DS-GA 3001.009 Applied Statistics: Homework #1

Due on Thursday, September 21, 2023

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

1. The Gamma distribution has a shape parameter $\alpha > 0$ and a scale parameter $\beta > 0$, with density given by

$$\Gamma_{\alpha,\beta}(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, \quad y > 0.$$

Here $\Gamma(\alpha)$ is the Gamma function - you only need to know that this is a function of α and will not need any further properties.

(a) Show that the family of Gamma distributions $\{\Gamma_{\alpha,\beta}(y)\}_{\alpha,\beta>0}$ belongs to the exponential family. Write down the expressions of $(\theta, T(y), A(\theta), h(y))$.

(b) Verify that the Gamma distribution is a conjugate prior for the Poisson family, i.e. if $\lambda \sim \Gamma_{\alpha,\beta}$ and $y \sim \mathrm{Poi}(\lambda)$, then $\lambda \mid y \sim \Gamma_{\alpha(y),\beta(y)}$.

2. Let $\{p_\theta(y)\}_{\theta\in\Theta}$ be an exponential family taking the standard form

$$p_\theta(y) = \exp(\langle \theta, T(y)\rangle - A(\theta))h(y), \quad y \in \mathcal{Y}.$$

Show that for $\mathcal{Y}_0 \subseteq \mathcal{Y}$, the conditional family $\{p_\theta(y \mid y \in \mathcal{Y}_0)\}_{\theta\in\Theta}$ is also an exponential family taking the form

$$p_\theta(y \mid y \in \mathcal{Y}_0) = \exp(\langle \theta, \widetilde{T}(y)\rangle - \widetilde{A}(\theta))\widetilde{h}(y), \quad y \in \mathcal{Y}.$$

Write down the expressions of $(\widetilde{T}(y), \widetilde{A}(\theta), \widetilde{h}(y))$.

3. Recall from the lecture that for an exponential family $p_\theta(y) = \exp(\langle \theta, T(y)\rangle - A(\theta))h(y)$, the family of conjugate priors has two parameters $\xi \in \mathbb{R}^d$ and $\tau > 0$, with density

$$\pi_{\xi,\tau}(\theta) = \exp(\langle \xi, \theta\rangle - \tau A(\theta))b(\xi, \tau).$$

(a) Using $\mathbb{E}_{\xi,\tau}[\nabla_\theta \log \pi_{\xi,\tau}(\theta)] = 0$ (you don't need to prove this), show that

$$\mathbb{E}_{\xi,\tau}[\nabla A(\theta)] = \frac{\xi}{\tau}.$$

(b) Given i.i.d. observations $y_1, \cdots, y_n \sim p_\theta(y)$, show that the posterior distribution takes the form

$$\pi_{\xi,\tau}(\theta \mid y_1, \cdots, y_n) = \pi_{\xi+\sum_{i=1}^n T(y_i), \tau+n}(\theta).$$

(c) Show that the posterior mean of $\mu_\theta = \nabla A(\theta)$ is

$$\mathbb{E}_{\xi,\tau}[\nabla A(\theta) \mid y_1, \cdots, y_n] = \frac{\tau}{\tau+n} \cdot \mathbb{E}_{\xi,\tau}[\nabla A(\theta)] + \frac{n}{\tau+n} \cdot \frac{1}{n}\sum_{i=1}^n T(y_i).$$

How would you interpret this result?

4. Coding: based on the instructions, complete the missing codes in the colab link. In your submission, be sure to submit a pdf with your codes, outputs, and colab link.

    - Colab link (you should make a copy before edits): `https://tinyurl.com/4z6eh9k4`
    - Dataset “College.csv” link: `https://tinyurl.com/yckewn3u`
    - For the meanings of the variables, consult Chapter 2, Exercise 8 of the ISLR book (`https://tinyurl.com/ykczmds5`)