Shrinkage Methods in Linear Regression

June 10, 2018

Benno Geißelmann

1 Motivation

Linear Regression has the aim to predict a dependent target variable by a combination of predictor variables (features) together with their according coefficients. The linear regression function is of the following form:

$$\hat{y}_i = w_0 + w_1 x_{i1} + \cdots + w_p x_{ip} + \varepsilon_i \tag{1}$$

where y_i is the ith target value ($\hat{y}_i$ is the ith predicted value) and x_i is the according ith feature vector. $w_0 \cdots w_p$ are the p coefficients of the regression function which need to be determined. ε_i is some random noise. P is the number of features and N is the number of training data. In some scenarios P can be very large, sometimes even larger than N. Therefore in these scenarios there needs to be performed some kind of feature selection in order to shrink the number of used features for the regression model. Otherwise the huge amount of predictor variables would lead to overfitting on the training data and therefore modelling the noise of the data rather than the underlying connections. By shrinking of the coefficients, the variance in the model is reduced. At the same time the bias of the model is increased (bias-variance-dilemma) with the aim to build a model which is able to perform well on unseen data. Lasso and Ridge regression are methods to shrink the amount of features or their according coefficients. In the next sections the concepts of Ridge and Lasso is introduced.

2 The Ridge model

The Rigde regression analysis method defines an approach by which it is possible to shrink the size of the coefficients of the regression function. Due to

this, overfitting is avoided and therefore the ability for generalization beyond the traing set of the regression model is improved. The optimization goal which Ridge is trying to achieve can be formulated like this:

$$\min_{w_0,w} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - w_0 - x_i^T w)^2 + \lambda \sum_{j=1}^{P} w_j^2 \right\} \tag{2}$$

The first part of the sum is the ordinary least squares (OLS) approach and in fact if you set $\lambda = 0$ you get OLS. The second part of the sum (penatly term) is introduced by Ridge. It defines a further restriction on the sum over all coefficients, weighted by some parameter λ. By λ there can be handled how big the influence of the penalty term is for the optimization. The bigger it is, the more influence the penalty term is going to have in the optimization.

3 The Lasso model

Lasso (least absolute shrinkage and selection operator) is comparable to Ridge except the difference, that the penatly term uses the absolute value instead of the square number:

$$\min_{w_0,w} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - w_0 - x_i^T w)^2 + \lambda \sum_{j=1}^{P} |w_j| \right\} \tag{3}$$

4 Difference between Ridge and Lasso Model

As stated before the difference between the Ridge approach and the Lasso approach the the usage of square number respectively the absolute value. Due to this, the Ridge approach tends to create small but non null coefficients (because squares of numbers less than 1 are small but bigger than 0) where as the Lasso approach tends to set more coefficients to 0. Therefore Lasso is stronger in doing actual feature selection. This might be an important reason why Lasso is the more popular approach for shrinkage in linear regression today.

5 Gradient descent for optimization of the Ridge approach

For finding a feasable solution for the above described optimization problem of the Ridge regression, there can be used gradient descent. The following algorithm describes this approach:

Algorithm 1: Gradient Descent for Ridge Regression

Input: iterations
Input: η
Result: the coefficients $w_0 \cdots w_n$
$Cost \leftarrow \frac{1}{N} \sum_{i=1}^{N} (y_i - w_0 - x_i^T w)^2 + \lambda \sum_{j=1}^{P} w_j^2$;
for *it in interations* **do**
  for w_i *in* w **do**
    $w_i \leftarrow w_i - \eta \frac{\partial Cost}{\partial w_i}$
  end
end

The algorithm gets the number of iterations and the step size η as input. In each iteration there is calculated the partial derivative of the cost function and a certain coefficient w_i. Using the negative value of this derivative multiplied by the stepsize, we are descending on the cost function.

6 Coordinate descent for optimization of the Lasso approach

Due to the usage of the absolute value in the optimization function of Lasso we cannot directly use gradient descent because the function is not differentiable at $w_j = 0$. There is an alternative approach for approximating a solution for this optimization problem called coordinate descent.