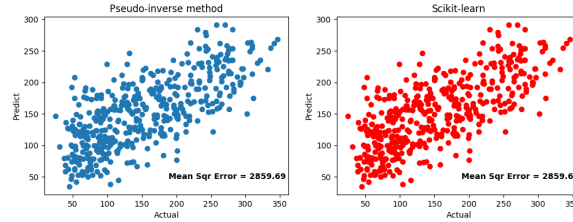# COMP6245 : Lab 4 Report

Thanakorn Panyapiang(tp2n19@soton.ac.uk)

1 Linear Least Square Regrssion

The result of the linear predictors using pseudo-inverse method and using scikit-learn is shown below.

Both methods give exactly the same result as can be seen from the identical mean square error.

2 Regularization

The gradient of L_2 regularization can be derived as follow:

$$
\begin{aligned}
\nabla_a E_{L2} &= \frac{\partial \|f - Ya\|_2^2 + \gamma\|a\|_2^2}{\partial a} \\
&= 2Y^T(f - Ya) + 2\gamma a
\end{aligned}
\tag{1}
$$

Equating the gradient to 0 gives the coefficient matrix a as follow

$$
\begin{aligned}
0 &= 2Y^T(f - Ya) + 2\gamma a \\
Y^T Ya - \gamma a &= Y^T f \\
a(Y^T Y - \gamma) &= Y^T f \\
a &= (Y^t Y - \gamma)^{-1} Y^T f
\end{aligned}
\tag{2}
$$

It can be observed that the coefficients of the predictor with L_2 regularization are significantly lower than the model without regularization. This is caused by adding a quadratic penalty of the weights to the error function. The coefficients of the model with and without L_2 regularization is shown below.
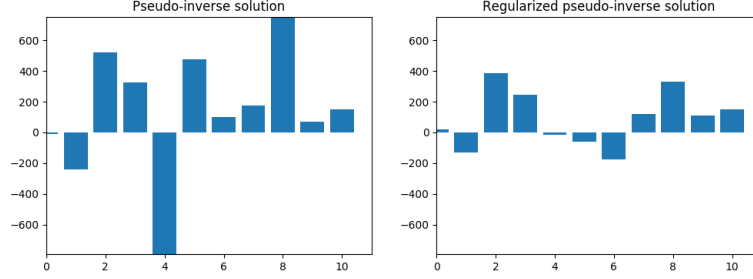
Figure 1

Moreover, when the regularization parameter(γ) increases, it can be noticed that the coefficients are reduced as can be seen on Figure 2.
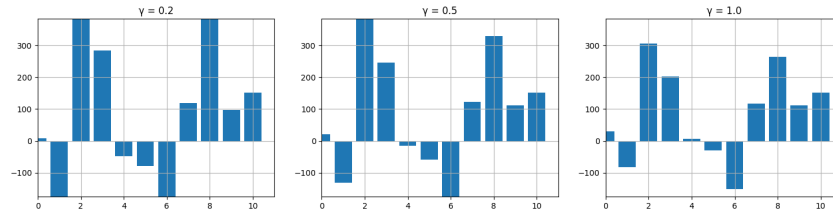

Figure 2

3 Sparse Regression

The coefficients of linear regression with Lasso regularization on different γ compare to Ridge regression is displayed in Figure 3.
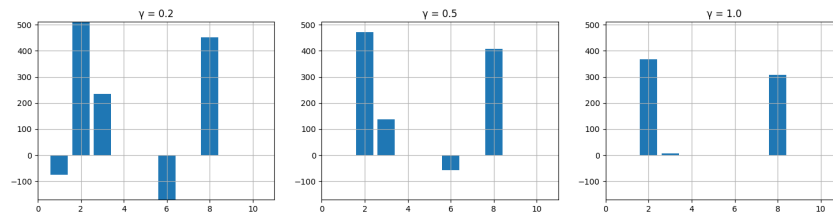

Figure 3

One observation on the lasso coefficients is that the numbers of zero coefficients increases along with γ.
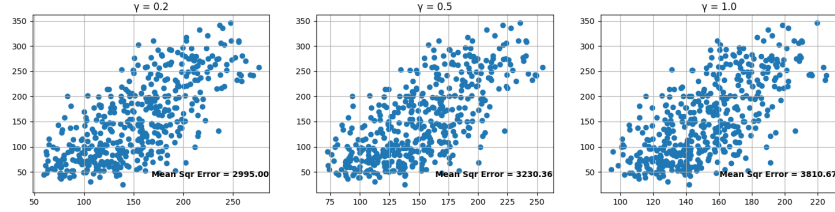
Figure 4

However, this coeeficients do not indicate that the attibutes which have 0 weight are not meaningful because the prediction error goes up as the numbers of zero coefficients increases as show in Figure 4. This suggests that, for the diabetes dataset, there are more useful features than the irrelevant ones so that when several parameters are removed the prediction becomes less accurate.

Lasso removes some useful features from the model because there are some features which are correlated to each other. Therefore, in order to minimize the penalty term$(\gamma||a||_1)$, Lasso chooses only one of these features. From the scatter plot below, it can be noticed that the 8th attribute, which has a non-zero coefficient, is positively correlated with the 0th and 4th attibutes, which have a zero coefficient.
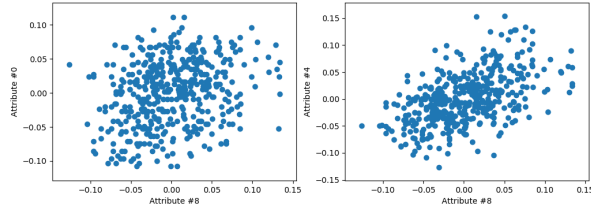

Figure 5

4 Solbility Prediction

The accuracy of linear regression with L2 regularization on solubility dataset is illustrated in the figure below.
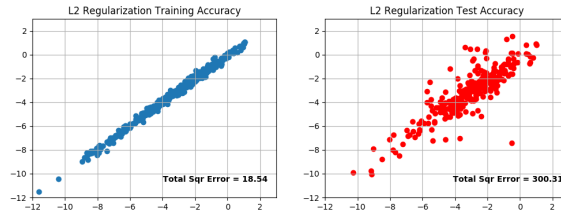

Figure 6

3

For Lasso, the models were slightly poorer on the training data but more accurate on the test set than L2 for all γ. The accuracy on test set and coefficients(non-zero only) are in the picture below. However, the total square error tends to increase together with the value of γ.
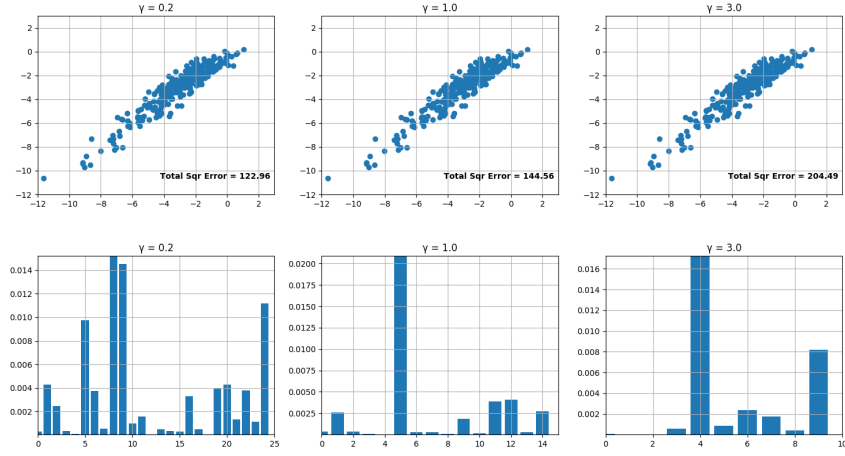

Figure 7

To compare the result with the Artificial Neural Network method proposed by Huuskonen, 160 and 50 records are randomly selected from the whole dataset to be a training set and a test set respectively. The result compare to the ANN result in the journal is shown in the scatter plot below. From the picture, it can be seen that the technique proposed by Huuskonen perform better on the test set.
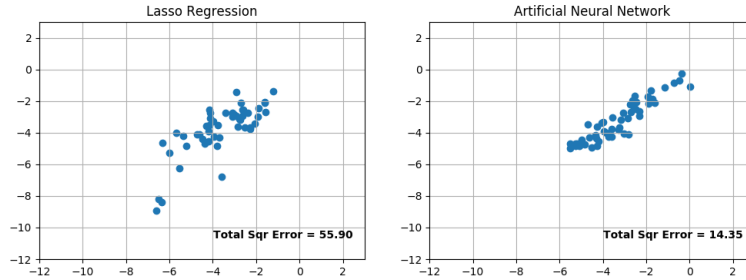

Figure 8

4