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 we have a dataset with n observations. Each observation consists of an input vector of features $\mathbf{x}_i$, a target value y_i, and potentially a noise term σ_i. We want to fit a linear model of the form:

$$\hat{y}_i = \mathbf{w}^T \mathbf{x}_i + b$$

where $\hat{y}_i$ is the predicted value for the i^{th} observation, $\mathbf{w}$ is the vector of weights (coefficients), b is the bias term, and $\mathbf{w}^T$ denotes the transpose of $\mathbf{w}$.

Important Note: While the formulation above shows the model for a single observation, Lasso regression typically deals with minimizing the error across all observations in the dataset. However, for a single observation, the summation doesn't strictly apply.

To calculate the total squared error for all observations, we can use the following:

$$\text{Error} = \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i + b - y_i)^2$$

Here, the summation () iterates over all n observations, and the squared error for each observation is calculated and summed to get the total error.

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

$$\frac{1}{2n} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i + b - y_i)^2$$

The rest of the formulation for Lasso regression with L1 regularization and the minimization problem remain the same as before:

$$\text{Minimize} \left(\frac{1}{2n} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i + b - y_i)^2 + \lambda \|\mathbf{w}\|_1 \right)$$

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, considering multiple observations, is:

$$\frac{1}{2n} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i + b - y_i)^2 + \lambda \|\mathbf{w}\|_1$$

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