Hongjue Zhao
3190104515
hongjue0830@zju.edu.cn

Optimization Midterm

Hongjue Zhao

Contents

1 Requirement **1**

2 Introduction to Ridge Regression **1**

2.1 Ordinary Least squares Method 1

2.2 Ridge Regression 2

3 Optimization Algorithms **3**

3.1 Gradient Method 3

3.2 Conjugate Gradient Method 3

3.3 quasi-Newton Method 3

1 Requirement

Based on small datasets *abalone*, *bodyfat* and *housing*, training a *ridge regression* model with algorithms Gradient Descent, Conjugate Descent, and quasi-Newton method respectively.

Requirements:

1. Implement algorithms (Gradient Descent, Conjugate Descent, and quasi-Newton method) using C/C++ programming language.
2. In this answer sheet, please briefly introduce
   (a) The ridge regression;
   (b) The training algorithms that you implement;
   (c) Experimental settings;
   (d) Experimental results.

   Please illustrate with diagrams, the Mean Squared Error of different algorithms at each iteration, and analyze the results in light of what you have learned in this course.
3. Compress the pdf file of this answer sheet and the C/C++ program into a zip file, and submit it.

Remark. Datasets can be downloaded from link.

2 Introduction to Ridge Regression

2.1 Ordinary Least squares Method

Suppose that we have input vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^\top$ and want to predict a real-valued output y. The linear regression model has the from

$$f(\boldsymbol{x}) = \beta + \sum_{j=1}^{p} x_j w_j = \langle \boldsymbol{w}, \boldsymbol{x}' \rangle , \tag{2.1}$$

in which $\boldsymbol{w} = (\beta, w_1, \ldots, w_p)^\top \in \mathbb{R}^{p+1}$ and $\boldsymbol{x}' = (1, x_1, \ldots, x_p)^\top \in \mathbb{R}^{p+1}$. The linear model either assumes that the regression function is linear, or that the linear model is reasonable approximation. Here $\boldsymbol{w}$ is unknown parameter vector.

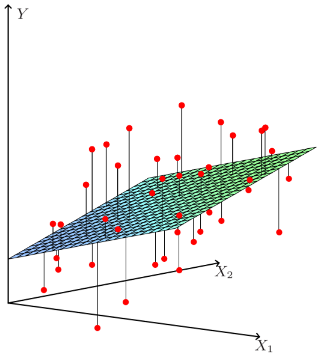

Figure 1: Linear model fitting with $\boldsymbol{x} \in \mathbb{R}^2$.

Generally speaking we have a training dataset with training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$ from which we would like to estimate the parameter vector $\boldsymbol{w}$, and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$. If we use *Ordinary Least Squares* (OLS) method, we will pick the $\boldsymbol{w}$ which can minimize *the residual sum of squares* (RSS)

$$\begin{aligned} \mathrm{RSS}(\boldsymbol{w}) &= \sum_{i=1}^{N} (y_i - f(\boldsymbol{x}_i))^2 \\ &= \sum_{i=1}^{N} \left(y_i - \beta - \sum_{j=1}^{p} x_{ij} w_j \right)^2 . \end{aligned} \tag{2.2}$$

Denote by $\boldsymbol{X} \in \mathbb{R}^{N \times (p+1)}$ the matrix with each row an input vector, and similarly, let $\boldsymbol{y} \in \mathbb{R}^N$ be the output vector in this training dataset. Then we can rewritte Eq (2.2) in as

$$\mathrm{RSS}(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \tag{2.3}$$

This is a quadratic function. Differentiating with respect to $\boldsymbol{w}$ we can get

$$\begin{aligned} \frac{\partial \mathrm{RSS}}{\partial \boldsymbol{w}} &= -2\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \\ \frac{\partial^2 \mathrm{RSS}}{\partial \boldsymbol{w} \partial \boldsymbol{w}^\top} &= 2\boldsymbol{X}^\top \boldsymbol{X}. \end{aligned} \tag{2.4}$$

Suppose that $\boldsymbol{X}$ has full column rank. Therefore, $\boldsymbol{X}^\top \boldsymbol{X}$ is positive definite. We can set the first derivative to 0

$$\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = 0 \tag{2.5}$$

to get the unique solution

$$\hat{\boldsymbol{w}}_{\mathrm{OLS}} = \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{2.6}$$

2.2 Ridge Regression

However, according to Eq (2.6), $\hat{\boldsymbol{w}}_{\mathrm{OLS}}$ depends on $\boldsymbol{X}^\top \boldsymbol{X}$. In some cases, $\boldsymbol{X}^\top \boldsymbol{X}$ may be *singular* or *nearly singular*. When there are many correlated variables in a linear regression model, their coefficients can become poorly determined and exhibit high variance. In those cases, we call $\boldsymbol{X}$

ill-conditioned. Small changes to elements of $\boldsymbol{X}$ will lead to large changes in $\boldsymbol{X}^\top \boldsymbol{X}$. In addition, $\hat{\boldsymbol{w}}_{\text{OLS}}$ may provide a good fit to the training data, but it will not fit sufficiently well to the test data.

In order to alleviate this problem, we introduce *ridge regression*. Ridge regression shrinks the regression coefficients by imposing a penalty on their size. The ridge coefficients minimize a penalized residual sum of squares,

$$\hat{\boldsymbol{w}}_{\text{ridge}} = \arg\min_{\boldsymbol{w}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta - \sum_{j=1}^{p} x_{ij} w_j \right)^2 + \lambda \sum_{j=1}^{p} w_j^2 \right\}. \tag{2.7}$$

Here $\lambda \geq 0$ is a *complexity parameter* that controls the amount of shrinkage.

Rewriting the criterion in Eq (2.7) in matrix form, we can obtain

$$\text{RSS} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \lambda \boldsymbol{w}^\top \boldsymbol{w}, \tag{2.8}$$

and

$$\begin{aligned} \frac{\partial \text{RSS}}{\partial \boldsymbol{w}} &= -2\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + 2\lambda \boldsymbol{w} \\ \frac{\partial^2 \text{RSS}}{\partial \boldsymbol{w} \partial \boldsymbol{w}^\top} &= 2\boldsymbol{X}^\top \boldsymbol{X} + 2\lambda. \end{aligned} \tag{2.9}$$

Let the first derivative in Eq (2.9) be zero, then the ridge regression solution can be expressed as

$$\hat{\boldsymbol{w}}_{\text{ridge}} = \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \tag{2.10}$$

where $\boldsymbol{I}$ is the $(p+1) \times (p+1)$ *identity matrix*. The solution adds a positive constant to the diagonal of $\boldsymbol{X}^\top \boldsymbol{X}$ before inversion, which makes this problem nonsingular, even if $\boldsymbol{X}^\top \boldsymbol{X}$ is not of full rank.

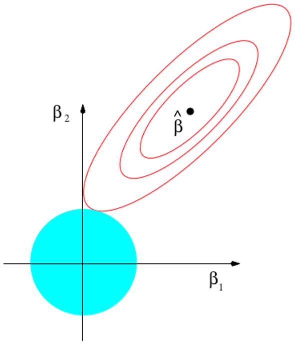

Figure 2: Estimation picture for ridge regression.

3 Optimization Algorithms

3.1 Gradient Method

3.2 Conjugate Gradient Method

3.3 quasi-Newton Method

Algorithm 1: Gradient Method

```
Input   : Objective function $f(\boldsymbol{x})$, gradient function $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$ and accuracy $\varepsilon$.
Output: Local minimum of $f(\boldsymbol{x})$: $\boldsymbol{x}$.
Initialization: Set $k = 0$ and initialize $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$.
while True do
    Calculate $f(\boldsymbol{x}^{(k)})$.
    Calculate gradient $\boldsymbol{g}_k = \boldsymbol{g}(\boldsymbol{x}^{(k)})$.
    if $\|\boldsymbol{g}_k\| < \varepsilon$ then
        Stop iteration and Let $\boldsymbol{x}^* = \boldsymbol{x}^{(k)}$.
    else
        Let $\boldsymbol{p} = -\boldsymbol{g}(\boldsymbol{x}^{(k)})$
    end
end
```