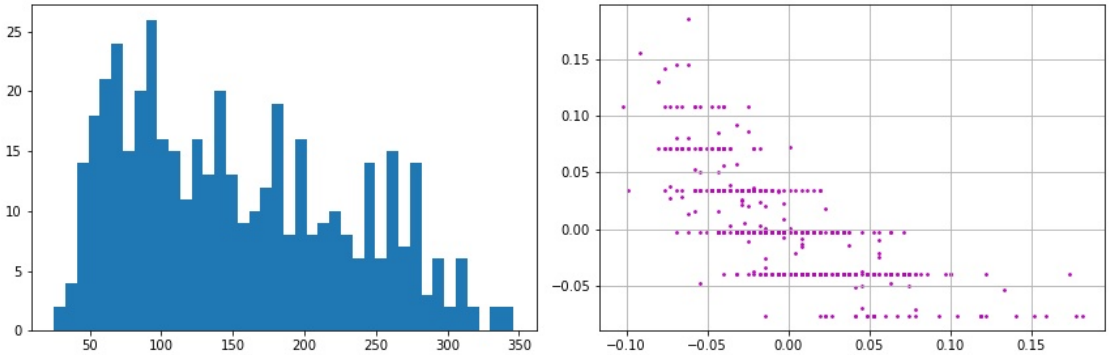
Foundations of Machine Learning Lab 4 Report

Abhilash Pulickal Scaria
aps1n21@soton.ac.uk
Student Id:33124639

Observation

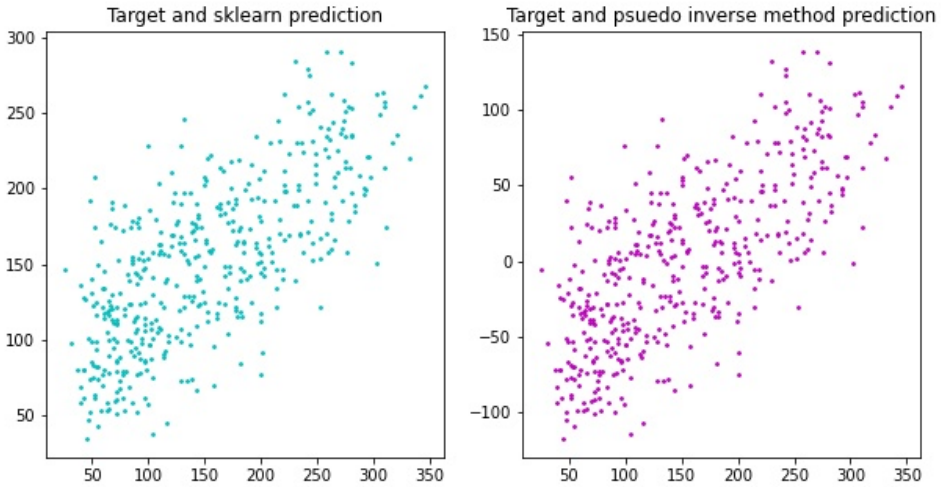
1)Diabetes dataset is loaded using the sklearn package. Using the first code snippet we can observe the histogram of target values and scatterplot of 6th and 7th feature. The results obtained while

Figure 1: Histogram and Scatterplot.



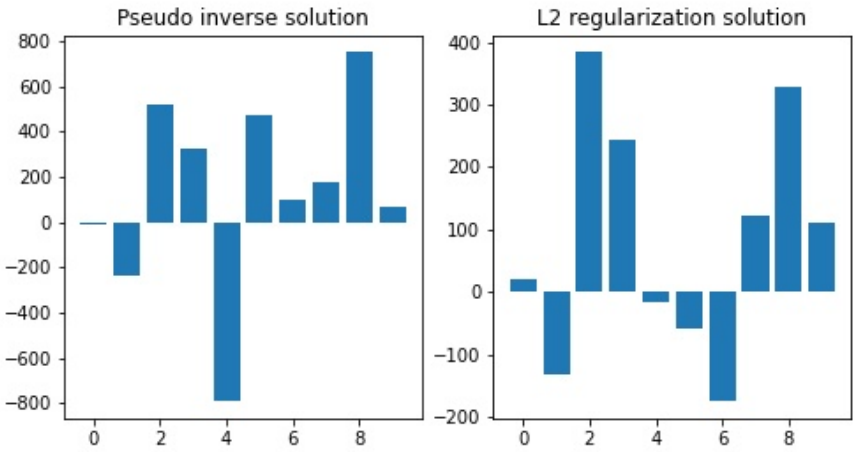
solving with pseudo inverse method and by using linear regression from scikitlearn package are given in Figure 2. It can be observed that prediction of psuedo inverse method is not as accurate as prediction of linear model from sklearn package.

Figure 2: Scatterplot.



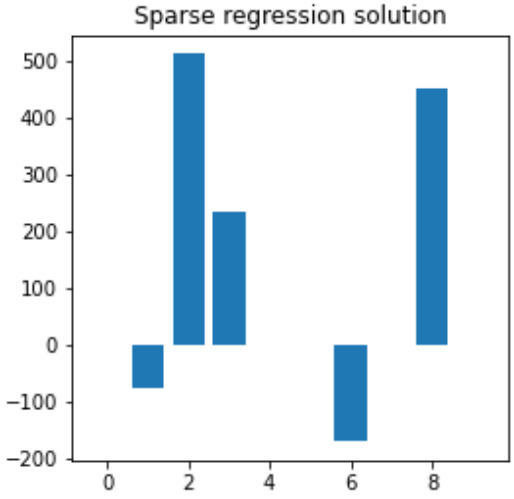
2)Using code snippet 3 we can implement L2 regularizer. Plots of weights for psuedo inverse solution and ridge regression(L2 regularization) is given in figure (3) . It can be observed that the values of weights decreses when regularization term is introduced. Having larger coefficients(weights) results in overfitting and introducing a regularization term can prevent this.

Figure 3: Bargraph of weights.



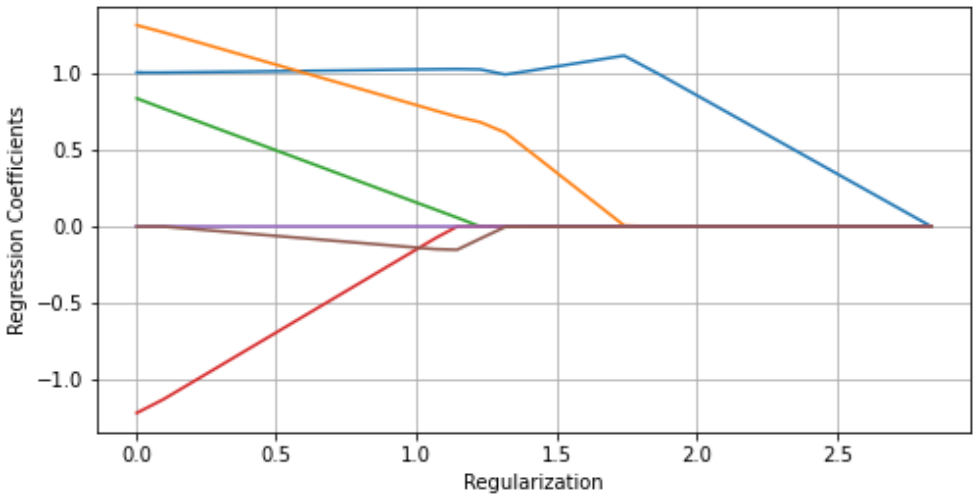
3)After implementing sparse regression solution using code snippet 4 we can plot bargraphs of weights for sparse regression. From figure (3) and (4) we can observe the values and number of non zero weights change with regularization parameter. By calculating Mean square error for each cases we can observe lasso regressor has the least error followed by the other two. For this particular case where we haven't split data and we are using the same data for testing pseudo inverse method is having less error than L2 regularizer, But in actual scenarios where we use seperate test and train data, L2 regularizer will have less error than pseudo inverse method. Sparse regression method can shrink some weights which doesn't have much impact on output to zero. In the case of diabetes dataset sparse regression has retained the features that influence the output the most which include bmi, bloodpressure, sex and 3 blood serum measurements.

Figure 4: Bargraph of weights.



4) Using code snippet 5 and 6 we can observe regularization path for six variables. The plot shows how regression coefficients(weights) change as regularization parameter λ is increased.

Figure 5: Reguralization path.



5)

Figure 6: Scatterplot of predicted solubility and true solubility for linear Regression.

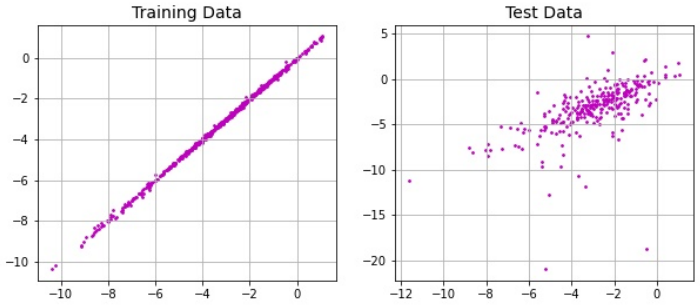
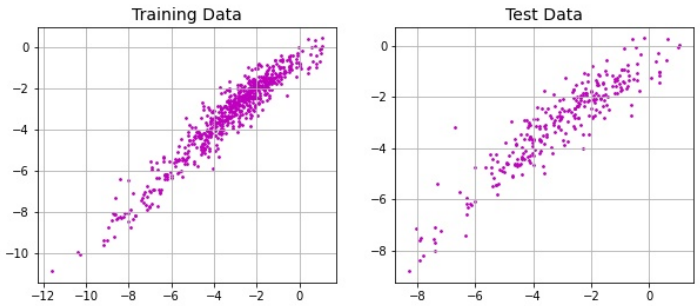


Figure 7: Scatterplot of predicted solubility and true solubility for lasso regression.



Using code snippet 7 we can analyse and predict solubility of chemical compounds using data from HusskonenSolubilityfeatures dataset. By implementing a linear regression model the scatterplot

obtained between true values and predicted values for test and train data is given in figure(6). After implementing the lasso regressor and plotting scatterplot figure(7) we can see that the test data prediction error has decreased and model has become more generalized. The change in number of nonzero coefficients can be observed from figure(8). The top ten features can be idenified by taking the features corresponding to top ten weights. The prediction accuracy obtained by using these ten features(figure 9) is almost as good or slightly better than that of psuedo inverse method trained with all features and L2 regularization. But it doesn't perform as good as ridge and lasso regressor implemented using sklearn. There are 80 non zero weights, if some more of the features corresponding to the weights are used then the accuracy may become better.

Figure 8: Bargraph of weights.

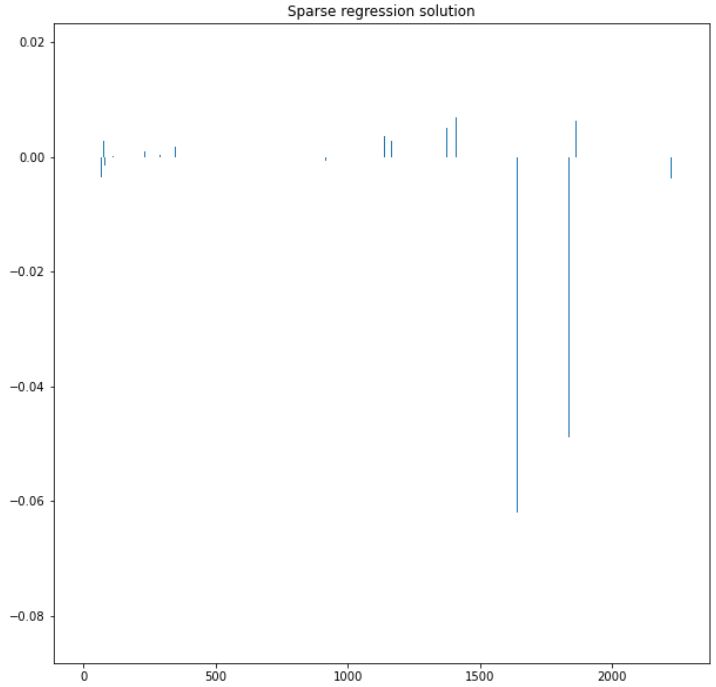


Figure 9: Scatterplot of predicted solubility and true solubility for regression using ten features.