P9120- Homework # 1 (online submission)

Assigned: September 19, 2024 Due: 9pm EST on October 5, 2024

Maximum points that you can score in this Homework is 20.

Please include all R/Python code you used to complete this homework.

1. (8 points) Let $\mathbf{X}$ denote an $n \times p$ matrix with each row an input vector and $\mathbf{y}$ denote an n-dimensional vector of the output in the training set. For fixed $q \geq 1$, define

$$\text{Bridge}_\lambda(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|^q$$

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

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

(a) Using the definition of convexity, show that $\text{Bridge}_\lambda(\boldsymbol{\beta})$ for $\lambda > 0$ is a convex function in $\boldsymbol{\beta}$, which is strictly convex for $q > 1$.

Definition: A function $f(x) : \mathcal{X} \to \mathbb{R}$ is *convex* if and only if for all $0 < t < 1$ and $x_1, x_2 \in \mathcal{X}$,

$$f[tx_1 + (1-t)x_2] \leq tf(x_1) + (1-t)f(x_2).$$

$f(x)$ is *strictly convex* if and only if for all $0 < t < 1$ and $x_1, x_2 \in \mathcal{X}$ such that $x_1 \neq x_2$,

$$f[tx_1 + (1-t)x_2] < tf(x_1) + (1-t)f(x_2).$$

(a) Using the definition of convexity, show that $\text{Bridge}_\lambda(\boldsymbol{\beta})$ for $\lambda > 0$ is a convex function in $\boldsymbol{\beta}$, which is strictly convex for $q > 1$.

Definition: A function $f(x) : \mathcal{X} \to \mathbb{R}$ is *convex* if and only if for all $0 < t < 1$ and $x_1, x_2 \in \mathcal{X}$,

$$f[tx_1 + (1-t)x_2] \leq tf(x_1) + (1-t)f(x_2).$$

$$\text{Bridge}_\lambda(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|^q$$

$f(x)$ is *strictly convex* if and only if for all $0 < t < 1$ and $x_1, x_2 \in \mathcal{X}$ such that $x_1 \neq x_2$,

$$f[tx_1 + (1-t)x_2] < tf(x_1) + (1-t)f(x_2).$$

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

Let $f_1(\beta) = (y - X\beta)^T(y - X\beta)$

$\because f_1(\beta) = (y - X\beta)^T(y - X\beta) = 2X^TX \geq 0 \quad \forall x \in \mathbb{R}$ (X^TX is gram matrix $\Rightarrow$ positive semi matrix)

$\therefore f_1(\beta)$ is a convex quadratic term

By definition, $\beta_1, \beta_2 \in \mathbb{R}^p$ & $t \in [0,1]$

$f_1(t\beta_1 + (1-t)\beta_2) \leq tf_1(\beta_1) + (1-t)f_1(\beta_2)$

$f_1(t\beta_1 + (1-t)\beta_2) = (y - X(t\beta_1 + (1-t)\beta_2))^T(y - X(t\beta_1 + (1-t)\beta_2))$

$= (y - tX\beta_1 - (1-t)X\beta_2)^T(y - tX\beta_1 - (1-t)X\beta_2)$

$= y^Ty - 2y^T(tX\beta_1 + (1-t)X\beta_2) + (tX\beta_1 + (1-t)X\beta_2)^T(tX\beta_1 + (1-t)X\beta_2)$

$f_1(\beta_1) = (y - X\beta_1)^T(y - X\beta_1)$ & $f_1(\beta_2) = (y - X\beta_2)^T(y - X\beta_2)$

$\therefore f_1(t\beta_1 + (1-t)\beta_2) \leq tf_1(\beta_1) + (1-t)f_1(\beta_2)$

$\Rightarrow f_1(\beta) = (y - X\beta)^T(y - X\beta)$ is convex function

Let $f_2(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|^q$, $\beta_1, \beta_2 \in \mathbb{R}^p$ $t \in [0,1]$

To show $\lambda \sum_{j=1}^{p} |t\beta_{1j} + (1-t)\beta_{2j}|^q \leq t\lambda \sum_{j=1}^{p} |\beta_{1j}|^q + (1-t)\lambda \sum_{j=1}^{p} |\beta_{2j}|^q$

We know $\lambda > 0$ so we can prove $|t\beta_{1j} + (1-t)\beta_{2j}|^q \leq t|\beta_{1j}|^q + (1-t)|\beta_{2j}|^q$

For $q=1$, $|t\beta_{1j}+(1-t)\beta_{2j}| \leq t|\beta_{1j}|+(1-t)|\beta_{2j}|$ ①

by trangle equility, ① holds so we already proved $f_2(\beta)$'s convexity

For $q>1$, by Jensen's inequality, $g(t\beta_{1j}+(1-t)\beta_{2j}) < tg(\beta_{1j})+(1-t)g(\beta_{2j})$

where $g(\beta_j)=|\beta_j|^q$ $\quad g'(\beta_j)=q|\beta_j|^{q-1}$

$g''(\beta_j)=q(q-1)|\beta_j|^{q-2}>0$ for $q>1$

$\therefore g(\beta_j)$ is strictly convex

$\therefore |t\beta_{1j}+(1-t)\beta_{2j}|^q < t|\beta_{1j}|^q+(1-t)|\beta_{2j}|^q$

$\Rightarrow f_2(\beta)$ is convex for $q=1$ & strictly convex for $q>1$

As a result, $\text{Bridge}_\lambda(\beta)$ is a convex function when $\lambda>0$ & $q\geq 1$

is strictly convex when $\lambda>0$ & $q>1$

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

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

$\lambda\sum_{j=1}^{p}|\beta_j|^q$ increases without bound as $\|\beta\|$ increases,

so there must exist at least one minimizer $\hat{\beta}(\lambda)$

From (a), we know when $q>1$, $\text{Bridge}_\lambda(\beta)$ is strictly convex, thus

the function has only one point where gradient is 0, where is an unique minimizer.

t_0

We know as λ increases, $\sum_{j=1}^{p}|\hat{\beta}_j(\lambda)|^q$ shrinks.

Therefore, for a large enough λ, $\sum_{j=1}^{p}|\hat{\beta}_j(\lambda)|^q \leq t_0$ satisfies.

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

(c) Show that for $q = 1$ there exists a minimizer and for all minimizers, $\hat{\boldsymbol{\beta}}(\lambda)$, the penalty function takes the same value

$$s(\lambda) \triangleq \sum_{j=1}^{p} |\hat{\beta}_j(\lambda)|^q \leq t_0.$$

Thus for $q \geq 1$, $s(\lambda)$ is well defined as a function of λ on the interval $(0, \infty)$.

for $q=1$, $f_\lambda(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|$ ~ LASSO penalty & we know $\frac{\partial^2 f_\lambda(\beta)}{\partial \beta^2} \geq 0$ $(\lambda > 0)$

so there exist a unique minimizer under certain conditions

Assume there are 2 distinct minimizer $\hat{\beta}_1(\lambda)$ & $\hat{\beta}_2(\lambda)$ which satisfies

$s_1(\lambda) = \sum_{j=1}^{p} |\hat{\beta}_{1j}(\lambda)|$ & $s_2(\lambda) = \sum_{j=1}^{p} |\hat{\beta}_{2j}(\lambda)|$ and $s_1(\lambda) \neq s_2(\lambda)$

$\tilde{\beta}(\lambda) = t\beta_1(\lambda) + (1-t)\beta_2(\lambda) \quad \forall t [0,1]$

we know $f_\lambda(\beta)$ is convex, so $\text{Bridge}_\lambda(\tilde{\beta}(\lambda)) \leq t\,\text{Bridge}_\lambda(\hat{\beta}_1(\lambda)) + (1-t)\,\text{Bridge}_\lambda(\hat{\beta}_2(\lambda))$

By L1 norm, $\sum_{j=1}^{p} |\tilde{\beta}_j(\lambda)| = \sum_{j=1}^{p} |t\hat{\beta}_{1j}(\lambda) + (1-t)\hat{\beta}_{2j}(\lambda)|$

$\sum_{j=1}^{p} |\tilde{\beta}_j(\lambda)| \leq t \sum_{j=1}^{p} |\hat{\beta}_{1j}(\lambda)| + (1-t) \sum_{j=1}^{p} |\hat{\beta}_{2j}(\lambda)|$

where $\sum_{j=1}^{p} |\hat{\beta}_{1j}(\lambda)| \neq \sum_{j=1}^{p} |\hat{\beta}_{2j}(\lambda)|$

$\therefore \sum_{j=1}^{p} |\tilde{\beta}_j(\lambda)| < t \sum_{j=1}^{p} |\hat{\beta}_{1j}(\lambda)| + (1-t) \sum_{j=1}^{p} |\hat{\beta}_{2j}(\lambda)|$

$\tilde{\beta}(\lambda)$ should be a minimizer but it still has smaller penalty value than $\hat{\beta}_1(\lambda)$ & $\hat{\beta}_2(\lambda)$

$\therefore$ All minimizer must give the same value.

(d) Show that minimizing $\text{Bridge}_\lambda(\boldsymbol{\beta})$ is equivalent to minimizing

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad \text{subject to } \sum_{j=1}^{p} |\beta_j|^q \leq s(\lambda).$$

To introduce a Lagrange multiple λ for constrainted optimization

$L(\beta,\lambda) = (y-X\beta)^T(y-X\beta) + \lambda(\sum_{j=1}^{p} |\beta j|^q - s(\lambda))$ where $\lambda > 0$

$\min L(\beta,\lambda) = \min\limits_{\beta} [(y-X\beta)^T(y-X\beta) + \lambda \sum_{j=1}^{p} |\beta j|^q - \lambda s(\lambda)]$

where there is a trade-off between min error & $\sum_{j=1}^{p} |\beta j|^q \leq s(\lambda)$

If there is no constraint, $\min\limits_{\beta} ((y-X\beta)^T(y-X\beta) + \lambda \sum_{j=1}^{p} |\beta j|^q)$

Since they are equivalent with or without constraint,

$\min\limits_{\beta} ((y-X\beta)^T(y-X\beta) + \lambda \sum_{j=1}^{p} |\beta j|^q) = \min\limits_{\beta} (y-X\beta)^T(y-X\beta)$ subject to $\sum_{j=1}^{p} |\beta j|^q \leq s(\lambda)$

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

(c) Show that for $q = 1$ there exists a minimizer and for all minimizers, $\hat{\boldsymbol{\beta}}(\lambda)$, the penalty function takes the same value

$$s(\lambda) \triangleq \sum_{j=1}^{p} |\hat{\beta}_j(\lambda)|^q \leq t_0.$$

Thus for $q \geq 1$, $s(\lambda)$ is well defined as a function of λ on the interval $(0, \infty)$.

(d) Show that minimizing $\text{Bridge}_\lambda(\boldsymbol{\beta})$ is equivalent to minimizing

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad \text{subject to} \sum_{j=1}^{p} |\beta_j|^q \leq s(\lambda).$$

2. (6 points) In this exercise, you will investigate the potential inconsistency problem of bootstrap. Let $X_j, j = 1, \ldots, p$, be p independent normal random variables with mean $\mu_j, j = 1, \ldots, p$, and standard deviation 1. Let $\mu_{max} = \max\{\mu_j : j = 1, \ldots, p\}$.

For given p, $\mu_j, j = 1, \ldots, p$, and a training set, the procedure below describes how to construct bootstrap confidence intervals for μ_{max}:

i Compute $\bar{X}_{max} = \max\{\bar{X}_j : j = 1, \ldots, p\}$, where $\bar{X}_j$ is the sample mean of $X_j, j = 1, \ldots, p$. Then $\bar{X}_{max}$ is an estimate of μ_{max} based on the training sample.

ii Generate $B = 1000$ bootstrap samples from the training sample. For each bootstrap sample b, compute the estimate of μ_{max} as previously. Denote the estimates as $\bar{X}_{max}^{(b)}, b = 1, \ldots, B$.

iii Construct 95% confidence intervals for μ_{max} using the three bootstrap inference methods presented in Lecture 3 (page 16 of lecture notes).

Now, consider scenarios (a) - (f):

(a) $p = 2$, $\mu_j = 1, j = 1, \ldots, p$.

(b) $p = 2$, $\mu_j = j, j = 1, \ldots, p$.

(c) $p = 5$, $\mu_j = 1, j = 1, \ldots, p$.

(d) $p = 5$, $\mu_j = j, j = 1, \ldots, p$.

(e) $p = 10$, $\mu_j = 1, j = 1, \ldots, p$.

(f) $p = 10$, $\mu_j = j, j = 1, \ldots, p$.

For each scenario, conduct simulations to investigate the performance of the bootstrap confidence intervals as follows.

i Generate $M = 1000$ training sets, each consisting $n = 100$ observations of $(X_1, \ldots, X_p)$.

ii For each training set, use the procedure described above to construct three confidence intervals for μ_{max}, corresponding to the three bootstrap inference methods.

iii For each bootstrap inference method, you will obtain $M = 1000$ confidence intervals, one from each training set. Compute the converge rate (i.e., the proportion of times that μ_{max} lies in the confidence intervals out of the 1000 replications). If the coverage rate is close to the nominal level of 95%, then bootstrap is consistent (i.e., valid); otherwise, bootstrap is not consistent.

Present the confidence interval coverage rate for each of the three bootstrap inference methods under each of the scenarios (a)-(f) and discuss your results.

3. (6 points) The prostate data described in Chapter 3 of [ESL] have been divided into a training set of size 67 and a test set of size 30. (https://hastie.su.domains/ElemStatLearn/data.html)

   Carry out the following analyses on the training set:

   (a) Best-subset linear regression with k chosen by 5-fold cross-validation.

   (b) Best-subset linear regression with k chosen by BIC.

   (c) Lasso regression with λ chosen by 5-fold cross-validation.

   (d) Lasso regression with λ chosen by BIC.

   For each analysis, compute and plot the cross-validation or BIC estimates of the prediction error as the model complexity increases as in Figure 3.7 (page 62 of [ESL]). Report the final estimated model as well as the test error and its standard error over the test set as in Table 3.3 (page 63 of [ESL]). Briefly discuss your results. (note: Std Error is calculated as $SD\big\{(Y_i-\hat{f}(X_i))^2, i \in \text{test set}\big\}/\sqrt{n_{test}}$).