MAT 258A, Final Project

Lasso Solvers: Quadratic Programming and LARS

Yilun Zhang
999486337

Abstract

Lasso[1], proposed by Tibshirani in 1996, is short for least absolute shrinkage and selection operator. It can estimate the regression coefficients and select them at the same time. But LASSO wasn't popular when it is invented. Because this method involves absolute constraint which is not derivable. Methods and computational power at that time can't solve lasso problem in a short time. In this project I will first show under modification, lasso can be rewritten into standard quadratic programming problem and then introduce the least angle regression method (LARS).

1 The Lasso Problem

Robert Tibshirani's 1996 paper R[1], introduces the lasso which can do the parameter estimation and selection at the same time. It sacrifice a little bias but greatly reduced prediction error often occurs in ordinary least square estimator, retains nice properties of both ridge and subset selection.

The most part of the paper introduces lasso based on a linear regression model. Suppose $(\boldsymbol{x}^i, y_i),\ i = 1, 2, \cdots, N$ are the data where y has mean 0 and $\boldsymbol{x}$ is normalized such that $\sum_i x_{ij}/N = 0$ and $\sum_i x_{ij}^2/N = 1$. Different from OLS(ordinary least square), lasso estimator is obtained by minimizing

$$\hat{\boldsymbol{\beta}} = argmin\{\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2\} \tag{1}$$

with constraint

$$||\boldsymbol{\beta}||_1 = \sum_j |\beta_j| \leq t \tag{2}$$

t is a prespecified parameter. By Lagrange form, this is equivalent to minimizing

$$\hat{\boldsymbol{\beta}} = argmin\{\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2 + \lambda||\boldsymbol{\beta}||_1\} \tag{3}$$

This means put a L_1 penalty on the coefficients. The small t or large λ will cause many coefficient set to be 0 by the property of L_1 norm. This will do the predictor selection automatically. This will be benefit if we believe the underlying model has sparse representation.

2 Quadratic Programming

Many previous works (such as Busa J, 2012 [2]) proves a simple method that can rewrite lasso into a standard quadratic programming problem that no absolute operation involves.

Consider

$$\begin{array}{ll} minimize_{\beta} & ||Y - X\boldsymbol{\beta}||_2^2 \\ subject\,to & ||\boldsymbol{\beta}||_1 < t \end{array}$$

this is equivalent to

$$\begin{array}{ll} minimize_{\beta} & \frac{1}{2}\boldsymbol{\beta}^T H\boldsymbol{\beta} + f^T\boldsymbol{\beta} \\ subject\,to & ||\boldsymbol{\beta}||_1 < t \end{array}$$

where $H = 2X^TX$, $f = -2X^TY$. Then let

$$\beta_i = \beta_i^+ - \beta_i^-$$

and

$$\beta_i^+ = \frac{|\beta_i| + \beta_i}{2} \quad \beta_i^- = \frac{|\beta_i| - \beta_i}{2}$$

obviously, $|\beta_i| = \beta_i^+ + \beta_i^-$. Then we have

$$\begin{array}{ll} minimize_{\beta} & \frac{1}{2}(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-)^T H(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-) + f^T(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-) \\ subject\,to & \sum_i[\beta_i^+ + \beta_i^-] < t,\, \beta_i^+,\, \beta_i^- \leq 0 \end{array}$$

The objective function becomes

$$\frac{1}{2}\begin{bmatrix} \boldsymbol{\beta}^+ \\ \boldsymbol{\beta}^- \end{bmatrix}^T \begin{bmatrix} H & -H \\ -H & H \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^+ \\ \boldsymbol{\beta}^- \end{bmatrix} + \begin{bmatrix} f^T & -f^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^+ \\ \boldsymbol{\beta}^- \end{bmatrix}$$

The constraint becomes

$$\begin{bmatrix} 1_{2p}^T \\ -I_{2p} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^+ \\ \boldsymbol{\beta}^- \end{bmatrix} \leq \begin{bmatrix} t \\ 0_{2p}^T \end{bmatrix}$$

where 1_{2p} is $2p$ by 1 vector with all 1, I_{2p} is $2p \times 2p$ identical matrix, 1_{2p} is $2p$ by 1 vector with all 0.

Let

$$Q = \begin{bmatrix} H & -H \\ -H & H \end{bmatrix} \quad c^T = \begin{bmatrix} f^T & -f^T \end{bmatrix} \quad A = \begin{bmatrix} 1_{2p}^T \\ -I_{2p} \end{bmatrix}$$

$$b = \begin{bmatrix} t \\ 0_{2p}^T \end{bmatrix} \quad x = \begin{bmatrix} \boldsymbol{\beta}^+ \\ \boldsymbol{\beta}^- \end{bmatrix}$$

The lasso problem can be written as standard quadratic programming problem:

$$\begin{array}{ll} minimize_x & \frac{1}{2}x^TQx + c^Tx \\ subject\,to & Ax \leq b \end{array}$$

3 The LARS Algorithm

Efron et al [3] proposed the least angle regression algorithm. This algorithm is computationally efficient. It only need p iterations for p predictors.

3.1 Notation

Suppose y is the response variable of interest. $x_1, x_2, \cdots x_p$ are observed predictor variables assumed to be linearly independent and nomalized as in section 1. $\mathcal{A}$ denotes a subset of index $(1, 2, \cdots, p)$, define the matrix

$$X_{\mathcal{A}} = (\cdots, s_jx_j, \cdots)_{j\in\mathcal{A}}$$

$$G_{\mathcal{A}} = X'_{\mathcal{A}}X_{\mathcal{A}}\, and\, A_{\mathcal{A}} = (1'_{\mathcal{A}}G^{-1}1_{\mathcal{A}})^{-1/2}.$$

The equiangular vector $u_{\mathcal{A}} = X_{\mathcal{A}}w_{\mathcal{A}}$, where $w_{\mathcal{A}} = A_{\mathcal{A}}G^{-1}1_{\mathcal{A}}$.

3.2 The LARS algorithm

1. Start with $\hat{\mu}_0 = 0$
2. In each step, update μ_i in the following way:
(1). Compute $\hat{c} = X'(y - \hat{\mu}_i)$ (2). Let $\hat{C} = max_j|\hat{c}_j|$ and add j into active set $\mathcal{A}$. (3). Set $s_j = sign(\hat{c}_j)$ for all j in $\mathcal{A}$. (4). Compute $X_{\mathcal{A}},\ A_{\mathcal{A}},\ u_{\mathcal{A}}$ defined in the last section, and

$$a = X'u_{\mathcal{A}}$$

coeffcients	quadratic programming	LARS
β_1	-1.38714880002693e-19	0
β_2	-3.00961394071294e-20	0
β_3	80.0607375117748	80.0607375117755
β_4	3.49816666282146e-18	0
β_5	-4.90071425713167e-19	0
β_6	-3.16049168444261e-20	0
β_7	-1.08009935664724e-18	0
β_8	7.05816673079248e-19	0
β_9	19.9392624882252	19.9392624882245
β_{10}	2.00773278278871e-19	0

Table 1: Fitting the diabetes data set by LARS and quadratic programming at $t = 100$

(5). Update

$$\hat{\mu}_{i+1} = \hat{\mu}_i + \hat{\gamma} u_{\mathcal{A}}$$

where

$$\hat{\gamma} = \overset{+}{\min_{j \in \mathcal{A}}} \left\{ \frac{\hat{C} - \hat{c}_j}{A_{\mathcal{A}} - a_j}, \frac{\hat{C} + \hat{c}_j}{A_{\mathcal{A}} + a_j} \right\}$$

3. Continue updating until p step, or the max correlation is 0, or reaches bound.

3.3 Lasso Modification

Efron et al [3] proves that under certain modification, the LARS solution is the lasso solution.

Let $\hat{d}$ be the $p \times 1$ vector with jth entry equaling $s_j w_{\mathcal{A}j}$ and all other entries to be 0. Let

$$\gamma_j = -\hat{\beta}_j / \hat{d}_j.$$

and

$$\tilde{\gamma} = \min_{\gamma_j > 0} (\gamma_j)$$

The lasso modification is if $\tilde{\gamma} < \hat{\gamma}$, stop the ongoing step at $\gamma = \tilde{\gamma}$ and remove $\tilde{j}$ from the calculation of next equiangular direction. i.e,

$$\hat{\mu}_{\mathcal{A}}^{+} = \hat{\mu}_{\mathcal{A}} + \tilde{\gamma} u_{\mathcal{A}}$$

and

$$\mathcal{A}^{+} = \mathcal{A} - \tilde{j}$$

4 Code and Result

The *QPlasso.m* implements quadratic programming method with algorithm 'interior-point-convex'. The *lars.m* (I referenced matlab package 'lar' and R packages 'lars', some piece of code are similar to the code in the package) is the matlab code implements LARS algorithm with lasso modification. The codes is available on `https://github.com/lunge111/MAT258A`.

The fitting result of these two method agrees with each other generally. One major difference is for those 0 coefficients, LARS will set them exactly equal to 0 but quadratic programming will yield some very small value. Table 1 shows different value fitted by these two methods on diabetes data set (the data set used in [3]).

Both two methods are very fast. The LARS finishes 400 fitting in 1 second, while quadratic programming does it in 2 seconds, averagely 9 iterations are used.

5 Reproduction of Paper Result

I reproduced Figure 1 in [3]. The result agrees with the paper.

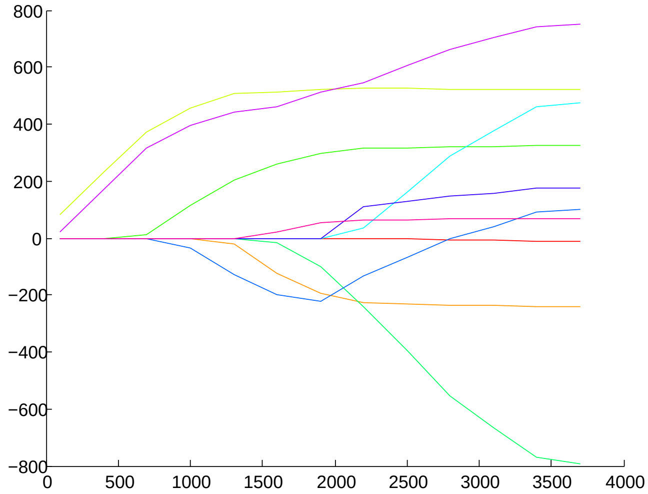

Figure 1: Estimates the regression coefficients $\beta_j, j = 1, 2, \cdots, 10$, for diabetes study, as a function of $t = \sum_j |\beta_j|$.

References

[1] Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso.* J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288).

[2] Busa J. (2012). *Solving quadratic programming problem with linear constraints containing absolute values[J].* Acta Electrotechnica et Informatica, 12(3): 11-18.

[3] Efron B, Hastie T, Johnstone I, et al. (2004) *Least angle regression[J].* The Annals of statistics, 32(2): 407-499.