Fundamentals of Linear Algebra and Optimization
Lasso Regression

Jean Gallier and Jocelyn Quaintance

CIS Department
University of Pennsylvania

jean@cis.upenn.edu

April 18, 2022

Scaling Ridge Regression

The main weakness of ridge regression is that the estimated weight vector w usually has many nonzero coefficients.

Scaling Ridge Regression

The main weakness of ridge regression is that the estimated weight vector w usually has many nonzero coefficients.

As a consequence, ridge regression does not scale up well.

Scaling Ridge Regression

The main weakness of ridge regression is that the estimated weight vector w usually has many nonzero coefficients.

As a consequence, ridge regression does not scale up well.

In practice we need methods capable of handling millions of parameters, or more.

Lasso Regression

A way to encourage sparsity of the vector w, which means that many coordinates of w are zero, is to replace the quadratic penalty function $\tau w^\top w = \tau \|w\|_2^2$ by the penalty function $\tau \|w\|_1$, with the ℓ^2-norm replaced by the ℓ^1-norm.

Lasso Regression

A way to encourage sparsity of the vector w, which means that many coordinates of w are zero, is to replace the quadratic penalty function $\tau w^\top w = \tau \left\| w \right\|_2^2$ by the penalty function $\tau \left\| w \right\|_1$, with the ℓ^2-norm replaced by the ℓ^1-norm.

This method was first proposed by Tibshirani arround 1996, under the name *lasso*, which stands for "least absolute selection and shrinkage operator."

Lasso Regression

A way to encourage sparsity of the vector w, which means that many coordinates of w are zero, is to replace the quadratic penalty function $\tau w^\top w = \tau \left\| w \right\|_2^2$ by the penalty function $\tau \left\| w \right\|_1$, with the ℓ^2-norm replaced by the ℓ^1-norm.

This method was first proposed by Tibshirani arround 1996, under the name *lasso*, which stands for "least absolute selection and shrinkage operator."

This method is also known as *ℓ^1-regularized regression*, but this is not as cute as "lasso," which is used predominantly.

Lasso Regression: Notational Convention

Given a set of training data $\{(x_1, y_1), \ldots, (x_m, y_m)\}$, with $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$, if X is the $m \times n$ matrix

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_m^\top \end{pmatrix},$$

in which the row vectors $x_i^\top$ are the rows of X, then *lasso regression* is the following optimization problem

Lasso Regression: Problem (lasso1)

Program (lasso1):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\xi^\top \xi + \tau \left\| w \right\|_1 \\ &\text{subject to} \\ &\qquad\qquad\quad y - Xw = \xi, \end{aligned}$$

minimizing over ξ and w, where $\tau > 0$ is some constant determining the influence of the regularizing term $\left\| w \right\|_1$.

Lasso Regression: (lasso1) Reduction

The difficulty with the regularizing term $\|w\|_1 = |w_1| + \cdots + |w_n|$ is that the map $w \mapsto \|w\|_1$ is **not** differentiable for all w.

Lasso Regression: (lasso1) Reduction

The difficulty with the regularizing term $\|w\|_1 = |w_1| + \cdots + |w_n|$ is that the map $w \mapsto \|w\|_1$ is **not** differentiable for all w.

This difficulty can be overcome by using subgradients, but the dual of the above program can also be obtained in an elementary fashion by using a trick, which is that if $x \in \mathbb{R}$, then

$$|x| = \max\{x, -x\}.$$

Lasso Regression: (lasso1) Reduction

The difficulty with the regularizing term $\|w\|_1 = |w_1| + \cdots + |w_n|$ is that the map $w \mapsto \|w\|_1$ is **not** differentiable for all w.

This difficulty can be overcome by using subgradients, but the dual of the above program can also be obtained in an elementary fashion by using a trick, which is that if $x \in \mathbb{R}$, then

$$|x| = \max\{x, -x\}.$$

Using this trick, by introducing a vector $\epsilon \in \mathbb{R}^n$ of *nonnegative* variables, we can rewrite lasso minimization as follows:

Lasso Regression: Program (lasso2)

Program lasso regularization (lasso2):

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\xi^\top \xi + \tau \mathbf{1}_n^\top \epsilon \\ \text{subject to} \quad & \\ & y - Xw = \xi \\ & w \leq \epsilon \\ & -w \leq \epsilon. \end{aligned}$$

minimizing over ξ, w and ϵ, with $y, \xi \in \mathbb{R}^m$, and $w, \epsilon, \mathbf{1}_n \in \mathbb{R}^n$.

Lasso Regression: Program (lasso2)

Program lasso regularization (lasso2):

$$
\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\xi^\top \xi + \tau \mathbf{1}_n^\top \epsilon \\
&\text{subject to} \\
&\qquad\qquad y - Xw = \xi \\
&\qquad\qquad\qquad w \leq \epsilon \\
&\qquad\qquad\quad -w \leq \epsilon.
\end{aligned}
$$

minimizing over ξ, w and ϵ, with $y, \xi \in \mathbb{R}^m$, and $w, \epsilon, \mathbf{1}_n \in \mathbb{R}^n$.
The constraints $w \leq \epsilon$ and $-w \leq \epsilon$ are equivalent to $|w_i| \leq \epsilon_i$ for $i = 1, \ldots, n$, so for an optimal solution we must have $\epsilon \geq 0$ and $|w_i| = \epsilon_i$, that is, $\|w\|_1 = \epsilon_1 + \cdots + \epsilon_n$.

Lasso Regression: Program (lasso1) Solution

The best way to solve lasso minimization is to use ADMM.

Lasso Regression: Program (lasso1) Solution

The best way to solve lasso minimization is to use ADMM.

Lasso minimization can be stated as the following optimization problem:

$$\text{minimize} \quad (1/2)\left\|Ax - b\right\|_2^2 + \tau \left\|x\right\|_1 ,$$

with $A = X$, $b = y$ and $x = w$, to conform with our original formulation.

Lasso Regression: Program (lasso1) Solution

The best way to solve lasso minimization is to use ADMM.

Lasso minimization can be stated as the following optimization problem:

$$\text{minimize} \quad (1/2)\left\|Ax - b\right\|_2^2 + \tau\left\|x\right\|_1,$$

with $A = X$, $b = y$ and $x = w$, to conform with our original formulation.

The *lasso minimization* is converted to the following problem in ADMM form:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\left\|Ax - b\right\|_2^2 + \tau\left\|z\right\|_1 \\ \text{subject to} & x - z = 0. \end{array}$$

Lasso Regression: ADMM Solution

Then the ADMM procedure is

$$
\begin{aligned}
x^{k+1} &= (A^\top A + \rho I)^{-1}(A^\top b + \rho(z^k - u^k)) \\
z^{k+1} &= S_{\tau/\rho}(x^{k+1} + u^k) \\
u^{k+1} &= u^k + x^{k+1} - z^{k+1}
\end{aligned}
$$

where $\rho > 0$ is some given constant.

Lasso Regression: ADMM Solution

Then the ADMM procedure is

$$
\begin{aligned}
x^{k+1} &= (A^\top A + \rho I)^{-1}(A^\top b + \rho(z^k - u^k)) \\
z^{k+1} &= S_{\tau/\rho}(x^{k+1} + u^k) \\
u^{k+1} &= u^k + x^{k+1} - z^{k+1}
\end{aligned}
$$

where $\rho > 0$ is some given constant.

Since $\rho > 0$, the matrix $A^\top A + \rho I$ is symmetric positive definite. Note that the x-update looks like a *ridge regression step*.

Soft Thresholding Operator

In the above procedure, the function S_c known as a *soft thresholding operator*. If $v \in \mathbb{R}$ it is given by

$$S_c(v) = \begin{cases} v - c & \text{if } v > c \\ 0 & \text{if } |v| \leq c \\ v + c & \text{if } v < -c. \end{cases}$$

Soft Thresholding Operator

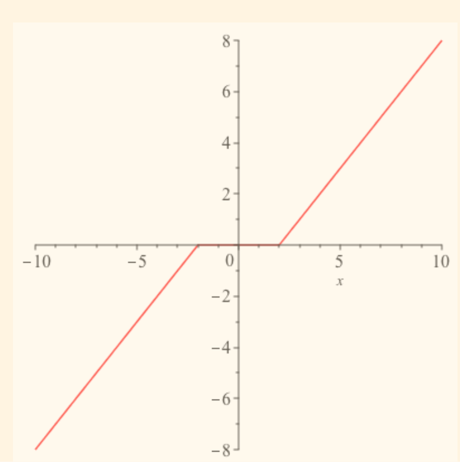

Figure 1: The graph of S_c (when $c = 2$).

Soft Thresholding Operator

The operator S_c is extended to vectors in $\mathbb{R}^n$ component wise, that is, if $x = (x_1, \ldots, x_n)$, then

$$S_c(x) = (S_c(x_1), \ldots, S_c(x_n)).$$

Soft Thresholding Operator

The operator S_c is extended to vectors in $\mathbb{R}^n$ component wise, that is, if $x = (x_1, \ldots, x_n)$, then

$$S_c(x) = (S_c(x_1), \ldots, S_c(x_n)).$$

The soft thresholding operator is one of the built-in functions in `Matlab`.