ORF 525: Statistical Foundations of Data Science

Jianqing Fan — Frederick L. Moore'18 Professor of Finance

Problem Set #2 Spring 2024

Due Friday, February 23 2024.

1. Consider the Lasso problem $\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1$, where $\lambda>0$ is a tuning parameter.

   (a) Let $\widehat{\boldsymbol{\beta}}$ be a minimizer of the Lasso problem with j^{th} componenet $\widehat{\beta}_j$. Denote $\mathbf{X}_j$ to be the j-th column of $\mathbf{X}$. Show that

   $$\begin{cases} \lambda = n^{-1}\mathbf{X}_j^T(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) & \text{if } \widehat{\beta}_j>0; \\ \lambda = -n^{-1}\mathbf{X}_j^T(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) & \text{if } \widehat{\beta}_j<0; \\ \lambda \geq |n^{-1}\mathbf{X}_j^T(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}})| & \text{if } \widehat{\beta}_j=0. \end{cases}$$

   (b) If $\lambda > \|n^{-1}\mathbf{X}^T\mathbf{Y}\|_\infty$, prove that $\widehat{\boldsymbol{\beta}}_\lambda=\mathbf{0}$, where $\widehat{\boldsymbol{\beta}}_\lambda$ is the minimizer of the Lasso problem with regularization parameter λ.

   (c) If $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ are both minimizers of the Lasso problem, show that they have the same prediction, i.e., $\mathbf{X}\widehat{\boldsymbol{\beta}}_1=\mathbf{X}\widehat{\boldsymbol{\beta}}_2$.
   Hint: Consider the vector $\widehat{\boldsymbol{\beta}}_\alpha=\alpha\widehat{\boldsymbol{\beta}}_1+(1-\alpha)\widehat{\boldsymbol{\beta}}_2$ for $\alpha\in(0,1)$. Show that $Q(\widehat{\boldsymbol{\beta}}_\alpha)$ does not depend on α by using the convexity, where $Q(\boldsymbol{\beta})=\frac{1}{2n}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1$. The same convexity argument entails that the loss $L(\alpha)=\frac{1}{2n}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}_\alpha\|_2^2$ does not depend on α and conclude the result from here.

2. Risk properties of Lasso.
   Let $R_n(\boldsymbol{\beta})=\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2/n$ and $R(\boldsymbol{\beta})=ER_n(\boldsymbol{\beta})$ be the empirical and theoretical risks, and $\widehat{\boldsymbol{\beta}}=\operatorname{argmin}_{\|\boldsymbol{\beta}\|_1\leq c} R_n(\boldsymbol{\beta})$ be the Lasso estimator which estimates $\boldsymbol{\beta}_0=\operatorname{argmin}_{\|\boldsymbol{\beta}\|_1\leq c} R(\boldsymbol{\beta})$.

   (a) Consider the in-sample risk $R_n(\widehat{\boldsymbol{\beta}})$ as an estimator of optimal risk $R(\boldsymbol{\beta}_0)$. Show that

   $$|R(\boldsymbol{\beta}_0)-R_n(\widehat{\boldsymbol{\beta}})| \leq \max_{\|\boldsymbol{\beta}\|_1\leq c}|R(\boldsymbol{\beta})-R_n(\boldsymbol{\beta})| \leq (1+c)^2\|\boldsymbol{\Sigma}^*-\mathbf{S}_n^*\|_{\max},$$

   where $\mathbf{Z}=\begin{pmatrix} Y \\ \mathbf{X}\end{pmatrix}$, $\boldsymbol{\Sigma}^*=E(\mathbf{Z}\mathbf{Z}^T)$ and $\mathbf{S}_n^*=n^{-1}\sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i^T$.
   Hint: Deal with two sides of the inequality separately. For example, $R(\boldsymbol{\beta}_0)-R_n(\widehat{\boldsymbol{\beta}}) = R(\boldsymbol{\beta}_0)-R_n(\boldsymbol{\beta}_0)+R_n(\boldsymbol{\beta}_0)-R_n(\widehat{\boldsymbol{\beta}}) \geq R(\boldsymbol{\beta}_0)-R_n(\boldsymbol{\beta}_0)$.

   (b) Suppose that $\|\mathbf{X}\|_\infty \leq b$ and $|Y|\leq b$ (bounded random variables). Use Hoeffding's inequality to show $\|\boldsymbol{\Sigma}^*-\mathbf{S}_n^*\|_{\max}=O_p(\sqrt{\frac{\log p}{n}})$.

   (c) Consider the lasso of form $\widehat{\boldsymbol{\beta}}=\operatorname{argmin}\{\frac{1}{2}R_n(\boldsymbol{\beta})+\lambda\|\boldsymbol{\beta}\|_1\}$.
   If $\lambda\geq\|2n^{-1}\mathbf{X}^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}_0)\|_\infty$, under the restricted eigenvalue condition

   $$\min_{3\|\boldsymbol{\Delta}_{\mathcal{S}_0}\|_1\geq\|\boldsymbol{\Delta}_{\mathcal{S}_0^c}\|_1} n^{-1}\|\mathbf{X}\boldsymbol{\Delta}\|_2^2/\|\boldsymbol{\Delta}\|_2^2 \geq a,$$

   show that with $\widehat{\boldsymbol{\Delta}}=\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0$ and $s=|\mathrm{Supp}(\boldsymbol{\beta}_0)|$,

   $$\|\widehat{\boldsymbol{\Delta}}\|_2 \leq 8a^{-1}\sqrt{s}\lambda \qquad \text{and} \qquad \|\widehat{\boldsymbol{\Delta}}\|_1 \leq 32a^{-1}s\lambda.$$

3. Concentration inequalities.

   (a) The random vector $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is called σ-sub-Gaussian if $E \exp\big(\mathbf{a}^T \boldsymbol{\varepsilon}\big) \le \exp\big(\|\mathbf{a}\|_2^2 \sigma^2/2\big)$, $\forall \mathbf{a} \in \mathbb{R}^n$. Show that $E\boldsymbol{\varepsilon} = \mathbf{0}$ and $\mathrm{var}(\boldsymbol{\varepsilon}) \le \sigma^2 \mathbf{I}_n$.

   Hint: Expand exponential functions as infinite series (actually, you only need the condition for $\mathbf{a}$ in a small neighborhood around 0)

   (b) Suppose that the random vector $\mathbf{X} - E\mathbf{X}$ is σ-sub-Gaussian and $S_n = \mathbf{1}^T\mathbf{X} = \sum_{i=1}^n X_i$. Show that

   $$P\Big(n^{-1/2}|S_n - ES_n| \ge t\Big) \le 2\exp\Big(-\frac{t^2}{2\sigma^2}\Big), \qquad t > 0.$$

   Hint: Use Chebyshev's inequality

   $$P\Big(n^{-1/2}(S_n - ES_n) \ge t\Big) \le \exp(-xt) E\exp\Big(xn^{-1/2}(S_n - ES_n)\Big)$$

   and optimize the choice of x after using the moment generating function of sub-Gaussian distributions.

   (c) For $\mathbf{X} \in \mathbb{R}^{n\times p}$ with the j-th column denoted by $\mathbf{X}_j \in \mathbb{R}^n$, suppose that $\|\mathbf{X}_j\|_2^2 = n$ for all j, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a σ-sub-Gaussian random vector. Show that there exists a constant $C > 0$ such that

   $$P\Big(\|n^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\|_\infty > \sqrt{2(1+\delta)}\sigma\sqrt{\frac{\log p}{n}}\Big) \le Cp^{-\delta}, \qquad \forall \delta > 0.$$

   Hint: Using the same argument as part (b), we can obtain $P\{|\mathbf{b}^T\boldsymbol{\varepsilon}| \ge t\} \le 2\exp\Big(-\frac{t^2}{2\sigma^2\|\mathbf{b}\|^2}\Big)$. You can use this without proof.

4. This problem intends to show that the gradient decent method for a convex function $f(\cdot)$ is a member of majorization-minimization algorithms and has a sublinear rate of convergence in terms of function values. From now on, the function $f(\cdot)$ is convex and let $\mathbf{x}^* \in \mathrm{argmin} f(\mathbf{x})$. Here we implicitly assume the minimum can be attained at some point $\mathbf{x}^\star \in \mathbb{R}^p$.

   (a) Suppose that $f''(\mathbf{x}) \le L\mathbf{I}_p$ and $\delta \le 1/L$. Show that the quadratic function $g(\mathbf{x}) = f(\mathbf{x}_{i-1}) + f'(\mathbf{x}_{i-1})^T(\mathbf{x} - \mathbf{x}_{i-1}) + \frac{1}{2\delta}\|\mathbf{x} - \mathbf{x}_{i-1}\|^2$ is a majorization of $f(\mathbf{x})$ at point $\mathbf{x}_{i-1}$, i.e., $g(\mathbf{x}) \ge f(\mathbf{x})$ for all $\mathbf{x}$ and also $g(\mathbf{x}_{i-1}) = f(\mathbf{x}_{i-1})$.

   (b) Show that gradient step $\mathbf{x}_i = \mathbf{x}_{i-1} - \delta f'(\mathbf{x}_{i-1})$ is the minimizer of the majorized quadratic function $g(\mathbf{x})$ and hence the gradient descend method can be regarded as a member of MM-algorithms. Use (a) to show that

   $$f(\mathbf{x}_i) \le g(\mathbf{x}_i) = f(\mathbf{x}_{i-1}) - \frac{1}{2\delta}\|\mathbf{x}_i - \mathbf{x}_{i-1}\|^2.$$

   (c) Show that

   $$f(\mathbf{x}_i) \le f(\mathbf{x}^*) + \frac{1}{2\delta}(\|\mathbf{x}_{i-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}^* - \mathbf{x}_i\|^2).$$

   Hint: By convexity, $f(\mathbf{x}^*) \ge f(\mathbf{x}_{i-1}) + f'(\mathbf{x}_{i-1})^T(\mathbf{x}^* - \mathbf{x}_{i-1})$ and substitute this into the second part of part (b).

(d) Conclude using (c) that $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \|\mathbf{x}_0 - \mathbf{x}^*\|^2/(2k\delta)$, namely gradient descent converges at a sublinear rate. (**Note**: The gradient descent method converges linearly if $f(\cdot)$ is strongly convex.)

5. Let us consider the Zillow data again. We drop the first 3 columns ("(empty)", "id", "date") and treat "zipcode" as a factor variable. Now, consider the variables

   (a) "bedrooms", "bathrooms", "sqft_living", and "sqft_lot" and their interactions and the remaining 14 variables in the data, including "zipcode". (We can use *model.matrix* to expand factors into a set of dummy variables.)

   (b) Add the following additional variables to (a): $X_{12} = I(view == 0)$, $X_{13} = L^2$, $X_{13+i} = (L - \tau_i)_+^2$, $i = 1, \cdots, 9$, where τ_i is $10 * i^{th}$ percentile and L is the size of living area ("sqft_living").

   Compute and compare out-of-sample R^2 using ridge regression, Lasso (using R package `glmnet`) and SCAD (using R package `ncvreg`) with regularization parameter chosen by 10 fold cross-validation. Set a random seed by `set.seed(525)` before excuting Lasso.