Lasso Regression

Problem Statement

In regression analysis, feature selection is often crucial for model interpretability and simplicity. To address this, we need an optimization framework that minimizes prediction errors while promoting sparsity in the model coefficients, effectively selecting the most relevant features for prediction.

Mathematical Formulation

Assume $\mathbf{X}$ is a matrix of feature vectors and $\mathbf{y}$ represents the target values corresponding to these feature vectors. We want to fit a linear model of the form:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w} + b$$

where $\hat{\mathbf{y}}$ is the predicted value, $\mathbf{w}$ is the vector of weights (coefficients), and b is the bias term.

The error between the predicted values $\hat{\mathbf{y}}$ and the actual values $\mathbf{y}$ can be measured using the residual sum of squares:

$$\text{Error} = \|\mathbf{X}\mathbf{w} + b - \mathbf{y}\|$$

To find the best fit, we minimize the squared error term:

$$\frac{1}{2n}\|\mathbf{X}\mathbf{w} + b - \mathbf{y}\|^2$$

where n is the number of observations.

However, if there are non-relevant features present in the data, the model may overfit. To mitigate overfitting and minimize the coefficients corresponding to the non-relevant features, we add a penalizing term, the L1 norm $\|\mathbf{w}\|_1$. Since the L1 norm penalizes smaller coefficients more than the L2 norm (used in Ridge regression), it tends to set the coefficients of non-relevant features to zero, promoting sparsity.

Hence, our optimization problem becomes:

$$\text{Minimize} \left(\frac{1}{2n}\|\mathbf{X}\mathbf{w} + b - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_1 \right)$$

where λ is the regularization parameter that controls the degree of sparsity.

Summary

In summary, Lasso regression combines the least squares error with L1 regularization to create a model that not only fits the data well but also promotes sparsity in the coefficients. This helps in identifying the most relevant features, leading to a more interpretable and simplified model. The key equation to minimize is:

$$\frac{1}{2n}\|\mathbf{X}\mathbf{w} + b - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_1$$

This approach effectively balances fitting the data with reducing the complexity of the model.