Derivation of Lasso Regression Equations

Mohamed Hozien

December 8, 2024

Contents

1 Hypothesis Function

The hypothesis function models the linear relationship between the input features X and the target y. It is defined as:

$$\hat{y} = Xw + b$$

Where:

- X is the feature matrix of size $m \times n$, where m is the number of samples and n is the number of features.
- w is the weight vector of size $n \times 1$.
- b is the bias term, which can be included as part of w if a column of ones is appended to X.

For a single training sample $\mathbf{x}_i$, the predicted value $\hat{y}_i$ is:

$$\hat{y}_i = \sum_{j=1}^{n} w_j x_{ij} + b$$

For all m samples, it is written in matrix form as:

$$\hat{y} = Xw + b$$

2 Cost Function

The cost function for Lasso Regression combines the Mean Squared Error (MSE) with an L1 regularization penalty. It is defined as:

$$J(w) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 + \lambda \sum_{j=1}^{n} |w_j|$$

Where:

- $\hat{y}_i$ is the predicted value for the i-th sample.
- y_i is the true value for the i-th sample.
- λ is the L1 regularization strength.
- w_j represents the weight (parameter) of the model for the j-th feature.

The MSE term is:

$$\text{MSE} = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2$$

The L1 regularization penalty is:

$$\text{L1 Penalty} = \lambda \sum_{j=1}^{n} |w_j|$$

The total cost function is:

$$J(w) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 + \lambda \sum_{j=1}^{n} |w_j|$$

3 Gradient Descent

To update the weights w and bias b, we compute the gradients of the cost function $J(w)$ with respect to each parameter w_j.

3.1 Gradient of MSE

The derivative of the MSE with respect to the weight w_j is:

$$\frac{\partial J}{\partial w_j} = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - y_i) x_{ij}$$

In vectorized form, it is written as:

$$\nabla_w \, \text{MSE} = \frac{1}{m} X^T (\hat{y} - y)$$

3.2 Gradient of L1 Regularization

The derivative of the L1 regularization term $|w_j|$ with respect to w_j is:

$$\frac{d}{dw_j} |w_j| = \text{sign}(w_j)$$

Thus, the total gradient of the L1 penalty is:

$$\nabla_w \, \text{L1} = \lambda \, \text{sign}(w)$$

3.3 Total Gradient

The total gradient of the cost function $J(w)$ with respect to w is the sum of the gradients of the MSE and the L1 penalty:

$$\nabla_w J = \nabla_w \mathrm{MSE} + \nabla_w \mathrm{L1}$$

Substituting the two gradients, we get:

$$\nabla_w J = \frac{1}{m} X^T(\hat{y} - y) + \lambda \,\mathrm{sign}(w)$$

4 Weight Update Rule

Using the gradient descent algorithm, the weight w is updated as follows:

$$w = w - \alpha \nabla_w J$$

Where α is the learning rate. Substituting the total gradient of $J(w)$, we get:

$$w = w - \alpha \left(\frac{1}{m} X^T(\hat{y} - y) + \lambda \,\mathrm{sign}(w) \right)$$

5 Summary of Key Equations

Here is a summary of the key equations for Lasso Regression.

5.1 Hypothesis

$$\hat{y} = Xw + b$$

5.2 Cost Function

$$J(w) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 + \lambda \sum_{j=1}^{n} |w_j|$$

5.3 Gradient of MSE

$$\nabla_w \mathrm{MSE} = \frac{1}{m} X^T(\hat{y} - y)$$

5.4 Gradient of L1 Regularization

$$\nabla_w \mathrm{L1} = \lambda \,\mathrm{sign}(w)$$

5.5 Total Gradient

$$\nabla_w J = \frac{1}{m} X^T (\hat{y} - y) + \lambda \, \mathrm{sign}(w)$$

5.6 Weight Update Rule

$$w = w - \alpha \, \nabla_w J$$