Linear Least Squares Regression

First we load the diabetes data from sklearn package, and in order to inspect some features and targets in this dataset, the targets distribution is plotted as histogram and the 7th, 8th dimension of the data are plotted as scatter as **Figure 1** shows.

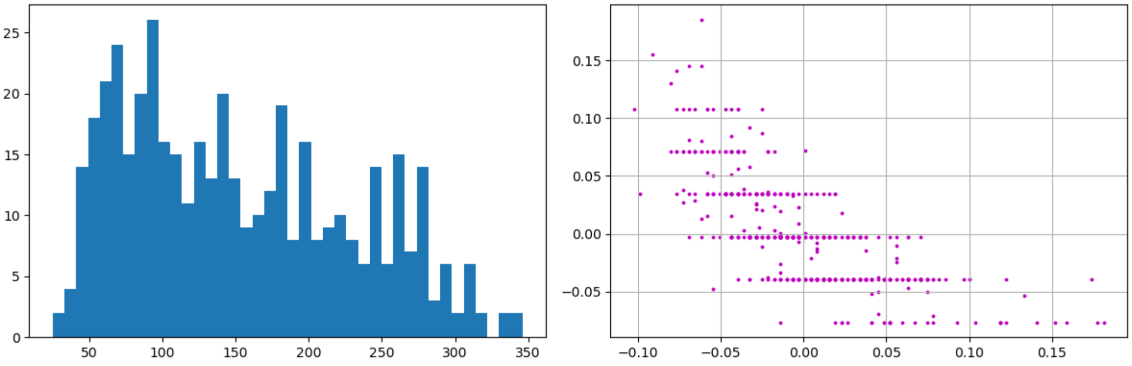

Figure 1: Diabetes dataset inspection

Then we use pseudo-inverse method($w = X^tX)^{-1}X^tt$) to directly compute the parameter of the linear predictor based on diabetes dataset and to perform the prediction which is shown as **Figure 2**. We also use imported sklearn predictor to predict as a contrast, the result is shown in **Figure 3**. If the points are closer to the diagonal means that the performance of the predictor is better since one axis represents the prediction value and another one represents the actual value. In this case, no significant difference could be detected between two results.

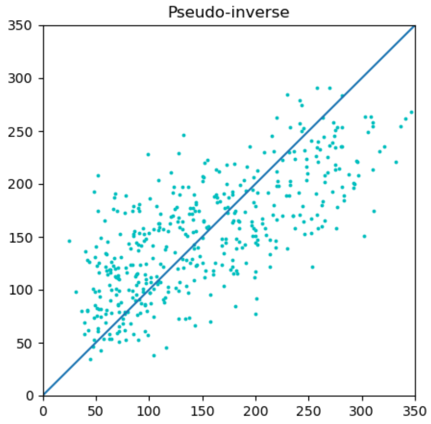

Figure 2: Pseudo-inverse

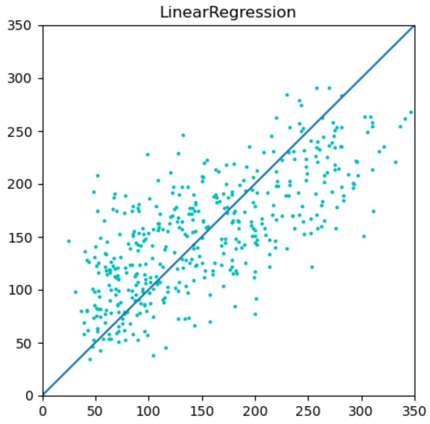

Figure 3: Linear Regression

Regularization

L2 regularization is a good way to prevent parameters from becoming too large while training and overfitting. As **Figure 4** shows, almost all parameters after regularization are smaller than before.

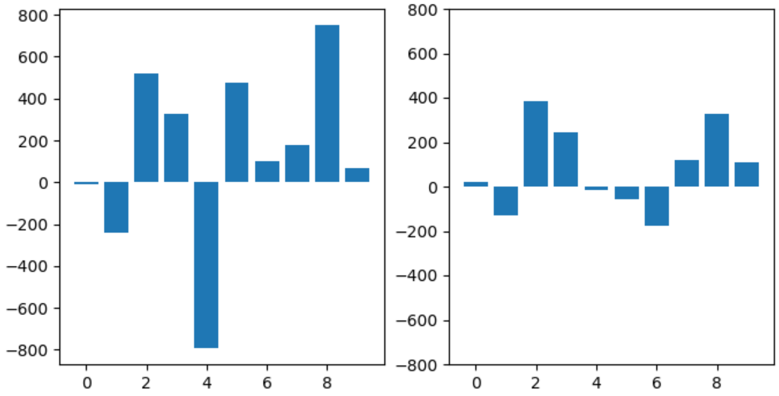

Figure 4: Parameters of pseudo-inverse solution before and after regularization

Sparse Regression

Sparse Regression is also helpful to prevent overfitting by simplify the model. By gradually increasing the value of alpha, the number of parameters(non-zero weights) decreased as **Figure 5** shows while the error increased from 2995(α=0.2) to 3152(α=0.4) and 3548(α=0.8).
And after looking at the attribute information of each feature, I think the features with non-zero weights are meaningful. The third, fourth and ninth feature attributes are BMI, avg blood pressure and serum triglycerides level respectively. These factors are very important to determine if someone is prone to have diabetes. The seventh feature is high-density lipoproteins which is beneficial to our health, but its weight is negative which means less high-density lipoproteins contributes to diabetes which makes sense as well.

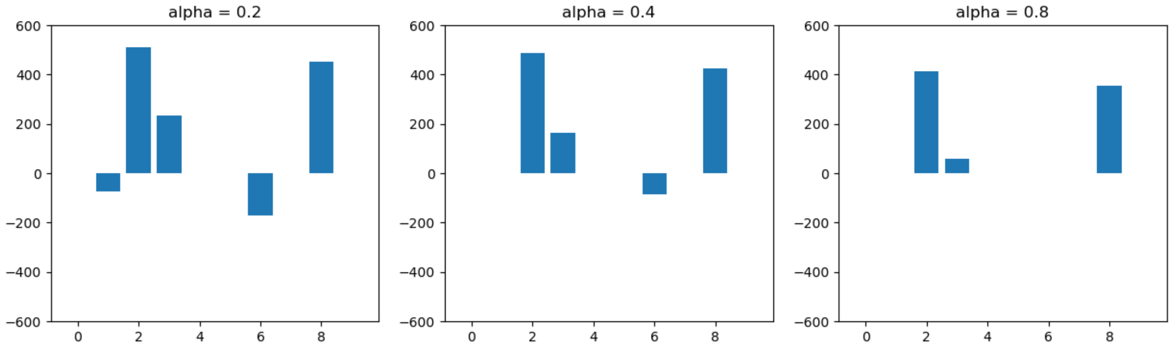

Figure 5: Parameters of Lasso solutions with different alpha

Regularization path

The regularization path of the Lasso solution is shown as **Figure 6**, which indicates that when α increased, the weight of parameters decreased until 0.

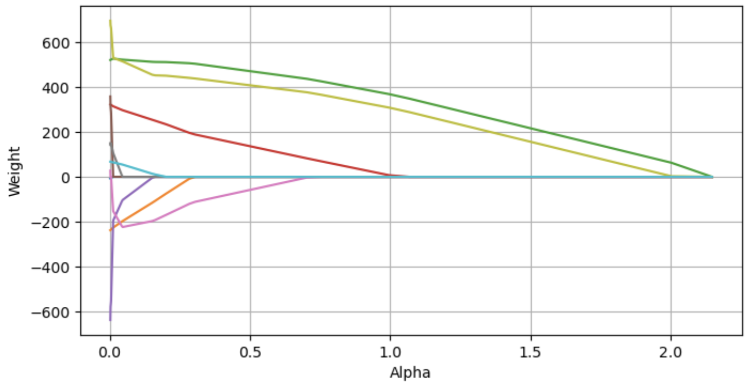

Figure 6: Regularization path

Solubility prediction

Figure 7 shows the comparison of the prediction results between training set and test set. The overfitting problem could be witnessed since the predictor works well on training data but not so well on test data. In order to reduce overfitting we need more data or reduce the number of parameters by using L1 regularization.

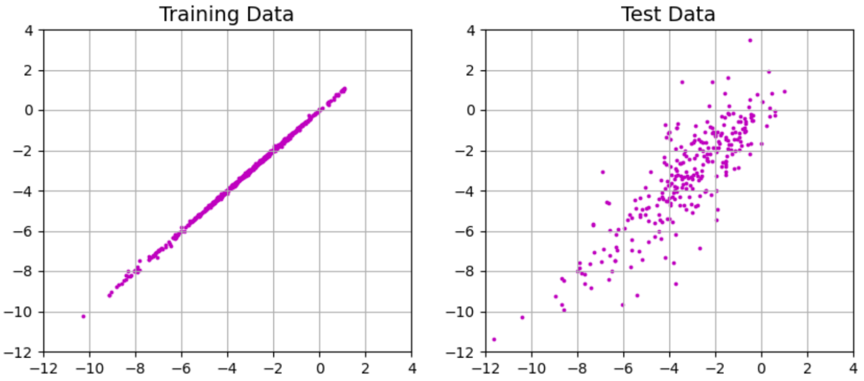

Figure 7: Prediction of the training set and test set

Figure 8 shows the prediction error on test set with increasing α. The error is computed by using mean squared error function. It is clear that with the increasing value of α, the error is increasing as well.

Figure 9 shows that with the increase of α, the number of non-zero coefficients are decreasing.

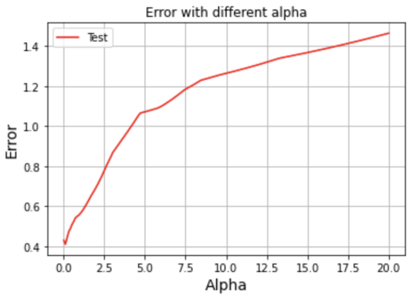

Figure 8: Prediction error on the test set

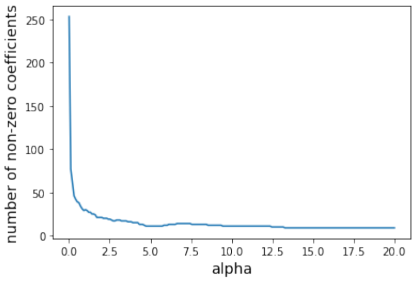

Figure 9: Number of non-zero weights

The top ten features I selected are SMTI, GMTIV, Wap ,IDMT, H_D/Dt, Wi_Dz(p), P_VSA_v_3, P_VSA_p_2, SAacc, Vx. The reason why I selected them is that they are the last ten features remained with non-zero weights during the process of increasing α. Since the features with non-zero weights are more meaningful, I believe these ten features could be better choices.
Figure 10 shows the parameters of these features.

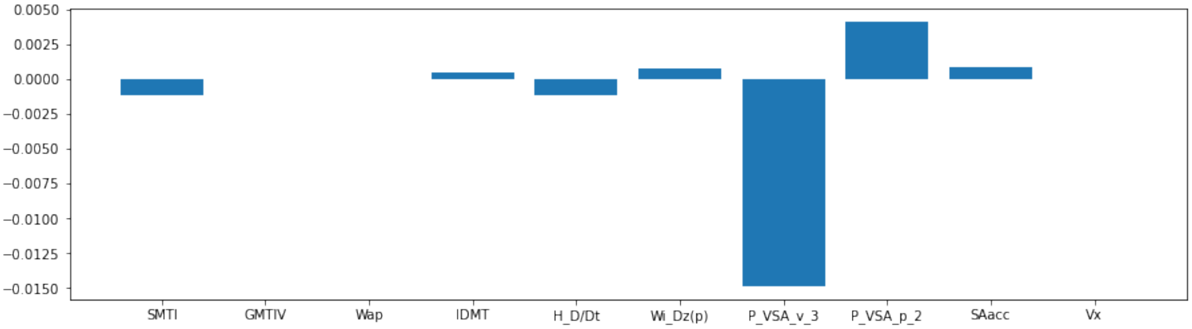

Figure 10: Parameters of the last ten features

And the error of predictor using L1 regularization and all features with $\alpha = 0.2$ is 0.47, meanwhile the error of predictor trained with selected top ten features using L2 regularization is 1.13. It is clear that the prediction error value of a predictor trained with selected ten features is higher.
From what I learn from the paper and during the whole process of lab, I think reducing the features of input data sometimes does not contribute to the performance of the model. And when we are using L1 regularization with a large value α as a way to prevent overfitting, it is likely that we could not get a more generalized model.