Math of Big Data, Summer 2018

Prof: Gu

Name: **Forest Kobayashi**

HW #: **3**

Day: **Mon.** **Tue.** **Wed.** **Thu.** **Fri.**

Date: **05/17/2018**

No.	**Points**	**Acknowledgments**
1		Kevin Cotton, Tim Player
2		Solutions
Total		

This Assignment is (check one):

[X] On Time [ ] Late, without deduction [ ] Late, with deduction

Comments: Feel free to work with other students, but make sure you write up the homework and code on your own (no copying homework *or* code; no pair programming). Feel free to ask students or instructors for help debugging code or whatever else, though.

Problem 1. (Murphy 2.16)

Suppose $\theta \sim \text{Beta}(a, b)$ such that

$$\mathbb{P}(\theta; a, b) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the Beta function and $\Gamma(x)$ is the Gamma function. Derive the mean, mode, and variance of θ.

Solution:

(a) We have

$$\begin{aligned}\mu &= \mathbb{E}[\theta \mid a, b] \\ &= \frac{1}{B(a,b)}\int_0^1 \theta \cdot \theta^{a-1}(1-\theta)^{b-1}\,\mathrm{d}\theta \\ &= \frac{1}{B(a,b)}\int_0^1 \theta^{a}(1-\theta)^{b-1}\,\mathrm{d}\theta\end{aligned}$$

note that the inside of the integral is just the probability density function for a beta distribution parameterized by $a+1$, b. Hence, the integral is just $B(a+1,b)$, and so

$$\begin{aligned}&= \frac{B(a+1,b)}{B(a,b)} \\ &= \frac{\Gamma(a+1)\Gamma(b)\Gamma(a+b)}{\Gamma(a+1+b)\Gamma(a)\Gamma(b)}\end{aligned}$$

By definition, $\Gamma(s+1) = s\Gamma(s)$, hence

$$\begin{aligned}&= \frac{a\Gamma(a)\Gamma(a+b)}{\Gamma(a)(a+b)\Gamma(a+b)} \\ &= \boxed{\frac{a}{a+b}}\end{aligned}$$

(b) The mode will correspond to the maximum in the probability density function. We apply the first derivative test:

$$\begin{aligned}\frac{\mathrm{d}\mathbb{P}(\theta; a,b)}{\mathrm{d}\theta} &= \frac{1}{B(a,b)}\Big((a-1)\theta^{a-2}(1-\theta)^{b-1} - (b-1)\theta^{a-1}(1-\theta)^{b-2}\Big) \\ &= 0\end{aligned}$$

and so

$$(a-1)\theta^{a-2}(1-\theta)^{b-1} = (b-1)\theta^{a-1}(1-\theta)^{b-2}$$

note that we have trivial solutions $\theta = 0, 1$. Supposing $\theta \neq 0, 1$,

$$\begin{aligned}(a-1)(1-\theta) &= (b-1)\theta \\ a - 1 &= (b-1+a-1)\theta \\ \frac{a-1}{a+b-2} &= \theta\end{aligned}$$

(c) The variance is given by

$$\text{var}(\theta) = \mathbb{E}\Big[(\theta-\mu)^2\Big]$$

$$
\begin{aligned}
&= \frac{1}{B(a,b)} \int_0^1 \left(\theta - \frac{a}{a+b}\right)^2 \theta^{a-1} (1-\theta)^{b-1} \, \mathrm{d}\theta \\
&= \frac{1}{B(a,b)} \int_0^1 \theta^{a+1} (1-\theta)^{b-1} - \frac{2a}{a+b} \theta^{a} (1-\theta)^{b-1} + \frac{a^2}{(a+b)^2} \theta^{a-1} (1-\theta)^{b-1} \, \mathrm{d}\theta \\
&= \frac{1}{B(a,b)} \left(B(a+2,b) - \frac{2a}{a+b} B(a+1,b) + \frac{a^2}{(a+b)^2} B(a,b) \right) \\
&= \frac{B(a+2,b)}{B(a,b)} - \frac{2a}{a+b} \cdot \frac{a}{a+b} + \frac{a^2}{(a+b)^2} \\
&= \frac{a(a+1)}{(a+b+1)(a+b)} - \frac{a^2}{(a+b)^2} \\
&= \frac{\left(a^2+a\right)(a+b) - a^2(a+b+1)}{(a+b+1)(a+b)^2} \\
&= \frac{\cancel{a^3} + \cancel{a^2 b} + \cancel{a^2} + ab - \cancel{a^3} - \cancel{a^2 b} - \cancel{a^2}}{(a+b+1)(a+b)^2} \\
&= \boxed{\frac{ab}{(a+b+1)(a+b)^2}}
\end{aligned}
$$

Problem 2. (Murphy 9)

Show that the multinomial distribution

$$\text{Cat}(\mathbf{x} \mid \boldsymbol{\mu}) = \prod_{i=1}^{K} \mu_i^{x_i}$$

is in the exponential family and show that the generalized linear model corresponding to this distribution is the same as multinomial logistic regression (softmax regression).

Solution:

We apply $\exp \log$:

$$\begin{aligned} \exp\left(\log\left(\text{Cat}(\mathbf{x} \mid \boldsymbol{\mu})\right)\right) &= \exp\left(\log\left(\prod_{i=1}^{K} \mu_i^{x_i}\right)\right) \\ &= \exp\left(\sum_{i=1}^{K} x_i \log\left(\mu_i\right)\right) \end{aligned} \tag{2}$$

note that

$$\begin{aligned} \sum_{i=1}^{K} x_i &= 1 \\ x_K &= 1 - \sum_{i=1}^{K-1} x_i \end{aligned}$$

and

$$\begin{aligned} \sum_{i=1}^{K} \mu_i &= 1 \\ \mu_K &= 1 - \sum_{i=1}^{K-1} \mu_i x \end{aligned}$$

hence we can express (2) by

$$\begin{aligned} \text{Cat}(\mathbf{x} \mid \boldsymbol{\mu}) &= \exp\left(\sum_{i=1}^{K-1} x_i \log\left(\mu_i\right) + \left(1 - \sum_{i=1}^{K-1} x_i\right) \log\left(1 - \sum_{i=1}^{K-1} \mu_i\right)\right) \\ &= \exp\left(\sum_{i=1}^{K-1} x_i \left[\log\left(\mu_i\right) - \log\left(1 - \sum_{i=1}^{K} \mu_i\right)\right] + \log\left(1 - \sum_{i=1}^{K-1} \mu_i\right)\right) \\ &= \exp\left(\sum_{i=1}^{K-1} \left[x_i \log\left(\frac{\mu_i}{1 - \sum_{i=1}^{K} \mu_i}\right)\right] + \log\left(\mu_K\right)\right) \\ &= \exp\left(\sum_{i=1}^{K-1} \left[x_i \log\left(\frac{\mu_i}{\mu_K}\right)\right] + \log\left(\mu_K\right)\right) \end{aligned}$$

hence, if

$$\boldsymbol{\eta} = \begin{bmatrix} \log\left(\frac{\mu_1}{\mu_K}\right) \\ \log\left(\frac{\mu_2}{\mu_K}\right) \\ \vdots \\ \log\left(\frac{\mu_{K-1}}{\mu_K}\right) \end{bmatrix}$$

then integers $\forall 0 \leq i \leq K-1$

$$\begin{aligned} \eta_i &= \log\left(\frac{\mu_i}{\mu_K}\right) \\ e^{\eta_i}\mu_K &= \mu_i \end{aligned}$$

and so

$$\begin{aligned} \mu_K &= 1 - \mu_K \sum_{i=1}^{K-1} e^{\eta_i} \\ \mu_K\left(1 + \sum_{i=1}^{K-1} e^{\eta_i}\right) &= 1 \\ \mu_K &= \frac{1}{1 + \sum_{i=1}^{K-1} e^{\eta_i}} \end{aligned}$$

hence

$$\mu_i = \frac{e^{\eta_i}}{1 + \sum_{i=1}^{K-1} e^{\eta_i}}$$

Thus, letting

$$T(\mathbf{x}) = \begin{bmatrix} \mathbb{I}\{x_1 = 1\} \\ \mathbb{I}\{x_2 = 1\} \\ \vdots \\ \mathbb{I}\{x_{K-1} = 1\} \end{bmatrix}$$

and $h(\mathbf{x}) = 1$, and

$$\begin{aligned} A(\boldsymbol{\eta}) &= -\log(\mu_K) \\ &= \log\left(\frac{1}{\mu_K}\right) \\ &= 1 + \sum_{i=1}^{K-1} e^{\eta_i} \end{aligned}$$

we see

$$\mathrm{Cat}(\mathbf{x} \mid \boldsymbol{\mu}) = h(\mathbf{x})\big(\boldsymbol{\eta}^\top T(\mathbf{x}) - A(\boldsymbol{\eta})\big)$$

where $\boldsymbol{\eta}$ is the softmax of $\boldsymbol{\mu}$.