Regression Methods

Oliver Zhao

1 Introduction

1.1 Regression Models

Regression models take in n independent variables represented by $\mathbf{X}$. A linear hyperplane described by the weight vector $\mathbf{w} = \{w_0, w_1, \ldots, w_n\}$ separates the feature space to generate the predicted value y_i of the ith data point.

$$y_i = w_0 + w_1 X_{1,i} + w_2 X_{2,i} + \cdots + w_n X_{n,i}. \tag{1.1}$$

1.2 Solving for Coefficients

The ordinary least squares (OLS) method can provide a precise solution. For example, in a data set with only one independent variable, where the vector $\mathbf{x} = \{x_1, \ldots, x_p\}$ contains p data entries, the two parameters α and β in the model

$$y_i = \alpha + \beta x_i + \epsilon_i, \tag{1.2}$$

Where ϵ_i is the error of the ith data entry x_i, can be calculated with

$$\begin{aligned} \beta &= \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \frac{\mathrm{Cov}[x, y]}{\mathrm{Var}[x]} \\ \alpha &= \bar{y} - \beta \bar{x} \end{aligned} \tag{1.3}$$

This method can be extended to a multi-variable case where matrix operations are used to derive the model parameter vector containing the coefficients for each feature. Sometimes the number of features are so numerous that gradient descent is used in lieu of the OLS method.

2 Types of Linear Regression

2.1 Linear Regression

This is often done through minimizing the mean squared error where the ideal weight vector $\hat{\mathbf{w}}$ is

$$\hat{\mathbf{w}} = \min_{\mathbf{w}} ||\mathbf{X}\mathbf{w} - \hat{\mathbf{y}}||^2, \tag{2.1}$$

Where $\hat{\mathbf{y}}$ is the ground truth.

2.2 Ridge Regression

Ridge Regression attempts introduces a shrinkage parameter λ_2, where the optimal weight vector $\hat{\mathbf{w}}$ is calculated as

$$\hat{\mathbf{w}} = \min_{\mathbf{w}} (||\mathbf{X}\mathbf{w} - \hat{\mathbf{y}}||^2 + \lambda_2 ||\mathbf{w}||^2). \tag{2.2}$$

2.3 Lasso Regression

Meanwhile, Lasso Regression is similar to Ridge Regression, except that the absolute size of the regression coefficients is penalized instead, or

$$\hat{\mathbf{w}} = \min_{\mathbf{w}} (||\mathbf{X}\mathbf{w} - \hat{\mathbf{y}}||^2 + \lambda_1 ||\mathbf{w}||). \tag{2.3}$$

2.4 ElasticNet Regression

ElasticNet Regression combines the Ridge Regression and Lasso Regression methods, where λ_1 and λ_2 are adjusted to allow for varying extents of Ridge Regression and Lasso Regression behavior.

$$\hat{\mathbf{w}} = \min_{\mathbf{w}} (||\mathbf{X}\mathbf{w} - \hat{\mathbf{y}}||^2 + \lambda_1 ||\mathbf{w}|| + \lambda_2 ||\mathbf{w}||^2). \tag{2.4}$$

3 Note on Regularization

It is worth noting that the shrinkage parameters λ_1 in Lasso Regression and λ_2 in Ridge Regression are forms of regularization, frequently called L1 and L2 regularization, respectively.

3.1 L1 Regularization

Based on Equation (2.3), we see that the regularization parameter λ_1 tends to push $\mathbf{w}$ towards $\mathbf{0}$, which can be crudely viewed as reducing the number of features in a model. In short, L1 regularization can be interpreted as an implicit feature selection method. Thus, it can be used to address overfitting by reducing model complexity.

3.2 L2 Regularization

In contrast to L1 regularization, the regularization parameter λ_2 in Equation (2.2) does not push the coefficient vector $\mathbf{w}$ to zero. Thus, while it does shrink the value of the coefficients, it does not have a feature selection mechanism. Instead, it is used to reduce the feature parameter variance, which can be used to solve multicollinearity problems.