the regularization paths for the 6 variables. It is observed that the value of each parameter decreased and the value of non-zero weights increased.

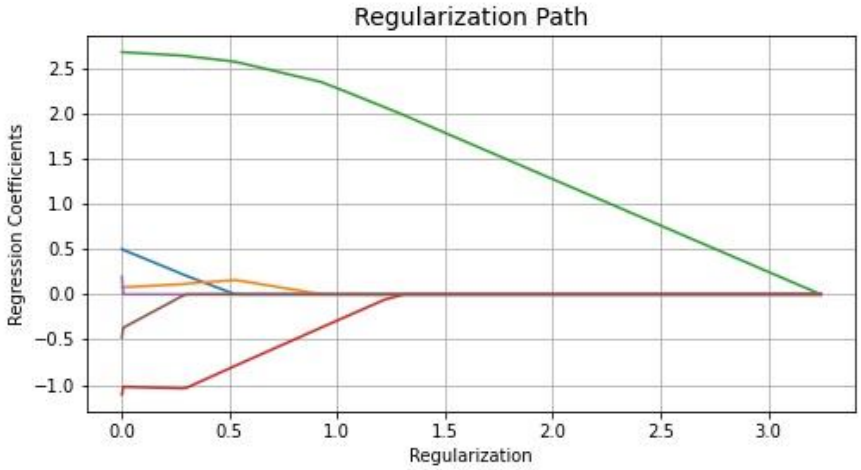

Figure 8: Regularization Path

4. Solubility Prediction

We are using the Husskonen Solubility Features for prediction here. This is quite a large dataset related to chemical compounds. We have loaded the dataset using pandas. In Figure 9, histogram for Log Solubility is plotted.

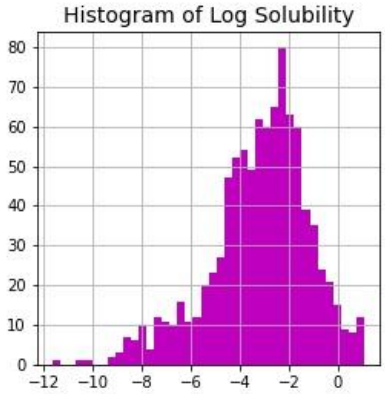

Figure 9: Histogram for Log Solubility

We have taken 0.3 as the test size that is testing set involves 30% of the data and the training set involves 70% of the data.

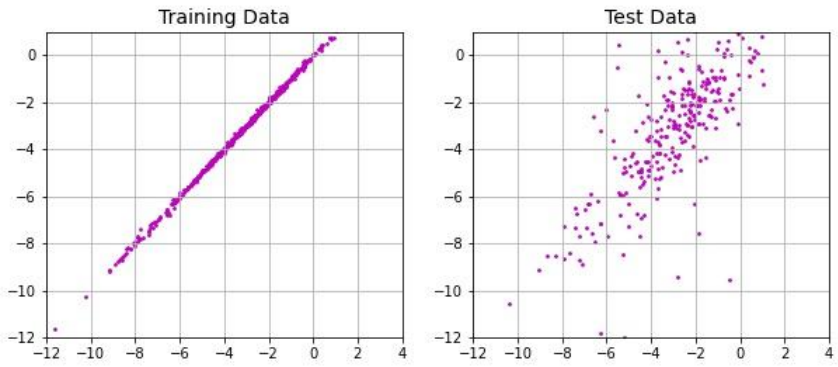

Figure 10: Training data and Test data scatter plot

In Figure 11, the parameter solution for each of the regularization model is showcased. As per observation, Lasso solution number of non-zero values weights are decreased and if this is reduced, over fitting problem will also be less.

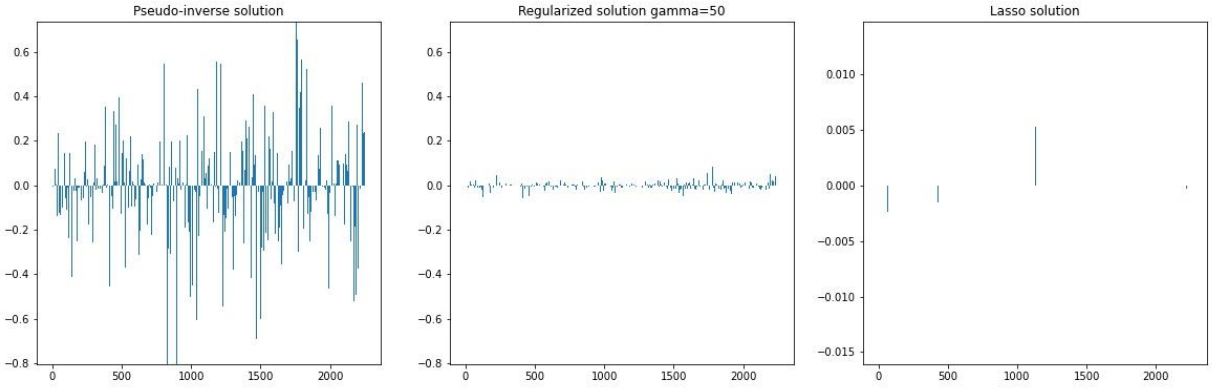

Figure 11: Parameter Solution for Pseudo Inverse, Regularised and Lasso Solutions

According to the papers, the main thing that needs to be focused is to have a balance between efficiency and accuracy. The accuracy mainly depends on the target that needs to be achieved and the efficiency depends on the dataset quality.