Lab Four: Linear Regression And Regularization

name: Huizhi Yu
number: ********
email: ********

1 Linear Least Squares Regression

In order to explore the diabetes dataset, we analyze the features and targets to understand their distributions and relationships. In **Figure 1**, the left column shows the target values' histogram. The right column shows a scatter plot to investigate the relationship between two selected features (feature 7 and feature 8).

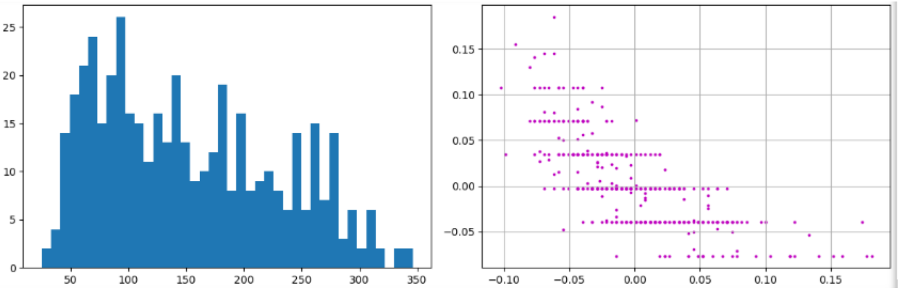

Figure 1: Distribution of Target Values and Feature Relationship in Diabetes Dataset

From the **histogram**, the distribution of targets indicates that most samples have relatively low diabetes indices, with only a few samples showing higher values. In the **scatter plot**, there appears to be no significant linear correlation between the two features. This lack of strong correlation implies that using these two features alone may not be effective for predicting the diabetes index.

We implement a linear regression model using the pseudo-inverse method to calculate the weight vector w, as shown in the following formula:

$$w = (X^T X)^{-1} X^T t$$

where X is the input matrix with size $N \times d$, and t is the target vector with size $N \times 1$. The weights calculated by this method can be used to generate predictions.

We also used the `LinearRegression` model from the `sklearn` library to solve the same linear regression problem and compared the prediction results of the two methods. From the **Figure 2**, we can see:

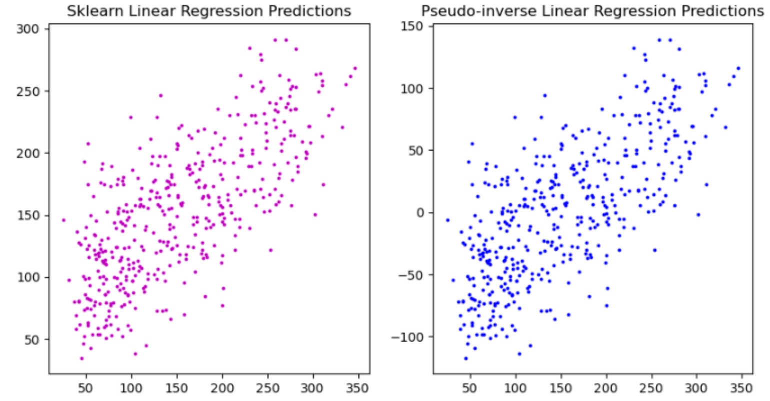

Figure 2: Comparison of Linear Regression Predictions Using Sklearn and Pseudo-inverse Method

The left graph shows the predictions from the `sklearn` model, and the right graph shows the predictions from the pseudo-inverse method. Looking at the distributions in both graphs, although there is some correlation between the predicted and actual values, they do not form a strict linear trend. This is especially true in the extreme high and low values, where there is considerable scattering. This suggests that there may be nonlinear characteristics in the data, and a simple linear model may not perfectly fit all data points.

2 Regularization

In L2 regularization, we add a regularization term, which is the sum of the squares of the weights w, to the MSE loss function:

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{d} w_j^2$$

where λ is the regularization parameter that controls the weight of the regularization term. A larger λ value will shrink the weights more strongly, making the model simpler, while a smaller λ reduces the effect of regularization. w_j is the weight of the j-th feature in the model.

By introducing the term $\lambda \sum_{j=1}^{d} w_j^2$, L2 regularization penalizes large weights, encouraging the model to choose smaller weights. This reduces the model's reliance on certain features, helping to prevent overfitting.

Figure 3 shows that some large weights, such as those for features 2, 4, and 8, decrease significantly. The direction of some feature weights also changes, such as feature 6 shifting from positive to negative. This indicates that regularization not only reduces the magnitude of the weights but also adjusts their direction to achieve a more balanced fit to the data. For features that initially have small weights, regularization nearly reduces their weights to zero, suggesting that these features contribute minimally to the model's predictions.

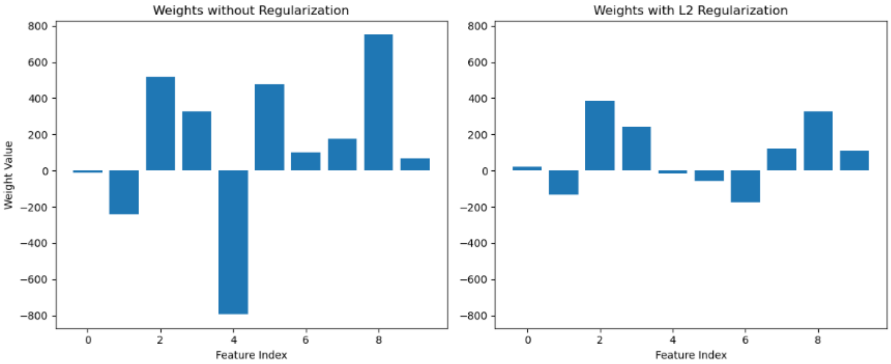

Figure 3: Effect of L2 Regularization on Feature Weights

3 Sparse Regression

To explore how L1 regularization (Lasso) affects model weights, we experimented with different regularization parameters α (0.1, 0.5, 1.0, 5.0, and 10.0). As shown in the **Figure 4**, as α increases, the weights in the model are gradually compressed to near zero, retaining only a few important features. Particularly, at higher values of α (5.0 and 10.0), almost all feature weights are zero, indicating that the model becomes overly sparse and may lose predictive power.

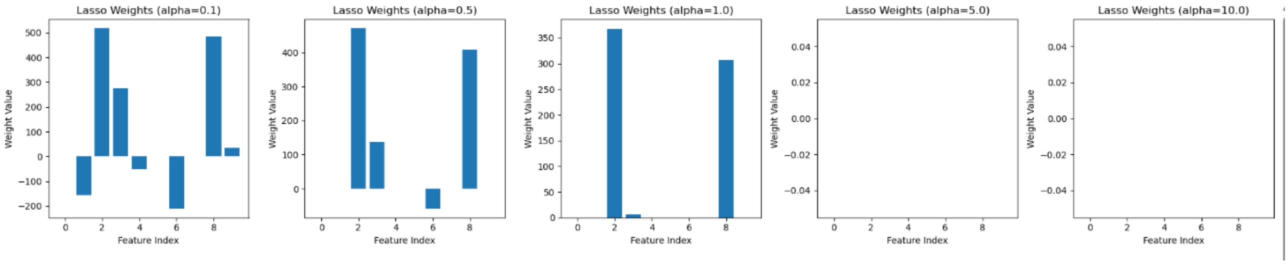

Figure 4: Feature Weight Distribution under Different L1 Regularization Parameters

To explore how L1 regularization affects the sparsity and predictive performance of the model, we experimented with different regularization parameters α (0.1, 0.5, 1.0, 5.0, and 10.0). The results are shown in the **Figure 5**:

As α increases, the number of non-zero weights in the model gradually decreases, indicating that L1 regularization effectively compresses the weights, retaining only a few important features. When α reaches higher values (e.g. 5.0 and 10.0), almost all weights are compressed to zero, suggesting that the model becomes overly sparse and may lose predictive power.

By comparing the mean squared error (MSE) under different α values, we observe that as α increases, the prediction error also gradually increases. Smaller α values (e.g. 0.1 and 0.5) allow more features to be retained, resulting in better model fit, while larger α values lead to reduced predictive ability as the model becomes overly simplified.

In the case of sparse regression, we observe that certain features retain non-zero weights even at higher α values, such as Feature 2, Feature 3, and Feature 8. According to the diabetes dataset documentation, these features represent medically significant factors, such as BMI and blood pressure, which are known to be closely related to diabetes progression. This suggests that L1 regularization is effective in selecting features that have a stronger relationship with the target variable, making these features more meaningful for predictive modeling.

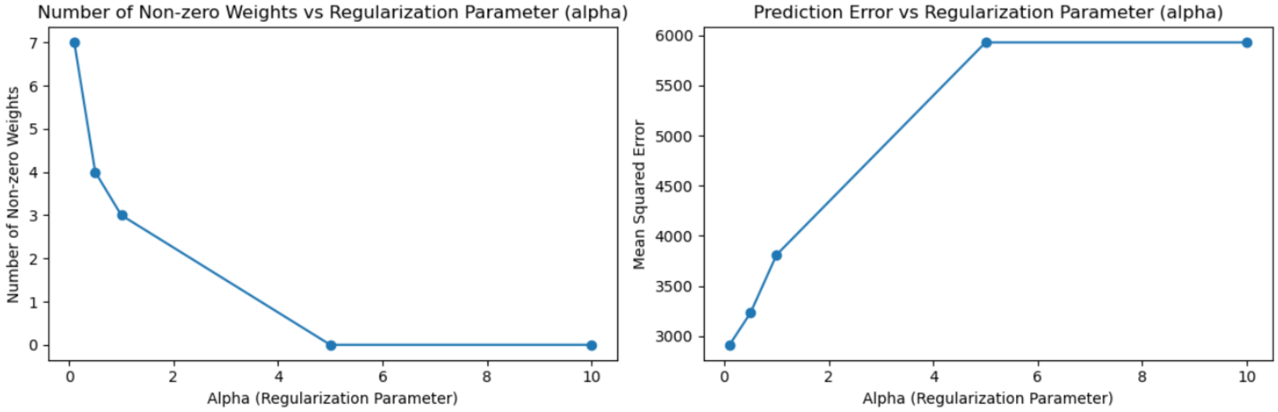

Figure 5: Effect of L1 Regularization Parameter on Number of Non-zero Weights and Prediction Error

We plot **Figure 6** to study the sparsity effect of L1 regularization. The horizontal axis represents the regularization parameter α, which increases gradually from left to right. When α is small, the regularization strength is low, and the feature weights in the model are relatively large. As α becomes larger, the regularization strength intensifies, causing L1 to compress more weights to zero. The vertical axis represents the regression weights of each feature. Positive coefficients indicate a positive correlation between the feature and the target variable, while negative coefficients indicate a negative correlation.

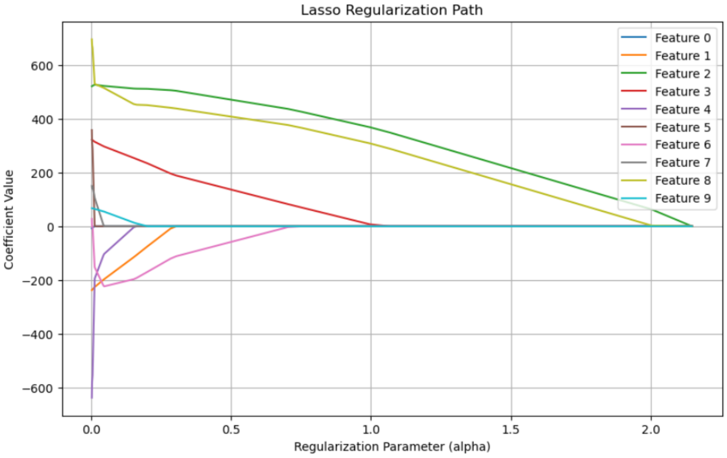

Figure 6: L1 Regularization Path

Each curve represents the path of a feature's weight as the value of α changes. As α increases, the regularization strength intensifies, causing many feature weights to be gradually compressed to zero. Features such as Feature 2, Feature 3, and Feature 8 retain relatively high non-zero values even at larger α values, indicating that these features are more important in the model. In contrast, features like Feature 0, Feature 1, and Feature 4 are compressed to zero early on, suggesting that these features have less impact on the model, and L1 regularization tends to eliminate them.

4 Solubility Prediction

4.1 Data loading and Model preparing

We loaded the data from the *Husskonen Solubility Features.xlsx* file and split the dataset into training and test sets, with the training set comprising 70% and the test set comprising 30% of the data. First, we implemented a linear regression model to fit the solubility features and generated predictions on both the training and test sets. For comparison purposes, we plotted scatter plots of the true vs. predicted solubility values for both the training and test sets (**Figure 7**).

To further control model complexity and avoid overfitting, we performed Lasso regularized regression (L1 regularization) on the data. By using different values of the regularization parameter α, we observed how the mean squared error (MSE) and the number of non-zero coefficients changed with α (**Figure 8**). Additionally,

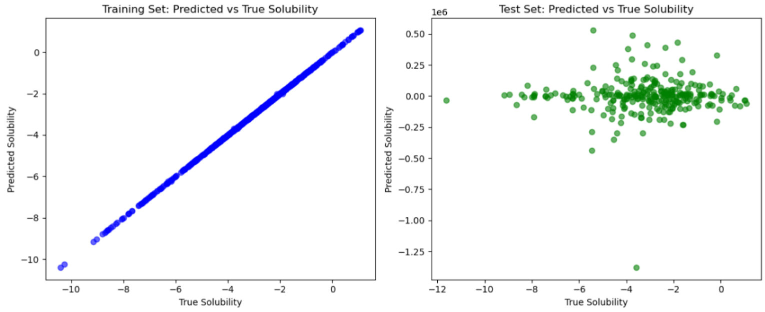

Figure 7: Scatter Plot of Predicted vs. True Solubility for Training and Test Sets

based on the results of Lasso regularization, we selected the top ten most important features affecting solubility and fitted a linear regression model using only these ten features. The predictive performance of this model was then compared with that of a Ridge regression model using all features.

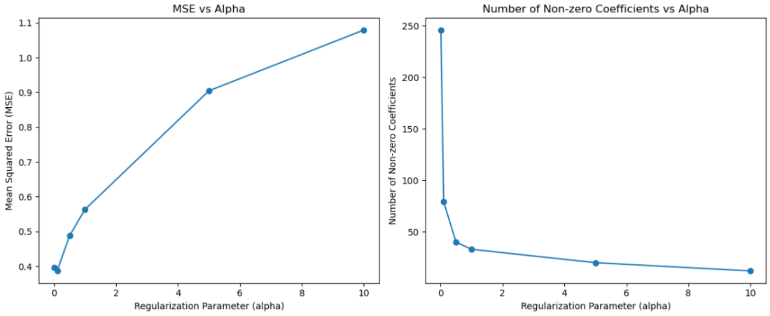

Figure 8: Effect of Regularization Parameter α on MSE and Number of Non-zero Coefficients in Lasso Regression

4.2 Results and Analysis

As shown in **Figure 7**, the linear regression model performs well on the training set, with the predicted values closely aligning with the true values along the diagonal line. However, on the test set, the prediction accuracy significantly decreases, suggesting that the model may be overfitting the training data.

Figure 8 shows that as the regularization parameter α increases, the mean squared error (MSE) on the test set also increases, indicating that excessive regularization can harm the predictive performance of the model. Meanwhile, the number of non-zero coefficients decreases with increasing α, demonstrating that Lasso regularization effectively compresses the weights of unimportant features to zero, achieving feature selection.

The prediction error (MSE) of the linear regression model using the top ten important features is **0.56**, while the MSE of the Ridge regression model using all features is **1.17**. This result suggests that the top ten important features are sufficient to explain the major variations in solubility, achieving a good balance between model complexity and predictive accuracy.

From the results of Lasso regularization, we can see that certain features retain non-zero coefficients even at higher levels of regularization, indicating that these features are more important for predicting solubility. Conversely, many features are compressed to zero at lower α values, suggesting that they have a minimal impact on the model's predictions.

4.3 Comparison with Previous Studies

The results of this experiment show that the top ten important features selected by Lasso regularization can achieve satisfactory solubility prediction performance without relying on all features. This method maintains high predictive accuracy while reducing model complexity. This is similar to the findings of Huuskonen et al., who also used specific structural features to improve solubility prediction accuracy.

Huuskonen et al. used a neural network model to predict solubility, achieving high predictive accuracy with an r^2 of 0.86 on the test set. They found that molecular topological features, such as E-state indices, are important in describing the structure of compounds. Our Lasso regularization approach also reveals similar key features, but employs a different regression method that is more interpretable and useful for feature selection.

Pirashvili et al. used Topological Data Analysis (TDA) to uncover the relationship between molecular structure and solubility. Their study emphasized that certain molecular structures, such as ring structures and chlorine content, are important for predicting solubility. In our Lasso regularization model, we observed a similar trend: certain features retained non-zero weights even at higher regularization parameters, indicating that they play an important role in solubility prediction. This feature selection capability allows us to achieve a good balance between model interpretability and predictive performance.