P9120 HW1 answer

Guojing Wu | UNI: gw2383

9/30/2019

1. Let X denote an n $\times$ p matrix with each row an input vector and y denote an n-dimensional vector of the output in the training set. For fixed $q \geq 1$, define

$$Bridge_\lambda(\beta) = (y - X\beta)^T(y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|^q$$

for $\lambda > 0$. Denote the minimal value of the penalty function over the least squares solution set by

$$t_0 = min_{\beta: X^T X\beta = X^T y} \sum_{j=1}^{p} |\beta_j|^q$$

(a) Show that $Bridge_\lambda(\beta)$ for $\lambda > 0$ is a convex function in β, which is strictly convex for q $>$ 1.

Denote $RSS = (y - X\beta)^T(y - X\beta)$, we then calculate its second derivative:

$$\frac{\partial^2 Bridge_\lambda(\beta)}{\partial \beta^2} = \frac{\partial^2 RSS}{\partial \beta^2} + \frac{\partial^2 \lambda \sum_{j=1}^{p} |\beta_j|^q}{\partial \beta^2}$$

$$\begin{aligned}
\frac{\partial RSS}{\partial \beta} &= -2X^T(y - X\beta) \\
\frac{\partial^2 RSS}{\partial \beta^2} &= 2X^T X \\
\frac{\partial \lambda \sum_{j=1}^{p} |\beta_j|^q}{\partial \beta_j} &= \lambda q |\beta_j|^{q-1} sign(\beta_j) \\
\frac{\partial^2 \lambda \sum_{j=1}^{p} |\beta_j|^q}{\partial \beta_j^2} &= sign(\beta_j)^2 \lambda q(q-1)|\beta_j|^{q-2} \\
&= \lambda q(q-1)|\beta_j|^{q-2}
\end{aligned}$$

- $X^T X$ is postive semi-definite matrix
- $\lambda q(q-1)|\beta_j|^{q-2} \geq 0$ for $q \geq 1, \lambda > 0$
- $\lambda q(q-1)|\beta_j|^{q-2} > 0$ for $q > 1$

So $Bridge_\lambda(\beta)$ for $\lambda > 0$ is a convex function in β, which is strictly convex for q $>$ 1.

(b) Show that for q > 1 there is a unique minimizer, $\hat{\beta}(\lambda)$, with $\sum_{j=1}^{p} |\hat{\beta}_j(\lambda)|^q \leq t_0$

$Bridge_{\lambda}(\beta)$ is a convex function, if there exist unique minimizer $\hat{\beta}$, then $\frac{\partial RSS}{\partial \hat{\beta}} = 0$

Denote $S_j(\beta, X, y) = \frac{\partial RSS}{\partial \beta_j}$ and $d(\gamma, \lambda, \beta) = \lambda q |\beta_j|^{q-1} sign(\beta_j)$, solving the above equation equals to solve $S_j(\beta, X, y) = -d(q, \lambda, \beta_j)$ for $j = 1, 2, ..., p$.

Rewrite β as (β_j, β^{-j}), where β^{-j} is a (p-1) vector consisting of the $\beta_i's$ other than β_j. Then we got:

$$
\begin{aligned}
S_j(\beta_j, \beta^{-j}, X, y) &= -d(q, \lambda, \beta_j) \\
S_j(\beta_j, \beta^{-j}, X, y) &= \frac{\partial (y - X\beta)^T (y - X\beta)}{\partial \beta_j} \\
&= \frac{\partial (y - \sum_{i=1}^{p} x_i \beta_i)^T (y - \sum_{i=1}^{p} x_i \beta_i)}{\partial \beta_{j \in (1,2,\ldots,p)}} \\
&= 2x_j^T x_j \beta_j + 2 \sum_{i \neq j} x_j^T x_i \beta_i - 2x_j^T y \\
-d(q, \lambda, \beta_j) &= \lambda q |\beta_j|^{q-1} sign(\beta_j)
\end{aligned}
$$

We can tell from above that:

- $S_j(\beta_j, \beta^{-j}, X, y)$ is linear function of β_j, with positive slope $2x_j^T x_j$
- from answer for (a) we know that, $-d(q, \lambda, \beta_j)$ is a nonlinear function of β_j and continuous, differentiable and monotonically decreasering for q > 1, except at $\beta_j = 0$ for $1 < q < 2$

So the equation above will have a unique solution.

(c)

(d)

2. We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain p + 1 models, containing 0, 1, 2, . . . , p predictors. Explain your answers:

(a) Which of the three models with k predictors has the smallest training RSS?

Best subset selection has the smallest training RSS. Because forward and backward selection results depend heavily on the path they choose.

(b) Which of the three models with k predictors has the smallest test RSS?

It varies by chance, the smallest test RSS could happened to any of them, most likely to be best subset selection though.

(c) True or False:

(i) The predictors in the k variable model identified by forward stepwise are a subset of the predictors in the (k+1) variable model identified by forward stepwise selection.

True.

(ii) The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the (k+1) variable model identified by backward stepwise selection.

True

(iii) The predictors in the k variable model identified by backward stepwise are a subset of the predictors in the (k+1) variable model identified by forward stepwise selection.

False

(iv) The predictors in the k variable model identified by forward stepwise are a subset of the predictors in the (k+1) variable model identified by backward stepwise selection.

False

(v) The predictors in the k variable model identified by best subset are a subset of the predictors in the (k+1) variable model identified by best subset selection.

False

3. Derive the entries in Table 3.4, the explicit forms for estimators in the orthogonal case.

The OLS estimator of β is $\hat{\beta}^{ols} = (X^TX)^{-1}X^Ty$, and since X columns are orthonormal, $X^TX = I$, so $\hat{\beta}^{ols} = X^Ty$.

a) for best subset

Extend $X_{n\times p}$ to $\mathbb{R}^{n\times N}$, we add N - p linearly independent additional orthonormal vectors $\tilde{x}_j$ with corresponding γ_j coefficients to the end, then y can write as:

$$y = \sum_{j=1}^{p} \hat{\beta}_j^{ols} x_j + \sum_{j=p+1}^{N} \gamma_j \tilde{x}_j$$

If we seek to approximate y with a subset of size M, then $\hat{y} = \sum_{j=1}^{p} I_j \hat{\beta}_j^{ols} x_j$, with $I_j = 1$ indicates x_j in the subset, and zero otherwise. Then

$$
\begin{aligned}
RSS &= ||y-\hat{y}||_2^2 \\
&= ||\sum_{j=1}^{p}\hat{\beta}_j^{\,ols}x_j + \sum_{j=p+1}^{N}\gamma_j\tilde{x}_j - \sum_{j=1}^{p}I_j\hat{\beta}_j^{\,ols}x_j||_2^2 \\
&= ||\sum_{j=1}^{p}(1-I_j)\hat{\beta}_j^{\,ols}x_j + \sum_{j=p+1}^{N}\gamma_j\tilde{x}_j||_2^2 \\
&= \sum_{j=1}^{p}(1-I_j)^2\hat{\beta}_j^{2^{\,ols}}||x_j||_2^2 + \sum_{j=p+1}^{N}\gamma_j^2||\tilde{x}_j||_2^2 \\
&= \sum_{j=1}^{p}(1-I_j)^2\hat{\beta}_j^{2^{\,ols}} + \sum_{j=p+1}^{N}\gamma_j^2
\end{aligned}
$$

Now we rank all the $\hat{\beta}_j^{\,ols}$ by their absolute value $(|\hat{\beta_{[1]}}^{ols}| \geq |\hat{\beta_{[2]}}^{ols}| \geq ... \geq |\hat{\beta}_{[M]}^{ols}| \geq |\hat{\beta}_{[M+1]}^{ols}| \geq ... \geq |\hat{\beta}_{[p]}^{ols}|)$, then

$$
\begin{aligned}
RSS &= \sum_{j=[1]}^{[M]}(1-I_j)^2\hat{\beta}_j^{2^{\,ols}} + \sum_{j=[M+1]}^{[p]}(1-I_j)^2\hat{\beta}_j^{2^{\,ols}} + \sum_{j=p+1}^{N}\gamma_j^2 \\
min\{RSS\} &= \sum_{j=[M+1]}^{[p]}(1-I_j)^2\hat{\beta}_j^{2^{\,ols}} + \sum_{j=p+1}^{N}\gamma_j^2 \\
with &: \\
\hat{\beta}_j^{\,best-subset} &= argmin_{\hat{\beta}^{ols},I}RSS \\
&= \hat{\beta}_j^{\,ols}\cdot I(|\hat{\beta}_j^{\,ols}| \geq |\hat{\beta}_{[M]}^{ols}|)
\end{aligned}
$$

b) for ridge regression

From previous conclusion, we can derive

$$
\begin{aligned}
\hat{\beta}^{ridge} &= (X^TX+\lambda I)^{-1}X^Ty \\
&= (I+\lambda I)^{-1}X^Ty \\
&= \frac{X^Ty}{1+\lambda} \\
&= \hat{\beta}^{ols}/(1+\lambda)
\end{aligned}
$$

c) for LASSO

$L(\beta) = ||y-X\beta||_2^2 + \lambda||\beta||_1$, here we set the objective function as $F(\beta) = \frac{1}{2}||y-X\beta||_2^2 + \lambda||\beta||_1$, then

$$
\begin{aligned}
F(\beta) &= \frac{1}{2}y^T y - y^T X\beta + \frac{1}{2}\beta^T\beta + \lambda||\beta||_1 \\
&= const. - y^T X\beta + \frac{1}{2}||\beta||^2 + \lambda||\beta||_1 \\
&= const. - (\hat{\beta}^{ols})^T\beta + \frac{1}{2}||\beta||^2 + \lambda||\beta||_1
\end{aligned}
$$

After remove the constant and consider each β_j individually, we get $f(\beta_j) = -\hat{\beta_j}^{ols}\beta_j + \frac{1}{2}\beta_j^2 + \lambda|\beta_j|$ and we need to minimize it. Then we need $f'(\beta_j) = -\hat{\beta_j}^{ols} + \beta_j + \lambda \cdot sign(\beta_j) = 0$

Here we consider 6 cases:

- $\beta_j < 0$ and $\hat{\beta_j}^{ols} > 0$, then $f'(\beta_j)$ will always < 0, there is no minimum
- $\beta_j \geq 0$ and $\lambda > \hat{\beta_j}^{ols} > 0$, then $f'(\beta_j)$ will always > 0, there is no minimum
- $\beta_j \geq 0$ and $\hat{\beta_j}^{ols} \geq \lambda$, then we get $argmin_{\beta_j} f'(\beta_j) = \hat{\beta_j}^{ols} - \lambda$
- $\beta_j > 0$ and $\hat{\beta_j}^{ols} < 0$, then $f'(\beta_j)$ will always > 0, there is no minimum
- $\beta_j \leq 0$ and $-\lambda < \hat{\beta_j}^{ols} < 0$, then $f'(\beta_j)$ will always < 0, there is no minimum
- $\beta_j \leq 0$ and $\hat{\beta_j}^{ols} \leq \lambda$, then we get $argmin_{\beta_j} f'(\beta_j) = \hat{\beta_j}^{ols} + \lambda$

Whenever there is no minimum, we discard the covariate, which means shrink the $\beta_j = 0$. So combine all those situations, $\hat{\beta}_j^{lasso} = sign(\hat{\beta_j}^{ols})(|\hat{\beta_j}^{ols}| - \lambda)_+$

Appendix

```
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE, comment = "")
library(tidyverse)
options(knitr.table.format = "latex")
theme_set(theme_bw())
```