Exercise of Supervised Learning: Regularization Part 2

Yawei Li

yawei.li@stat.uni-muenchen.de

Date

Exercise 1

For a design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and the vector of targets $\mathbf{y} \in \mathbb{R}^n$, consider Lasso regression, i.e.,

$$\arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} 0.5||\mathbf{X}\boldsymbol{\theta} - \mathbf{y}||_2^2 + \lambda||\boldsymbol{\theta}||_1.$$

where $\lambda > 0$ is the regularization parameter.
(a) Since there exists no analytical solution to Lasso regression in general, we want to find a procedure similar to gradient descent that should converge to the true solution.

We go through the questions in subsequent slides. (Write the optimization objective on whiteboard.)

Exercise 1 (a) (i)

Question (a) (i): Explain why $\mathcal{R}_{\text{reg}}$ is not differentiable.

Solution: Because $\lambda||\boldsymbol{\theta}||_1$ is not differentiable at $\boldsymbol{\theta} = \mathbf{0}$.

Exercise 1 (a) (ii)

Question (a) (ii): Show that $\mathcal{R}_{\text{reg}}$ is convex. *Hint: The sum of convex function is convex.*

Solution:

1. $0.5||\mathbf{X}\boldsymbol{\theta} - \mathbf{y}||_2^2$ is convex, since it is quadratic.
2. $\lambda||\theta_1||_1$ is also convex (since it is a norm).
3. Sum of convex functions is convex. So $\mathcal{R}_{\text{reg}}$ is convex.

Exercise 1 (a) (iii)

Question (a) (iii): Find $\rho_j, z_j \in \mathbb{R}$ for which

$$0.5 \frac{\partial}{\partial \theta_j} ||\mathbf{X}\boldsymbol{\theta} - \mathbf{y}||_2^2 = -\rho_j + \theta_j z_j.$$

Solution:

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} ||\mathbf{X}\boldsymbol{\theta} - \mathbf{y}||_2^2 &= \frac{\partial}{\partial \theta_j} 0.5 \sum_{i=1}^{n} \left(y^{(i)} - \sum_{k=1}^{p} \mathbf{x}_k^{(i)} \theta_k \right)^2 \\
&= -\sum_{i=1}^{n} \mathbf{x}_j^{(i)} \left(y^{(i)} - \sum_{k=1}^{p} \mathbf{x}_k^{(i)} \theta_k \right) \\
&= \underbrace{-\sum_{i=1}^{n} \mathbf{x}_j^{(i)} \left(y^{(i)} - \sum_{k \neq j} \mathbf{x}_k^{(i)} \theta_k \right)}_{=\rho_j} + \theta_j \underbrace{\sum_{i=1}^{n} \left(\mathbf{x}_j^{(i)} \right)^2}_{=z_j}
\end{aligned}$$

Exercise 1 (a) (iv)

Question (a) (iv): In this situation, we can use the so-called sub-derivative which we denote with ∂f for a real-valued convex continuous function f. The subderivative maps a point $\theta \in \mathbb{R}$ to and interval

- and if f is differentiable at $\tilde{\theta} \in \mathbb{R}$, then $\partial f(\tilde{\theta}) = \{\frac{\mathrm{d}}{\mathrm{d}\theta} f(\tilde{\theta})\}$,
- and for $f(\theta) = \lambda|\theta|$ and $\lambda > 0$, it holds that $\partial f(\theta) = \begin{cases} \{-\lambda\} & \text{for } \theta < 0, \\ [-\lambda, \lambda] & \text{for } \theta = 0, \\ \{\lambda\} & \text{for } \theta > 0 \end{cases}$
- and for f, g real-valued convex functions with $\partial f(\tilde{\theta}) = [a, b], \partial g(\tilde{\theta}) = [c, d]$,

$$\partial(f + g)(\tilde{\theta}) = [a + c, b + d]$$

  where $b \geq a$ and $d \geq c$.

With this compute teh sub-derivative of $\mathcal{R}_{\text{reg},\boldsymbol{\theta}_{\neq j}}$ w.r.t. θ_j, i.e., $\partial_{\theta_j} \mathcal{R}_{\text{reg},\boldsymbol{\theta}_{\neq j}}$ where $\mathcal{R}_{\text{reg},\boldsymbol{\theta}_{\neq j}} : \mathbb{R} \to \mathbb{R}_{\geq 0}, \theta_j \longmapsto \mathcal{R}_{\text{reg}}(\theta_1, \ldots, \theta_j, \ldots, \theta_p)$ for constants $\boldsymbol{\theta}_{\neq j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)^T$. *Hint: Use (a) (iii).*

Exercise 1 (a) (iv): Continued

Solution: Recall that $\mathcal{R}_{\text{reg}} = 0.5||\mathbf{X}\boldsymbol{\theta} - \mathbf{y}||_2^2 + \lambda||\boldsymbol{\theta}||_1$ and $\frac{\partial}{\partial\theta_j}0.5||\mathbf{X}\boldsymbol{\theta} - \mathbf{y}||_2^2 = -\rho_j + \theta_j z_j$. Therefor,

$$\partial_{\theta_j}\mathcal{R}_{\text{reg},\boldsymbol{\theta}_{\neq j}} = \begin{cases} \{-\rho_j + \theta_j z_j - \lambda\} & \text{for } \theta_j < 0, \\ [-\rho_j - \lambda, -\rho_j + \lambda] & \text{for } \theta_j = 0; \\ \{-\rho_j + \theta_j z_j + \lambda\} & \text{for } \theta_j > 0. \end{cases}$$

Exercise 1 (a) (v)

Question (a) (v): For a real-valued convex function f, the global minimum (if it exists) can be characterized in the following way:
A point $\theta^* \in \mathbb{R}$ is the global minimum of f if and only if $0 \in \partial f(\theta^*)$. With this show that

$$\theta_j^* \in \arg\min_{\theta_j \in \mathbb{R}} \mathcal{R}_{\text{reg},\boldsymbol{\theta}_{\neq j}} \Leftrightarrow \theta_j^* = \begin{cases} \frac{\rho_j + \lambda}{z_j} & \text{for } \rho_j < -\lambda \\ 0 & \text{for } -\lambda \leq \rho_j \leq \lambda \\ \frac{\rho_j - \lambda}{z_j} & \text{for } \rho_j > \lambda \end{cases}$$

Exercise 1 (a) (v): Continued

Solution: we need to show that

$$0 \in \partial_{\theta_j} \mathcal{R}_{\text{reg},\boldsymbol{\theta}_{\neq j}} = \begin{cases} \{-\rho_j + \theta_j z_j - \lambda\} & \text{for } \theta_j < 0, \\ [-\rho_j - \lambda, -\rho_j + \lambda] & \text{for } \theta_j = 0; \\ \{-\rho_j + \theta_j z_j + \lambda\} & \text{for } \theta_j > 0. \end{cases}$$

where $z_j = \sum_{i=1}^{n} \left(\mathbf{x}_j^{(i)}\right)^2 \geq 0$.

- If $\theta_j < 0$, then $-\rho_j + \theta_j z_j - \lambda = 0$ leads to $\theta_j^* = \frac{\rho_j + \lambda}{z_j}$. Since $\theta_j^* < 0$ and $z_j \geq 0$, we have $\rho_j < -\lambda$;
- If $\theta_j = 0$, then $\theta_j^* = 0$, and $-\rho_j - \lambda \leq 0 \leq -\rho_j + \lambda$ leads to $-\lambda \leq \rho_j \leq \lambda$, ;
- If $\theta_j > 0$ then $-\rho_j + \theta_j z_j + \lambda = 0$ leads to $\theta_j^* = \frac{\rho_j - \lambda}{z_j}$. Since $\theta_j^* > 0$, we have $\rho_j > \lambda$.

Exercise 1 (a) (v): Continued

Summarize everthing up:

$$\theta_j^* \in \arg\min_{\theta_j \in \mathbb{R}} \mathcal{R}_{\text{reg},\boldsymbol{\theta}_{\neq j}} \Leftrightarrow \theta_j^* = \begin{cases} \frac{\rho_j + \lambda}{z_j} & \text{for } \rho_j < -\lambda \\ 0 & \text{for } -\lambda \leq \rho_j \leq \lambda \\ \frac{\rho_j - \lambda}{z_j} & \text{for } \rho_j > \lambda \end{cases}$$

Exercise 1 (a) (vi)

Question (a) (vi): Plot θ_j^* as a function of ρ_j for $\rho_j \in [-5, 5], \lambda = 1, z_j = 1$. (This function is called soft thresholding operator).

Solution: **Show the standard solution.**

Exercise 1 (b)

Question (b): Find for non-singular $\mathbf{X}^T\mathbf{X}$ the matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ for which

$$\mathbf{A}^T\mathbf{X}^T\mathbf{X}\mathbf{A} = \mathbf{I}$$

Hint: $\mathbf{X}^T\mathbf{X}$ is positive definite.
Solution: Since $\mathbf{X}^T\mathbf{X}$ is positive definite, there exist an orthogonal matrix $\mathbf{V}$ and a diagonal matrix $\mathbf{D}$ with $D_{ii} > 0$ such that

$$\mathbf{V}\mathbf{D}\mathbf{V}^T = \mathbf{X}^T\mathbf{X}.$$

So

$$\mathbf{A}^T\mathbf{X}^T\mathbf{X}\mathbf{A} = \mathbf{A}^T\mathbf{V}\mathbf{D}\mathbf{V}^T\mathbf{A} = \mathbf{I}$$

To solve the equation for $\mathbf{A}$, we substitute the variable $\mathbf{V}^T\mathbf{A} = \mathbf{P}$. Then,

$$\mathbf{P}^T\mathbf{D}\mathbf{P} = \mathbf{I}$$

Since $\mathbf{D}$ is a diagnonal matrix, it can be seen that $\mathbf{P} = \mathbf{D}^{-0.5}$. Hence $\mathbf{V}^T\mathbf{A} = \mathbf{P}$. Because $\mathbf{V}$ is orthogonal, we have $\mathbf{A} = \mathbf{V}\mathbf{D}^{-0.5}$.

Exercise 1 (c) (i)

(c) For a design matrix with orthornormal columns, i.e., $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, exists an analytical minimizer of the Lasso regression $\hat{\boldsymbol{\theta}}_{Lasso} = (\hat{\theta}_{Lasso,1}, \ldots, \hat{\theta}_{Lasso,p})^T$ that is given by

$$\hat{\theta}_{Lasso,i} = \mathrm{sign}(\hat{\theta}_i) \max\left\{\left|\hat{\theta}_i - \lambda, 0\right|\right\}, \quad i = i, \ldots, p,$$

where $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is the minimizer of the unregularized empirical risk (w.r.t. the L2 loss).
Under the assumption that $\mathbf{X}^T\mathbf{X}$ is non-singular, your colleague proposes to project $\mathbf{X}$ with $\mathbf{A}$ from (b), i.e., use $\tilde{\mathbf{A}} = \mathbf{XA}$ and then apply the analytical solution given here.

(i) Show that this does not generally solve the original Lasso regression.

Hint: You only need to check under which condition $\nabla_{\boldsymbol{\theta}} 0.5||\mathbf{X}\boldsymbol{\theta} - \mathbf{y}||_2^2 = \nabla_{\boldsymbol{\theta}} 0.5||\mathbf{XA}\boldsymbol{\theta} - \mathbf{y}||_2^2$. *The proof can be finished with a subgradient argument regarding stationarity, which you do not have to do.*

Exercise 1 (c) (i): Continued

Solution:

$$\nabla_{\boldsymbol{\theta}} 0.5||\mathbf{XA}\boldsymbol{\theta} - \mathbf{y}||_2^2 = \mathbf{AX}^T\mathbf{XA}\boldsymbol{\theta} - \mathbf{A}^T\mathbf{X}^T\mathbf{y}$$
$$\nabla_{\boldsymbol{\theta}} 0.5||\mathbf{X}\boldsymbol{\theta} - \mathbf{y}||_2^2 = \mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - \mathbf{X}^T\mathbf{y}$$

To let the gradient be equal, it must be satisfied that

$$\mathbf{A} = \mathbf{I}.$$

In other words, if $\mathbf{A} \neq \mathbf{I}$, then the analytical solution does not solve the original Lasso regression.

Exercise 1 (c) (ii)

Question (c) (ii): How could you adapt the penalty term such that the solution to the projected problem is equivalent to the original Lasso regression? In this case, can we still solve for parameters independently?
TODO: It is weird in the solution that chosing $||\mathbf{A}\boldsymbol{\theta}||_1$ as regularization term. Because this will yeild $\boldsymbol{\theta}^* = \mathbf{A}^T\boldsymbol{\theta}_{\text{Orignal Lasso}}$, and it is different as $\boldsymbol{\theta}_{\text{Original Lasso}}$.

Exercise 1 (c) (iii)

Question (c) (iii): Does the procedure proposed in (c) perform variable selection?

Solution:

- We can see this procedure as a projection of variables followed by a Lasso regression.
- Hence, we select variables in these projected coordinates.
- But this does not imply that the solution in the original coordinates $\mathbf{A}\theta^*$ will be sparse.

Exercise 1 (c) (iii)

Question (c) (iii): Does the procedure proposed in (c) perform variable selection?

Solution:

- We can see this procedure as a projection of variables followed by a Lasso regression.
- Hence, we select variables in these projected coordinates.
- But this does not imply that the solution in the original coordinates $\mathbf{A}\boldsymbol{\theta}^*$ will be sparse.

Exercise 1 (d)

(d) You are given the following code to compare the quality of the projected Lasso regression vs. the regular Lasso regression. Complete the missing code of the algorithms and interpret the result.

Solution: show the sandard solution.