DS-GA 3001.009 Applied Statistics: Homework #3 Solutions

Due on Thursday, October 5, 2023

Please hand in your homework via Gradescope (entry code: RKXJN2) before 11:59 PM.

1. Let $p_\theta(y) = \exp(\theta \cdot T(y) - A(\theta))h(y)$ be a general one-dimensional exponential family, and consider a GLM $y_i \sim p_{\theta_i}$ with $\theta_i = \langle \beta, x_i \rangle$ for all $1 \le i \le n$.

   (a) Let $\widehat{\beta}$ be the MLE of β based on $(x_1, y_1), \cdots, (x_n, y_n)$. Prove Hoeffding's formula:

   $$D_+(\widehat{\beta}; \beta) = 2 \log \frac{p_{\widehat{\beta}}(y_1, \cdots, y_n \mid x_1, \cdots, x_n)}{p_\beta(y_1, \cdots, y_n \mid x_1, \cdots, x_n)}$$

   holds for any β, where $D_+(\cdot; \cdot)$ denotes the total deviance.

   (b) Let $\widehat{\beta}_0$ be the MLE restricted to a subset $\beta \in \Theta_0 \subseteq \mathbb{R}^p$; assume that Hoeffding's formula also holds for $\widehat{\beta}_0$ whenever $\beta \in \Omega_0$. Prove the deviance additivity theorem:

   $$D_+(\widehat{\beta}; \widehat{\beta}_0) = D_+(\widehat{\beta}; \beta) - D_+(\widehat{\beta}_0; \beta), \quad \forall \beta \in \Omega_0.$$

Solution:

(a) It holds that

$$\begin{aligned}
&2 \log \frac{p_{\widehat{\beta}}(y_1, \cdots, y_n \mid x_1, \cdots, x_n)}{p_\beta(y_1, \cdots, y_n \mid x_1, \cdots, x_n)} \\
&= 2 \sum_{i=1}^n \log \frac{p_{\widehat{\beta}}(y_i \mid x_i)}{p_\beta(y_i \mid x_i)} \\
&= 2 \sum_{i=1}^n \log \frac{\exp(\langle \widehat{\beta}, x_i \rangle T(y_i) - A(\langle \widehat{\beta}, x_i \rangle))h(y_i)}{\exp(\langle \beta, x_i \rangle T(y_i) - A(\langle \beta, x_i \rangle))h(y_i)} \\
&= 2 \sum_{i=1}^n \Big(A(\langle \beta, x_i \rangle) - A(\langle \widehat{\beta}, x_i \rangle) \Big) - 2 \left\langle \beta - \widehat{\beta}, \sum_{i=1}^n x_i T(y_i) \right\rangle \\
&\overset{(1)}{=} 2 \sum_{i=1}^n \Big(A(\langle \beta, x_i \rangle) - A(\langle \widehat{\beta}, x_i \rangle) \Big) - 2 \left\langle \beta - \widehat{\beta}, \sum_{i=1}^n x_i A'(\langle \widehat{\beta}, x_i \rangle) \right\rangle \\
&= 2 \sum_{i=1}^n \Big(A(\langle \beta, x_i \rangle) - A(\langle \widehat{\beta}, x_i \rangle) - A'(\langle \widehat{\beta}, x_i \rangle)\langle \beta - \widehat{\beta}, x_i \rangle \Big) = D_+(\widehat{\beta}; \beta),
\end{aligned}$$

where (1) is due to the estimating equation for GLM.

(b) By Hoeffding's formula for $\widehat{\beta}$ and $\widehat{\beta}_0$, we have

$$\begin{aligned}
&D_+(\widehat{\beta}; \beta) - D_+(\widehat{\beta}_0; \beta) \\
&= 2 \log \frac{p_{\widehat{\beta}}(y_1, \cdots, y_n \mid x_1, \cdots, x_n)}{p_\beta(y_1, \cdots, y_n \mid x_1, \cdots, x_n)} - 2 \log \frac{p_{\widehat{\beta}_0}(y_1, \cdots, y_n \mid x_1, \cdots, x_n)}{p_\beta(y_1, \cdots, y_n \mid x_1, \cdots, x_n)}
\end{aligned}$$

$$= 2\log \frac{p_{\widehat{\beta}}(y_1, \cdots, y_n \mid x_1, \cdots, x_n)}{p_{\widehat{\beta}_0}(y_1, \cdots, y_n \mid x_1, \cdots, x_n)}$$
$$= D_+(\widehat{\beta}; \widehat{\beta}_0),$$

where the last step follows from Hoeffding's formula for $\widehat{\beta}$.

2. Let $\pi = (\pi_1, \cdots, \pi_k)$ be a probability vector, i.e. $\pi_j \geq 0$ for all $j = 1, \cdots, k$, $\sum_{j=1}^k \pi_j = 1$. Let p_π denote the statistical model $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \pi$ with sample size n.
   (a) Write out the log-likelihood $\ell_\pi(Y_1, \cdots, Y_n) = \log p_\pi(Y_1, \cdots, Y_n)$.
   (b) Let $(\pi_1, \cdots, \pi_{k-1})$ be the free parameters, and $\pi_k = 1 - \sum_{j=1}^{k-1} \pi_j$. Show that the score function $s_\pi = (s_{\pi,1}, \cdots, s_{\pi,k-1})$ is given by
   $$s_{\pi,j}(Y_1, \cdots, Y_n) = \sum_{i=1}^n \left(\frac{\mathbb{1}(Y_i = j)}{\pi_j} - \frac{\mathbb{1}(Y_i = k)}{\pi_k} \right).$$
   (c) Verify that the Fisher information matrix $I(\pi)$ is given by
   $$I(\pi) = n\left(\operatorname{diag}(\pi_1^{-1}, \cdots, \pi_{k-1}^{-1}) + \frac{\mathbf{1}\mathbf{1}^\top}{\pi_k} \right),$$
   where $\mathbf{1} \in \mathbb{R}^{k-1}$ is the column vector consisting of all ones.
   (d) Using the Woodbury matrix identity (consult wikipedia), compute $I(\pi)^{-1}$. Compare your result with your answer to 2(a) in HW2. What do you find?

Solution:

(a) We have
$$\ell_\pi(Y_1, \cdots, Y_n) = \sum_{i=1}^n \log p_\pi(Y_i) = \sum_{i=1}^n \sum_{j=1}^k \mathbb{1}(Y_i = j) \log \pi_j.$$

(b) Using $\pi_k = 1 - \sum_{j=1}^{k-1} \pi_j$, for $j = 1, \cdots, k-1$ we have
$$s_{\pi,j}(Y_1, \cdots, Y_n) = \frac{\partial \ell_\pi(Y_1, \cdots, Y_n)}{\partial \pi_j} = \sum_{i=1}^n \left(\frac{\mathbb{1}(Y_i = j)}{\pi_j} - \frac{\mathbb{1}(Y_i = k)}{\pi_k} \right).$$

(c) Based on the expression of s_π, it is clear that
$$\frac{\partial s_{\pi,j}(Y_1, \cdots, Y_n)}{\partial \pi_\ell} = \sum_{i=1}^n \left(-\frac{\mathbb{1}(Y_i = j) \cdot \mathbb{1}(j = \ell)}{\pi_j^2} - \frac{\mathbb{1}(Y_i = k)}{\pi_k^2} \right).$$
Therefore,
$$I(\pi)_{j,\ell} = \mathbb{E}\left[-\frac{\partial s_{\pi,j}(Y_1, \cdots, Y_n)}{\partial \pi_\ell} \right] = n\left(\frac{\mathbb{1}(j = \ell)}{\pi_j} + \frac{1}{\pi_k} \right).$$
This coincides with the claimed matrix form.

(d) By Woodbury matrix identity, it holds that

$$\begin{aligned}
I(\pi)^{-1} &= \frac{1}{n}\left(\operatorname{diag}(\pi_1^{-1},\cdots,\pi_{k-1}^{-1}) + \frac{\mathbf{1}\mathbf{1}^\top}{\pi_k}\right)^{-1} \\
&= \frac{1}{n}\left(\operatorname{diag}(\pi_1,\cdots,\pi_{k-1}) - \pi_{\sim k}\left(\pi_k + \sum_{j=1}^{k-1}\pi_j\right)\pi_{\sim k}^\top\right) \\
&= \frac{\operatorname{diag}(\pi_{\sim k}) - \pi_{\sim k}\pi_{\sim k}^\top}{n},
\end{aligned}$$

where $\pi_{\sim k}$ denotes the column vector $(\pi_1,\cdots,\pi_{k-1})$. This result is identical to the covariance matrix of the MLE in 2(a) of HW2, showing that the MLE is efficient in this example (attaining the Cramér-Rao lower bound).

3. Coding: we will implement Lindsey's method for density estimation. Given $z_1,\cdots,z_{200} \sim p_Z$ (in the experiment we set $p_Z = \mathcal{N}(0.5, 1)$), we aim to fit p_Z using

$$p_\theta(z) \propto \exp\left(\sum_{j=1}^{5}\theta_j z^j\right) h(z)$$

with $h(z) = \exp(-z^2/2)$. In other words, the fitted exponent is a degree-5 polynomial of z. In this problem, we will:

(a) use Lindsey's method to fit a full model $\theta \in \mathbb{R}^5$;

(b) use model selection techniques (AIC and Lasso) to fit a reduced model.

Based on the inline instructions, fill in the missing codes in `https://tinyurl.com/mr34wr63`. Be sure to submit a pdf with your codes, outputs, and colab link.

Solution: see `https://tinyurl.com/r2vntd38`.

4. (Bonus question, 5 pts) In this problem we show that the map

$$(x,y) \mapsto g(x,y) = \log\left(\frac{1}{1+e^{-x}} - \frac{1}{1+e^{-y}}\right), \quad x,y \in \mathbb{R}, x \geq y$$

is concave, which implies the concavity of the MLE objective in the ordered logit model. To this end we use the following Prékopa-Leindler inequality.

Theorem 1 (Prékopa-Leindler). *If $(u,v) \mapsto f(u,v) \in [0,\infty)$ is log-concave for $u \in \mathbb{R}^m, v \in \mathbb{R}^n$, the partial integration $u \mapsto \int_{\mathbb{R}^n} f(u,v)dv$ is also log-concave.*

(a) For $x \geq y, t \in \mathbb{R}$, show that

$$f(x,y,t) = \frac{e^t}{(1+e^t)^2}\mathbb{1}(y \leq t \leq x)$$

is log-concave in (x,y,t).

(b) Use Prékopa-Leindler to conclude that $g(x, y)$ is log-concave in (x, y).

(c) Use the above program to prove that $(x, y) \mapsto \log(\Phi(x) - \Phi(y))$ is jointly concave in $(x, y) \in \mathbb{R}^2$ with $x \geq y$, where Φ is the CDF of the standard normal distribution. Choosing $y \to -\infty$, this gives an alternative proof that $x \mapsto \log \Phi(x)$ is concave.

Solution:

(a) First we show that the map $t \mapsto h(t) = e^t/(1+e^t)^2$ is log-concave. This directly follows from the concavity of

$$t \mapsto \log h(t) = t - 2\log(1+e^t).$$

For the original function f, assume that $(x, y, t) = \lambda(x_1, y_1, t_1) + (1-\lambda)(x_2, y_2, t_2)$. Without loss of generality we may assume that $y_1 \leq t_1 \leq x_1$ and $y_2 \leq t_2 \leq x_2$; otherwise one of $f(x_1, y_1, t_1)$ and $f(x_2, y_2, t_2)$ is zero, and

$$f(x, y, t) \geq f(x_1, y_1, t_1)^\lambda f(x_2, y_2, t_2)^{1-\lambda}$$

clearly holds. Under the above assumptions, we have $y \leq t \leq x$ as well, and

$$\begin{aligned}
&\log f(x, y, t) - \lambda \log f(x_1, y_1, t_1) - (1-\lambda) f(x_2, y_2, t_2) \\
&= \log h(t) - \lambda \log h(t_1) - (1-\lambda)\log h(t_2) \geq 0,
\end{aligned}$$

where the last step follows from the log-concavity of h. This completes the proof.

(b) Since

$$g(x, y) = \int_{\mathbb{R}} f(x, y, t)dt,$$

by Prékopa-Leindler we conclude that $g(x, y)$ is log-concave in (x, y).

(c) Note that the same analysis above goes through if $x \mapsto \Phi'(x)$ is log-concave. This is straightforward as

$$\log \Phi'(x) = -\frac{x^2 + \log(2\pi)}{2}$$

is clearly concave.