ORF 525: Statistical Foundations of Data Science

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

Problem Set #3 Spring 2021

Due Monday, March 15, 2021.

1. (a) The conditional distribution is the conditional trinomial. By definition, one has

$$\mathbb{P}(Y=0|\mathbf{X}) = \mathbb{P}(Z \le c_1|\mathbf{X}) = \mathbb{P}\left(\boldsymbol{\beta}^\top\mathbf{X}+\epsilon \le c_1|\mathbf{X}\right) = \mathbb{P}\left(\epsilon \le c_1-\boldsymbol{\beta}^\top\mathbf{X}|\mathbf{X}\right) = F\left(c_1-\boldsymbol{\beta}^\top\mathbf{X}\right).$$

Similarly, we can get

$$\mathbb{P}(Y=1|\mathbf{X}) = F\left(c_2-\boldsymbol{\beta}^\top\mathbf{X}\right) - F\left(c_1-\boldsymbol{\beta}^\top\mathbf{X}\right)$$

and

$$\mathbb{P}(Y=2|\mathbf{X}) = 1-F\left(c_2-\boldsymbol{\beta}^\top\mathbf{X}\right).$$

(b) Let $\mathcal{L}(\boldsymbol{\beta}, c_1, c_2)$ be the log-likelihood function of the random sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Then we have

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\beta}, c_1, c_2)\\
&=\sum_{i=1}^n \log\left\{\left[F\left(c_1-\boldsymbol{\beta}^\top\mathbf{x}_i\right)\right]^{\mathbb{I}\{y_i=0\}}\left[F\left(c_2-\boldsymbol{\beta}^\top\mathbf{x}_i\right)-F\left(c_1-\boldsymbol{\beta}^\top\mathbf{x}_i\right)\right]^{\mathbb{I}\{y_i=1\}}\left[1-F\left(c_2-\boldsymbol{\beta}^\top\mathbf{x}_i\right)\right]^{\mathbb{I}\{y_i=2\}}\right\}\\
&=\sum_{i=1}^n \mathbb{I}\{y_i=0\}\log F\left(c_1-\boldsymbol{\beta}^\top\mathbf{x}_i\right)+\sum_{i=1}^n \mathbb{I}\{y_i=1\}\log\left[F\left(c_2-\boldsymbol{\beta}^\top\mathbf{x}_i\right)-F\left(c_1-\boldsymbol{\beta}^\top\mathbf{x}_i\right)\right]\\
&\quad+\sum_{i=1}^n \mathbb{I}\{y_i=2\}\log\left[1-F\left(c_2-\boldsymbol{\beta}^\top\mathbf{x}_i\right)\right].
\end{aligned}$$

Here $\mathbb{I}\{\cdot\}$ is the indicator function. Note that the log-likelihood function $\mathcal{L}(\boldsymbol{\beta}, c_1, c_2)$ is only defined for $c_1 < c_2$.

(c) Softmax is a standard way to generalize logistic regression to multiple categories. It has the following form:

$$\mathbb{P}(Y=k|\mathbf{X}) = \frac{e^{\beta_k^\top\mathbf{X}}}{\sum_{k=1}^K e^{\beta_k^\top\mathbf{X}}}.$$

It is easy to check that $\sum_{k=1}^K \mathbb{P}(Y=k|\mathbf{X}) = 1$ and all the probabilities are between 0 and 1.

2. (a) Direct calculation gives $\ell_n(\boldsymbol{\beta}) = \phi^{-1}\sum_{i=1}^n[b(\mathbf{X}_i^T\boldsymbol{\beta}) - Y_i\mathbf{X}_i^T\boldsymbol{\beta}] + C$, where C does not depend on $\boldsymbol{\beta}$. Hence $\nabla^2\ell_n(\boldsymbol{\beta}) = \phi^{-1}\sum_{i=1}^n b''(\mathbf{X}_i^T\boldsymbol{\beta})\mathbf{X}_i\mathbf{X}_i^T$ and $\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}) = [\nabla^2\ell_n(\boldsymbol{\beta})]^{-1} = \phi[\sum_{i=1}^n b''(\mathbf{X}_i^T\widehat{\boldsymbol{\beta}})\mathbf{X}_i\mathbf{X}_i^T]^{-1} = \phi[\sum_{i=1}^n b''(\widehat{\theta}_i)\mathbf{X}_i\mathbf{X}_i^T]^{-1}$.

(b) For logistic regression, we have $b(t) = \log(1+e^t)$, $b''(t) = \frac{e^t}{(1+e^t)^2}$, and $\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}) = \phi[\sum_{i=1}^n \frac{e^{\mathbf{X}_i^T\widehat{\boldsymbol{\beta}}}}{(1+e^{\mathbf{X}_i^T\widehat{\boldsymbol{\beta}}})^2}\mathbf{X}_i\mathbf{X}_i^T]^{-1}$. For Poisson regression, we have $b(t) = e^t$, $b''(t) = e^t$, and $\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}) = \phi[\sum_{i=1}^n e^{\mathbf{X}_i^T\widehat{\boldsymbol{\beta}}}\mathbf{X}_i\mathbf{X}_i^T]^{-1}$.

(c) The formulation is $\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\|\boldsymbol{\beta}\|_1$, s.t. $\|\nabla\ell_n(\boldsymbol{\beta})\|_\infty \le \gamma_n$, where $\gamma_n > 0$ is a tuning parameter. From $\nabla\ell_n(\boldsymbol{\beta}) = \phi^{-1}\sum_{i=1}^n \mathbf{X}_i[b'(\mathbf{X}_i^T\boldsymbol{\beta}) - Y_i]$ we can write the optimization problem more explicitly: $\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\|\boldsymbol{\beta}\|_1$, s.t. $\|\sum_{i=1}^n \mathbf{X}_i[b'(\mathbf{X}_i^T\boldsymbol{\beta}) - Y_i]\|_\infty \le \phi\gamma_n$.

(d) The fact that $b'(t) = \frac{e^t}{1+e^t}$ leads to the answer

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{\beta}\|_1, \text{ s.t. } \Big\|\sum_{i=1}^n \mathbf{X}_i[\frac{e^{\mathbf{X}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{X}_i^T\boldsymbol{\beta}}} - Y_i]\Big\|_\infty \le \phi\gamma_n.$$

3. (a) Direct calculation gives $\nabla\ell_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \mathbf{X}_i[b'(\mathbf{X}_i^T\boldsymbol{\beta}) - Y_i]$ and thus $\nabla\ell_n(\boldsymbol{\beta}^*) = n^{-1}\sum_{i=1}^n \varepsilon_i\mathbf{X}_i$.

(b) Let $R(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ (clearly decomposable) and $\overline{\mathcal{M}} = \mathcal{M} = \{\boldsymbol{\beta}\in\mathbb{R}^p : \boldsymbol{\beta}_{\mathcal{S}^c} = \mathbf{0}\}$. Then $R^*(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_\infty$ and $\overline{\mathcal{M}}^\perp = \mathcal{M}^\perp = \{\boldsymbol{\beta}\in\mathbb{R}^p : \boldsymbol{\beta}_{\mathcal{S}} = \mathbf{0}\}$. Since the assumption $\lambda_n \ge 2\|\nabla\ell_n(\boldsymbol{\beta}^*)\|_\infty$ translates to $R^*(\boldsymbol{\beta}^*) \le \lambda_n/2$, Proposition 5.3 implies that $R(\boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp}) \le 3R(\boldsymbol{\Delta}_{\overline{\mathcal{M}}}) + 4R(\boldsymbol{\beta}^*_{\mathcal{M}^\perp})$. Note that for any $\boldsymbol{\beta}\in\mathbb{R}^p$ we have $\boldsymbol{\beta}_{\mathcal{M}} = \boldsymbol{\beta}_{\overline{\mathcal{M}}} = \boldsymbol{\beta}_{\mathcal{S}}$ and $\boldsymbol{\beta}_{\mathcal{M}^\perp} = \boldsymbol{\beta}_{\overline{\mathcal{M}}^\perp} = \boldsymbol{\beta}_{\mathcal{S}^c}$. Hence $\|\boldsymbol{\Delta}_{\mathcal{S}^c}\|_1 \le 3\|\boldsymbol{\Delta}_{\mathcal{S}}\|_1 + 4\|\boldsymbol{\beta}^*_{\mathcal{S}^c}\|_1$.

(c) It is easily seen that $\psi(\overline{\mathcal{M}}) = \sqrt{|\mathcal{S}_n|} = \sqrt{s_n}$. Theorem 5.8 then forces

$$\|\boldsymbol{\Delta}\|_2^2 \le \frac{9\lambda_n^2}{4\kappa_L^2}s_n + \frac{4\lambda_n}{\kappa_L}\|\boldsymbol{\beta}^*_{\mathcal{S}^c}\|_1 \lesssim \lambda_n^2 s_n + \lambda_n\|\boldsymbol{\beta}^*_{\mathcal{S}^c}\|_1 \lesssim \lambda_n^2 s_n + \lambda_n^2 \lesssim \lambda_n^2 s_n.$$

This implies that $\|\boldsymbol{\Delta}\|_2 \lesssim \lambda_n\sqrt{s_n}$ and $\|\boldsymbol{\Delta}_{\mathcal{S}}\|_1 \le \sqrt{s_n}\|\boldsymbol{\Delta}_{\mathcal{S}}\|_2 \le \sqrt{s_n}\|\boldsymbol{\Delta}\|_2 \lesssim \lambda_n s_n$. It follows from Part (b) that $\|\boldsymbol{\Delta}_{\mathcal{S}^c}\|_1 \le 3\|\boldsymbol{\Delta}_{\mathcal{S}}\|_1 + 4\|\boldsymbol{\beta}^*_{\mathcal{S}^c}\|_1 \lesssim \lambda_n s_n$. Finally the proof is completed by the triangle equality $\|\boldsymbol{\Delta}\|_1 \le \|\boldsymbol{\Delta}_{\mathcal{S}}\|_1 + \|\boldsymbol{\Delta}_{\mathcal{S}^c}\|_1$.

4. (a) By definition of $G(\mathbf{x}|\mathbf{x}_0)$, we have

$$\begin{aligned} G(\mathbf{x}|\mathbf{x}_0) &= f(\mathbf{x}_0) + f'(\mathbf{x}_0)^\top(\mathbf{x}-\mathbf{x}_0) + \frac{1}{2\delta}\|\mathbf{x}-\mathbf{x}_0\|^2 + \lambda\|\mathbf{x}\|_1 \\ &\le f(\mathbf{x}) + \frac{1}{2\delta}\|\mathbf{x}-\mathbf{x}_0\|^2 + \lambda\|\mathbf{x}\|_1 = F(\mathbf{x}) + \frac{1}{2\delta}\|\mathbf{x}-\mathbf{x}_0\|^2, \end{aligned}$$

where the last inequality follows from the convexity, i.e. $f(\mathbf{x}_0) + f'(\mathbf{x}_0)^\top(\mathbf{x}-\mathbf{x}_0) \le f(\mathbf{x})$.

(b) Since $G(\mathbf{x}|\mathbf{x}_{i-1})$ is a majorization of $F(\mathbf{x})$, one has

$$F(\mathbf{x}_i) \le G(\mathbf{x}_i|\mathbf{x}_{i-1}) \le \min_{\mathbf{x}} G(\mathbf{x}|\mathbf{x}_{i-1}),$$

where the last inequality arises due to the definition of $\mathbf{x}_i$. It is straightforward to see that

$$\begin{aligned} F(\mathbf{x}_i) &\le \min_w G(w\mathbf{x}^* + (1-w)\mathbf{x}_{i-1}|\mathbf{x}_{i-1}) \\ &= \min_w \left\{F(w\mathbf{x}^* + (1-w)\mathbf{x}_{i-1}) + \frac{1}{2\delta}\|w\mathbf{x}^* + (1-w)\mathbf{x}_{i-1} - \mathbf{x}_{i-1}\|^2\right\} \\ &\le \min_w \left\{wF(\mathbf{x}^*) + (1-w)F(\mathbf{x}_{i-1}) + \frac{w^2}{2\delta}\|\mathbf{x}^* - \mathbf{x}_{i-1}\|^2\right\}. \end{aligned}$$

(c) First, by the optimality of $\mathbf{x}^*$, one knows there exists a subgradient $\mathbf{y}$ of $\|\cdot\|_1$ at $\mathbf{x}^*$ such that

$$f'(\mathbf{x}^*) + \lambda\mathbf{y} = \mathbf{0}.$$

In addition, we have

$$\begin{aligned}
F\left(\mathbf{x}_{i-1}\right)-F\left(\mathbf{x}^*\right) &= f\left(\mathbf{x}_{i-1}\right)-f\left(\mathbf{x}^*\right)+\lambda\left\|\mathbf{x}_{i-1}\right\|_1-\lambda\left\|\mathbf{x}^*\right\|_1 \\
&\overset{\text{(i)}}{\geq} f'\left(\mathbf{x}^*\right)^\top\left(\mathbf{x}_{i-1}-\mathbf{x}^*\right)+\frac{\sigma}{2}\left\|\mathbf{x}^*-\mathbf{x}_{i-1}\right\|^2+\lambda\left\|\mathbf{x}_{i-1}\right\|_1-\lambda\left\|\mathbf{x}^*\right\|_1 \\
&\overset{\text{(ii)}}{\geq} f'\left(\mathbf{x}^*\right)^\top\left(\mathbf{x}_{i-1}-\mathbf{x}^*\right)+\frac{\sigma}{2}\left\|\mathbf{x}^*-\mathbf{x}_{i-1}\right\|^2+\lambda\left\langle\mathbf{y},\mathbf{x}_{i-1}-\mathbf{x}^*\right\rangle \\
&\overset{\text{(iii)}}{=} \frac{\sigma}{2}\left\|\mathbf{x}^*-\mathbf{x}_{i-1}\right\|^2.
\end{aligned}$$

Here (i) uses the fact that $f(\cdot)$ is a strongly convex function, (ii) results from the convexity of $\|\cdot\|_1$ and the fact that $\mathbf{y}$ is a subgradient of $\|\cdot\|_1$ at $\mathbf{x}^*$, and (iii) follows from the identity $f'(\mathbf{x}^*)+\lambda\mathbf{y}=\mathbf{0}$ we proved above. This finishes the proof.

(d) Combining the results in (b) and (c), we have

$$\begin{aligned}
F\left(\mathbf{x}_i\right)-F\left(\mathbf{x}^*\right) &\leq \min_w\left\{(1-w)\left[F\left(\mathbf{x}_{i-1}\right)-F\left(\mathbf{x}^*\right)\right]+\frac{w^2}{\delta\sigma}\left[F\left(\mathbf{x}_{i-1}\right)-F\left(\mathbf{x}^*\right)\right]\right\} \\
&= \left(1-\frac{\delta\sigma}{4}\right)\left[F\left(\mathbf{x}_{i-1}\right)-F\left(\mathbf{x}^*\right)\right].
\end{aligned}$$

Here the last line results from the choice $w=\frac{1}{2}\delta\sigma$ which minimizes the right hand side of the first line.

5. Cf code

6. Cf code