Derivation of Lasso and Ridge Regression Equations

Mohamed Hozien

December 8, 2024

Contents

1 Lasso Regression

1.1 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.

1.2 2. Cost Function

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

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

1.3 3. Gradient Descent

The total gradient of the cost function $J(w)$ with respect to w is:

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

Using gradient descent, the weight w is updated as:

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

1.4 4. Summary of Key Equations

- **Hypothesis:** $\hat{y} = Xw + b$
- **Cost Function:** $J(w) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 + \lambda \sum_{j=1}^{n} |w_j|$
- **Gradient of MSE:** $\nabla_w \, \mathrm{MSE} = \frac{1}{m} X^T (\hat{y} - y)$

- **Gradient of L1 Regularization:** $\nabla_w \, \mathrm{L1} = \lambda \, \mathrm{sign}(w)$
- **Total Gradient:** $\nabla_w J = \frac{1}{m} X^T (\hat{y} - y) + \lambda \, \mathrm{sign}(w)$
- **Weight Update Rule:** $w = w - \alpha \, \nabla_w J$

2 Ridge Regression

2.1 1. Objective of Linear Regression

The objective of standard linear regression is to minimize the cost function:

$$J(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$$

where:

- $\mathbf{X}$ is the input matrix of size $n \times m$ (n samples, m features)
- $\mathbf{y}$ is the target vector of size $n \times 1$
- $\mathbf{w}$ is the weight vector of size $m \times 1$

2.2 2. Ridge Regression Objective

Ridge regression modifies the linear regression cost function by adding an L_2-norm regularization term:

$$J_{\text{ridge}}(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

where $\lambda > 0$ is the regularization strength.

2.3 3. Augmented Input Matrix

To include the bias term b, we augment $\mathbf{X}$ with a column of ones:

$$\mathbf{X}_{\text{aug}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$$

2.4 4. Regularized Normal Equation

The ridge regression solution is derived by setting the gradient of the cost function to zero:

$$\nabla J_{\text{ridge}}(\mathbf{w}) = -\mathbf{X}\text{aug}^T(\mathbf{y} - \mathbf{X}\text{aug}\mathbf{w}) + \lambda\mathbf{w} = 0$$

Solving for $\mathbf{w}$, we get:

$$\mathbf{w} = (\mathbf{X}\text{aug}^T\mathbf{X}\text{aug} + \lambda\mathbf{I})^{-1}\mathbf{X}_{\text{aug}}^T\mathbf{y}$$

where:

- $\mathbf{I}$ is the identity matrix of size $(m+1) \times (m+1)$
- The top-left entry of $\mathbf{I}$ is set to 0 to avoid regularizing the bias term b

2.5 5. Prediction Equation

For a new input $\mathbf{X}_{\text{new}}$, the predicted output is:

$$\hat{\mathbf{y}} = \mathbf{X}_{\text{new, aug}} \cdot \mathbf{w}$$

where $\mathbf{X}_{\text{new, aug}}$ includes an additional column of ones for the bias.

2.6 6. Summary of Key Equations

- **Linear Regression Cost:** $J(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$
- **Ridge Regression Cost:** $J_{\text{ridge}}(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$
- **Augmented Form:** $\mathbf{X}_{\text{aug}}$ adds a bias term to $\mathbf{X}$
- **Ridge Regression Solution:** $\mathbf{w} = (\mathbf{X}\text{aug}^T\mathbf{X}\text{aug} + \lambda\mathbf{I})^{-1}\mathbf{X}_{\text{aug}}^T\mathbf{y}$
- **Prediction:** $\hat{\mathbf{y}} = \mathbf{X}_{\text{new, aug}} \cdot \mathbf{w}$