Homework 1 STOR 767: Theory Part

Due Sep 8, 2020

Student Name: Leo Li (PID: 730031954)

Instructions

- Edit this LaTeX file with your solutions and generate a PDF file from it. Upload **both the tex and the pdf** file to Sakai.
- Use proper fonts for a clear presentation:
  x for an observed value; X for a random variable; $\boldsymbol{X}$ for a vector; $\mathbf{X}$ for a matrix.
- You are allowed to work with other students but homework should be in your own words. Identical solutions will receive a **0** in grade and will be investigated.

1. Consider the ridge and LASSO regression on $\boldsymbol{y}_{n\times 1}$ and $\mathbf{X}_{n\times d}$ where $d \leq n$ and $\mathbf{X}$ has orthonormal columns. For the j-th covariate, $1 \leq j \leq d$, derive $\hat{\boldsymbol{\beta}}^{\lambda}_{Ridge,j}$ and $\hat{\boldsymbol{\beta}}^{\lambda}_{LASSO,j}$ as functions of $\hat{\boldsymbol{\beta}}_{LS}$ for fixed λ.

SOLUTION.

- **Ridge Estimator:** If the design matrix $\mathbf{X}$ has orthonormal columns, then the ridge estimator to the j^{th} covariate can be written as $\hat{\boldsymbol{\beta}}^{\lambda}_{Ridge,j} = \hat{\boldsymbol{\beta}}_{LS,j}/(1+\lambda)$. The reason is that, if the design matrix $\mathbf{X}$ has orthonormal columns, then we have that $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, and the least square estimator can be expressed as:

$$\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y} = \mathbf{X}^T\boldsymbol{y}$$

In addition, the ridge estimator can be written as

$$\hat{\boldsymbol{\beta}}^{\lambda}_{Ridge} = argmin_{\boldsymbol{\beta}\in\Re^d}(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

The above optimization problem can be solved by taking derivative of the objective function with respect to $\boldsymbol{\beta}$:

$$\frac{\partial}{\partial\boldsymbol{\beta}}\{(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}\} = 2\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}-\boldsymbol{y}) + 2\lambda\boldsymbol{\beta} = \mathbf{0}$$

Solving the above equation gives that $\hat{\boldsymbol{\beta}}^{\lambda}_{Ridge} = (\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{y}$. In the situation in which the design matrix has the orthonormal columns (i.e., $\mathbf{X}^T\mathbf{X} = \mathbf{I}$), we have that,

$$\hat{\boldsymbol{\beta}}^{\lambda}_{Ridge} = \mathbf{X}^T\boldsymbol{y}/(1+\lambda) = \hat{\boldsymbol{\beta}}_{LS}/(1+\lambda)$$

so that we have, $\hat{\boldsymbol{\beta}}^{\lambda}_{Ridge,j} = \hat{\boldsymbol{\beta}}_{LS,j}/(1+\lambda)$, as desired. $\square$

- **LASSO Estimator:** If the design matrix $\mathbf{X}$ has orthonormal columns, then the LASSO estimator to the j^{th} covariate can be written as $\hat{\boldsymbol{\beta}}^{\lambda}_{LASSO,j} = sign(\hat{\boldsymbol{\beta}}_{LS,j})(|\hat{\boldsymbol{\beta}}_{LS,j}|-\lambda)_{+}$. In other words, the LASSO estimator of the j^{th} covariate is linked to the least square estimator through a soft thresholding function. The reason is that, if the design matrix $\mathbf{X}$ has orthonormal columns, the we have that $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, and the least square estimator can be expressed as:

$$\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y} = \mathbf{X}^T\boldsymbol{y}$$

In addition, the LASSO estimator can be written as

$$\hat{\boldsymbol{\beta}}^{\lambda}_{LASSO} = argmin_{\boldsymbol{\beta} \in \Re^d} \frac{1}{2} ||\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1.$$

In the above optimization problem, although the objective function, $f(.)$, is non-differentiable, it is indeed convex, which can be minimized at the point in which $0 \in \partial f$. Hence, let $\hat{\boldsymbol{\beta}}$ be the solution to the optimization problem, we have that,

$$\mathbf{X}^T(\mathbf{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}) + \lambda \times sign(\hat{\boldsymbol{\beta}}) = \mathbf{0}$$

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}^T\boldsymbol{y} + \lambda \times sign(\hat{\boldsymbol{\beta}}) = \mathbf{0}$$

Since the design matrix has orthonormal columns, $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, so that we have,

$$\hat{\boldsymbol{\beta}} = \mathbf{X}^T\boldsymbol{y} - \lambda \times sign(\hat{\boldsymbol{\beta}})$$

so that,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_j &= \begin{cases} (\mathbf{X}^T\boldsymbol{y})_j + \lambda \text{ , if } (\mathbf{X}^T\boldsymbol{y})_j < -\lambda \\ \mathbf{0} \text{ , if } |(\mathbf{X}^T\boldsymbol{y})_j| < \lambda \\ (\mathbf{X}^T\boldsymbol{y})_j - \lambda \text{ , if } (\mathbf{X}^T\boldsymbol{y})_j > \lambda \end{cases} \\ &= sign((\mathbf{X}^T\boldsymbol{y})_j)(|(\mathbf{X}^T\boldsymbol{y})_j| - \lambda)_+ \\ &= sign(\hat{\boldsymbol{\beta}}_{LS,j})(|\hat{\boldsymbol{\beta}}_{LS,j}| - \lambda)_+, \end{aligned}$$

as desired. $\square$

2. State and prove the Hoeffding's inequality for i.i.d. bounded variables.

SOLUTION.

Hoeffding's inequality states that, if $X_1, ..., X_n$ are bounded i.i.d. random variables such that $|X_i| \leq c \leq \infty$, then we have that,

$$P(|\frac{1}{n}\sum_{i=1}^{n} X_i - E(X_i)| > t) \leq 2\exp(-\frac{nt^2}{2c^2}).$$

PROOF.

Before proving Hoeffding's inequality, let us start by proving the following claim:

- **Claim:** If A is a bounded random variable such that $|A| \leq 1$ and $E(A) = 0$. Then, for any constant $\lambda > 0$, we have that $E(\exp(\lambda A)) \leq \exp(\lambda^2/2)$.

- **Proof:** Let $p = (1 + A)/2$, which is bounded by $[0, 1]$. Since $\exp(\lambda A)$ is a convex function of λ, then we have that,

$$\exp(\lambda A) = \exp(\lambda(\frac{1+A}{2} - \frac{1-A}{2})) = \exp(p\lambda + (1-p)(-\lambda)) \leq p\exp(\lambda) + (1-p)\exp(-\lambda).$$

By simplifying the expression, we have that,

$$p\exp(\lambda)+(1-p)\exp(-\lambda) = \frac{1+A}{2}\exp(\lambda)+(1-\frac{1+A}{2})\exp(-\lambda) = \frac{e^{\lambda}+e^{-\lambda}}{2}+\frac{A(e^{\lambda}-e^{-\lambda})}{2}.$$

Then, since $E(A) = 0$, we have that $E(\exp(\lambda A)) \leq \frac{e^{\lambda}+e^{-\lambda}}{2}$. We also have that,

$$\frac{e^{\lambda}+e^{-\lambda}}{2} = \frac{1}{2}\sum_{n=0}^{\infty}\frac{\lambda^n}{n!} + \frac{1}{2}\sum_{n=0}^{\infty}\frac{(-\lambda)^n}{n!} = \frac{1}{2}\sum_{n=0}^{\infty}\frac{\lambda^{2n}}{(2n)!} \leq \sum_{n=0}^{\infty}\frac{\lambda^{2n}}{2^n(n!)} = \sum_{n=0}^{\infty}\frac{(\lambda^2/2)^n}{(n!)} = e^{\lambda^2/2},$$

where the inequality holds because,

$$2(2n)! = 2\prod_{i=1}^{2n} i \geq \prod_{i=1}^{n}(2i) = 2^n(n!).$$

Therefore, we have that $E(\exp(\lambda A)) \leq \exp(\lambda^2/2)$, as desired. □

Once we finish proving this useful claim, we can formally start to prove Hoeffding's inequality. Let $\tilde{X}_i = X_i/c$ and $\tilde{X} = \sum_{i=1}^{n}\tilde{X}_i$, so that $\tilde{X}_i \in [-1, 1]$. Also let $X_i^* = \tilde{X}_i - E(\tilde{X}_i)$, and let

$X^* = \sum_{i=i}^n X_i^*$, so we have that $E(X_i^*) = 0$ and $X_i^* \in [-1, 1]$. In addition, let $Y_i = \exp(\lambda X_i^*)$, and $Y = \exp(\lambda X^*)$. Now, we have that,

$$Y = \exp(\lambda X^*) = \exp(\lambda \sum_{i=i}^n X_i^*) = \prod_{i=1}^n \exp(\lambda X_i^*) = \prod_{i=1}^n Y_i,$$

and that,

$$E(Y_i) = E(\exp(\lambda X_i^*)) \leq \exp(\lambda^2/2) \text{ (by the claim that we proved earlier.)}$$

Since $\{X_1, ..., X_n\}$ is mutually independent, so is $\{\tilde{X}_1, ..., \tilde{X}_n\}$, $\{X_1^*, ..., X_n^*\}$, and $\{Y_1, ..., Y_n\}$. By independence, we have that

$$E(Y) = \prod_{i=1}^n E(Y_i) = \prod_{i=1}^n E(\exp(\lambda X^*)) \leq \prod_{i=1}^n \exp(\lambda^2/2) = \exp(\lambda^2 n/2).$$

Now, let us consider the probability that $X^* \geq t$:

$$\begin{aligned} P(X^* \geq t) &= P(\exp(\lambda X^*) \geq \exp(\lambda t)) \text{ (because } e^{\lambda x} \text{ is a monotone function of } x.) \\ &\leq \frac{E(\exp(\lambda X^*))}{\exp(\lambda t)} \text{ (by Markov Inequality.)} \\ &= E(Y) \times \exp(-\lambda t) \text{ (by definition of } Y.) \\ &\leq \exp(\lambda^2 n/2 - \lambda t) \text{ (by the inequality proved above.)} \\ &= \exp(-\frac{t^2}{2n}) \text{ (by optimization and get } \lambda = t/n.) \end{aligned}$$

Now, we have already showed that $P(X^* \geq t) = \exp(-\frac{t^2}{2n})$. By using similar reasoning, we can show that $P(-X^* \geq t) = \exp(-\frac{t^2}{2n})$. Then, we have that,

$$P(|\tilde{X} - E(\tilde{X})| \geq t) = P(|X^*| \geq t) = P(X^* \geq t) + P(X^* \leq t) \leq 2\exp(-\frac{t^2}{2n})$$

Since $\tilde{X} = \sum_{i=1}^n \tilde{X}_i = \frac{1}{c} \sum_{i=1}^n X_i$, we have that,

$$\begin{aligned} P(|\tilde{X} - E(\tilde{X})| \geq t) &= P(|\frac{1}{c} \sum_{i=1}^n X_i - E(\frac{1}{c} \sum_{i=1}^n X_i)| \geq t) = P(|\frac{1}{c} \sum_{i=1}^n X_i - \frac{n}{c} E(X_i)| \geq t) \\ &= P(|\frac{1}{n} \sum_{i=1}^n X_i - E(X_i)| \geq \frac{c}{n} t) \leq 2\exp(-\frac{t^2}{2n}) \end{aligned}$$

Take $t' = \frac{c}{n} t$ then $t = \frac{n}{c} t'$, we have that,

$$P(|\frac{1}{n} \sum_{i=1}^n X_i - E(X_i)| \geq t') \leq 2\exp(-\frac{(\frac{n}{c}t')^2}{2n}) = 2\exp(-\frac{nt'^2}{2c^2}),$$

as desired. $\square$