Fitting Generalized Lasso Models and Post-Selection Inference for the Lasso

Taeyoung Chang

2021. 12. 2

CONTENTS

1. Fitting Generalized Lasso Models
2. Post-selection Inference for the Lasso

Notation

- x_i : p-dimensional vector for i-th observation of predictor variables
- $\mathbf{x}_j$: n-dimensional vector for j-th predictor of n observations
- $s_j \in \text{sign}(\beta_j)$: subgradient of $|\beta_j|$

$$s_j = \begin{cases} 1 & \beta_j > 0 \\ -1 & \beta_j < 0 \\ [-1, 1] & \beta_j = 0 \end{cases}$$

  - $\mathbf{s} = (s_1, \cdots, s_p)$

Motivation

- So far we have focused on the Lasso for squared-error loss, and exloited the piecewise-linearity of its coefficient profile to efficiently compute the entire path.
- Unfortunately this is not the case for most other loss functions.
  - Obtaining the coefficient path is potentially more costly.

Logistic regression example

- We will use logistic regression as an example.
- Use loss function L which is the negative log-likelihood.
- The problem is given as

$$\text{minimize}_{\beta \in \mathbb{R}^p,\, \beta_0 \in \mathbb{R}} \; -\left\{ \frac{1}{n} \sum_{i=1}^{n} y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i) \right\} + \lambda \|\beta\|_1$$

$$\text{where } y_i \overset{indep}{\sim} \text{Bern}(\mu_i) \;\text{ and }\; \text{logit}(\mu_i) = \beta_0 + x_i^T \beta \quad \forall\, i = 1, \cdots, n$$

The solution satisfies the subgradient condition

- As in the case of the lasso for squared-error loss , the solution should satisfy the subgradient condition.

$$\frac{\partial}{\partial \beta} f(\beta, \beta_0) = \mathbf{0} \quad \text{and} \quad \frac{\partial}{\partial \beta_0} f(\beta, \beta_0) = 0$$

where $f(\beta, \beta_0)$ is the given objective function.

- We shall taking advantage of

$$\frac{\partial}{\partial \beta} \mu_i = \mu_i(1 - \mu_i)x_i \quad \text{and} \quad \frac{\partial}{\partial \beta_0} \mu_i = \mu_i(1 - \mu_i)$$

Derivation of the subgradient condition

- First condition

$$
\begin{aligned}
&\frac{\partial}{\partial \beta} f(\beta, \beta_0) = \mathbf{0} \\
&\Leftrightarrow \frac{\partial}{\partial \beta} - \frac{1}{n} \sum_{i=1}^{n} y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i) + \lambda \|\beta\|_1 = \mathbf{0} \\
&\Leftrightarrow -\frac{1}{n} \sum_{i=1}^{n} y_i (1 - \mu_i) x_i - (1 - y_i) \mu_i x_i + \lambda \mathbf{s} = \mathbf{0} \\
&\Leftrightarrow -\frac{1}{n} \sum_{i=1}^{n} (y_i - \mu_i) x_i + \lambda \mathbf{s} = \mathbf{0} \\
&\Leftrightarrow -\frac{1}{n} \langle \mathbf{x}_j, \mathbf{y} - \mu \rangle + \lambda s_j = 0 \quad \forall\, j = 1, \cdots p
\end{aligned}
$$

Derivation of the subgradient condition

- Second condition

$$
\begin{aligned}
&\frac{\partial}{\partial \beta_0} f(\beta, \beta_0) = 0 \\
&\Leftrightarrow \frac{\partial}{\partial \beta_0} - \frac{1}{n} \sum_{i=1}^{n} y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i) + \lambda \|\beta\|_1 = 0 \\
&\Leftrightarrow -\frac{1}{n} \sum_{i=1}^{n} y_i (1 - \mu_i) - (1 - y_i) \mu_i = 0 \\
&\Leftrightarrow -\frac{1}{n} \sum_{i=1}^{n} (y_i - \mu_i) = 0 \\
&\Leftrightarrow \frac{1}{n} \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \mu_i
\end{aligned}
$$

Solution path on λ grid

- The nonlinearity of μ_i in β_j results in piecewise nonlinear coefficient profiles.
- Hence, we settle for a solution path on a sufficiently fine grid of values for λ
- The largest value of λ we need to consider is

$$\lambda_{max} = \max_{j=1,\cdots p} |\langle \mathbf{x}_j, \mathbf{y} - \bar{y}\mathbf{1} \rangle|$$

  - This is because it is the smallest value of λ for which $\hat{\beta} = 0$ and $\hat{\beta}_0 = \text{logit}(\bar{y})$

Solution path on λ grid

- A reasonable sequence is 100 values $\lambda_1 > \lambda_2 > \cdots > \lambda_{100}$ equally spaced on the log-scale from λ_{max} down to $\varepsilon\lambda_{max}$ where ε is some small fraction such as 0.001
- An approach that has proven to be surprisingly efficient is path-wise coordinate descent.

Coordinate descent

- For the problem

$$\text{minimize } f(\mathbf{x})$$

with convex and differentiable function $f : \mathbb{R}^m \to \mathbb{R}$, coordinatewise minimization can yield a global minimization.

$$f(\mathbf{x}^* + \delta e_i) \geq f(\mathbf{x}^*) \quad \forall\, \delta > 0\ ,\ i = 1, \cdots, m \ \Rightarrow f(\mathbf{x}^*) = \min f(\mathbf{x})$$

where e_i is the i-th standard basis vector of $\mathbb{R}^m$

- We can also use coordinate descent for the problem

$$\text{minimize } f(\mathbf{x})$$

where $f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x}) = g(\mathbf{x}) + \sum_{i=1}^{n} h_i(x_i)$ with g being convex and differentiable and h_i being convex. Here, the nonsmooth part h is called separable

Coordinate descent

- Coordinate descent method is proceeded as the following :
  1. Take initial value $\mathbf{x}^{(0)} \in \mathbb{R}^m$
  2. Iterate

     $$x_i^{(k)} = \text{argmin}_{x_i} f(x_1^{(k)}, \cdots, x_{i-1}^{(k)}, x_i, x_{i+1}^{(k-1)}, \cdots, x_m^{(k-1)}) \quad \forall\ i = 1, \cdots, m$$

     for step $k = 1, 2, \cdots$ and so on until convergence.

Coordinate descent

- Coordinate descent example : linear regression
- minimize $\frac{1}{2}\|y - X\beta\|_2^2$ over β_i with all $\beta_j \quad \forall\, j \neq i$ are fixed.
- Using $\frac{\partial \beta}{\partial \beta_i} = e_i$ where e_i is i-th standard basis of $\mathbb{R}^p$

$$
\begin{aligned}
&\hat{\beta}_i \text{ minimizes } \frac{1}{2}\|y - X\beta\|_2^2 \text{ over } \beta_i \text{ with all } \beta_j \quad \forall\, j \neq i \text{ are fixed} \\
&\Leftrightarrow \frac{\partial}{\partial \beta_i}\frac{1}{2}\|y - X\beta\|_2^2 = 0 \quad \text{at } \beta_i = \hat{\beta}_i \\
&\Leftrightarrow \frac{\partial \beta}{\partial \beta_i}\frac{\partial}{\partial \beta}\frac{1}{2}\|y - X\beta\|_2^2 = 0 \quad \text{at } \beta_i = \hat{\beta}_i \\
&\Leftrightarrow e_i^T(X^T X\beta - X^T y) = 0 \quad \text{at } \beta_i = \hat{\beta}_i \\
&\Leftrightarrow \mathbf{x}_i^T(X\beta - y) = \mathbf{x}_i^T(X_i\beta_i + X_{-i}\beta_{-i} - y) = 0 \quad \text{at } \beta_i = \hat{\beta}_i \\
&\Leftrightarrow \hat{\beta}_i = \frac{\mathbf{x}_i^T(y - X_{-i}\beta_{-i})}{\mathbf{x}_i^T\mathbf{x}_i}
\end{aligned}
$$

Coordinate descent

- Coordinate descent example : the Lasso problem for squared-error loss
- minimize $\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$ over β_i with all $\beta_j \quad \forall\, j \neq i$ are fixed.
- By similar logic we used for the linear regression case , solution $\hat{\beta}_i$ should satisfy

$$\hat{\beta}_i + \frac{\lambda}{\|\mathbf{x}_i\|_2^2} s_i = \frac{\mathbf{x}_i^T (y - X_{-i}\beta_{-i})}{\mathbf{x}_i^T \mathbf{x}_i}$$

- We have the solution $\hat{\beta}_i$ given as

$$\hat{\beta}_i = S_{\lambda/\|\mathbf{x}_i\|_2^2}\Big(\frac{\mathbf{x}_i^T (y - X_{-i}\beta_{-i})}{\mathbf{x}_i^T \mathbf{x}_i}\Big) = \frac{1}{\mathbf{x}_i^T \mathbf{x}_i} S_\lambda\big(\mathbf{x}_i^T (y - X_{-i}\beta_{-i})\big)$$

where $S_\lambda(x)$ is soft-thresholding defined as

$$S_\lambda(x) = \begin{cases} x - \lambda & \text{if } x > \lambda \\ 0 & \text{if } -\lambda \leq x \leq \lambda \\ x + \lambda & \text{if } x < \lambda \end{cases}$$

Pathwise coordinate descent

- Outer loop
  - Find optimal value β for each λ_k in the order of $\lambda_1 > \lambda_2 > \cdots > \lambda_{100}$
  - By starting at λ_1 , where all parameters are zero, we use warm starts in computing the solutions at the decreasing sequence of λ values.
    - resulting β for λ_k is used as an initial value of coordinate descent algorithm for λ_{k+1}

Pathwise coordinate descent

- Inner loop
  - For each value λ_k , solve the lasso problem for one β_j only, holding the others fixed. This is done by coordinate descent. One or several coordinate cycles are implemented until the estimates stabilize.
  - Store the nonzero coefficients in the active set $\mathcal{A}$. (The active set grows slowly as λ decreases.)
  - Iterates coordinate descent using only those variables until convergence.
  - One more sweep through all the variables to check optimality conditions. If there is a variable not satisfying the condition, then add it in active set $\mathcal{A}$ and go back to the first step of inner loop.

Comments

- The R package glmnet employs a 'proximal-Newton' strategy at each value λ_k, which takes advantage of a weighted least-squares and coordinate descent.
- We can consider another penalty term called as 'elastic net' penalty which bridges the gap between the lasso and ridge regression. It is defined as
$$P_\alpha(\beta) = \frac{1}{2}\{(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1\}$$
for some $\alpha \in [0, 1]$
    - When the predictors are excessively correlated, the lasso performs somewhat poorly.
    - Elastic net can be used as an alternative in that case

CONTENTS

1 Fitting Generalized Lasso Models

2 Post-selection Inference for the Lasso

Post-selection Inference

- Inference is generally difficult for adaptively selected models.
- Suppose we have fit a lasso regression model with a particular value for λ , which ends up selecting a subset $\mathcal{A}$ of size $|\mathcal{A}| = k$ of p available variables.
- Question : interest in the population regression parameters using the full set of p predictors VS interest is restricted to the population regression parameters using only the subset $\mathcal{A}$

Post-selection Inference

- Focus on the second case
- The idea is to condition on the selected set $\mathcal{A}$ itself, and then perform conditional inference on the unrestricted (not lasso-shrunk) regression coefficients of the response on only the variables in $\mathcal{A}$
- For the case of the lasso with squared-error loss, using the fact about convexity along with delicate Gaussian conditioning arguments, it leads to truncated Gaussian and t-distributions for parameters of interest.

Reference

Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Institute of Mathematical Statistics Monographs. Cambridge University Press.

- Lecture note for Coordinate Descent by Ryan Tibshirani
- Convex optimization for All