Derivation of Lasso, Ridge, and Elastic Net Regression Equations

Mohamed Hozien (202201507) Youssef Tarek (202201545)
Ziad Moataz (202201252)

December 10, 2024

Contents

1 Lasso Regression

Mohamed Hozien

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:

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

Using gradient descent, the weight w is updated as:

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

1.4 4. Pros and Cons

Pros:

- **Feature Selection:** Lasso can drive some feature coefficients exactly to zero, effectively performing feature selection. This is particularly useful when dealing with high-dimensional data.
- **Sparse Model:** It produces a sparse model with fewer non-zero coefficients, which can be beneficial for interpretability and performance in some cases.

Cons:

- **Less Stable with Highly Correlated Features:** Lasso can behave unpredictably when there are highly correlated features, as it tends to randomly select one of the correlated features and discard the rest.
- **Underfitting with Large Numbers of Features:** If the number of predictors is much larger than the number of observations, Lasso may underperform because it forces many coefficients to be exactly zero, possibly missing important relationships.

2 Ridge Regression

Youssef Tarek

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:

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

2.3 3. Regularized Normal Equation

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

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

2.4 4. Pros and Cons

Pros:

- **No Feature Selection:** Ridge regression doesn't set coefficients to zero, but it shrinks them, making it suitable when all features are believed to contribute to the model.
- **Stable with Correlated Features:** Ridge tends to perform better than Lasso when dealing with multicollinearity (correlated features), as it keeps all features in the model.

Cons:

- **Lacks Feature Selection:** While Ridge shrinks coefficients, it doesn't perform feature selection, so it may still include many irrelevant features in the model.
- **Model Interpretation:** Since coefficients are never zero, the model can be harder to interpret when the number of features is large.

3 Elastic Net Regression

Ziad Moataz

3.1 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.

3.2 2. Cost Function

The cost function for Elastic Net combines the Mean Squared Error (MSE) with L1 (Lasso) and L2 (Ridge) regularization penalties. It is defined as:

$$J(w) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$$

Where:

- $\|w\|_1 = \sum_{j=1}^{n} |w_j|$ is the L1 norm of w
- $\|w\|_2^2 = \sum_{j=1}^{n} w_j^2$ is the squared L2 norm of w
- λ_1 and λ_2 are hyperparameters controlling the L1 and L2 penalties, respectively

3.3 3. Gradient of Cost Function

The gradient of the cost function with respect to w is:

$$\frac{\partial J}{\partial w} = \frac{1}{m} X^T (\hat{y} - y) + \lambda_1 \operatorname{sign}(w) + 2\lambda_2 w$$

3.4 4. Gradient of Cost Function with Respect to b

The gradient with respect to b is:

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

3.5 5. Update Rules

The weights w and bias b are updated using gradient descent as follows:

$$w = w - \alpha \frac{\partial J}{\partial w}$$

$$b = b - \alpha \frac{\partial J}{\partial b}$$

Where α is the learning rate.

3.6 6. Key Equations Summary

- **Hypothesis Function:** $\hat{y} = Xw + b$
- **Cost Function:** $J(w) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$
- **Gradient of Cost Function:** $\frac{\partial J}{\partial w} = \frac{1}{m} X^T (\hat{y} - y) + \lambda_1 \operatorname{sign}(w) + 2\lambda_2 w$

3.7 7. Pros and Cons

Pros:

- **Feature Selection and Shrinking:** Elastic Net combines the strengths of both Lasso and Ridge, allowing for both feature selection and coefficient shrinking.
- **Better Performance with Highly Correlated Features:** Elastic Net performs better than Lasso when there are many correlated features.

Cons:

- **Model Complexity:** The model has two regularization parameters, making it more complex and requiring more careful tuning.
- **Computational Cost:** The addition of both L1 and L2 penalties may increase computational cost compared to Lasso and Ridge.

theoretical Information are included in the Readme file on Github https://github.com/Mohamed-Mohamed-Hozien/Math-303-Report-2