Math189R SU17
Homework 3
Wednesday, May 24, 2017

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.

**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 $\theta$.

According to the definition of Beta Function and Gamma function:

$$B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}\,d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \quad \Gamma(x+1) = x\Gamma(x).$$

Then, let's compute the mean of $\theta$.

$$E(\theta) = \int_0^1 \theta P(\theta; a,b)\,d\theta = \int_0^1 \theta\left(\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\right)d\theta$$

$$= \frac{1}{B(a,b)}\int_0^1 \theta^a(1-\theta)^{b-1}\,d\theta = \frac{B(a+1,b)}{B(a,b)}$$

$$= \left[\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right] = \left[\frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right]$$

$$= \frac{a}{a+b}.$$

Then, by property of variance, $Var[\theta] = E[(\theta - E[\theta])^2] = E[\theta^2] - E[\theta]^2$.

Thus, let's compute $E[\theta^2]$.

$$E[\theta^2] = \int_0^1 \theta^2\left(\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\right)d\theta$$

$$= \frac{1}{B(a,b)}\int_0^1 \theta^{a+1}(1-\theta)^{b-1}\,d\theta = \frac{B(a+2,b)}{B(a,b)}$$

$$= \left[\frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right] = \left[\frac{a(a+1)\Gamma(a)\Gamma(b)}{(a+b)(a+b+1)\Gamma(a+b)}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right]$$

$$= \frac{a(a+1)}{(a+b)(a+b+1)}$$

Then, $Var[\theta] = E[\theta^2] - E[\theta]^2$

$$= \frac{a(a+1)}{(a+b)(a+b+1)} - \frac{a^2}{(a+b)^2} = \frac{a(a+1)(a+b) - a^2(a+b+1)}{(a+b)^2(a+b+1)} = \frac{a^3 + a^2b + a^2 + ab - a^3 - a^2b - a^2}{(a+b)^2(a+b+1)} = \frac{ab}{(a+b)^2(a+b+1)}.$$

At last, lets compute the mode.

We need to find when $\nabla_\theta P(\theta; a, b) = 0$. As $\frac{1}{B(a,b)}$ is constant and won't affect the optimizing value, let's ignore the term.

Thus, $\nabla_\theta P(\theta; a, b) = \nabla_\theta \left[\theta^{a-1}(1-\theta)^{b-1}\right] = 0 = (a-1)\theta^{a-2}(1-\theta)^{b-1} - (b-1)\theta^{a-1}(1-\theta)^{b-2} = 0$.

So $(a-1)\theta^{a-2}(1-\theta)^{b-1} = (b-1)\theta^{a-1}(1-\theta)^{b-2}$, $(a-1)(1-\theta) = (b-1)\theta$, $(a+b-2)\theta = a-1$, $\theta^* = \frac{a-1}{a+b-2}$.

**2** (**Murphy 9**) Show that the multinomial distribution

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

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).

The exponential family has the form $P(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$.

Thus, we just need to rewrite the distribution with exponential and logarithm.

$$\text{Cat}(x|\mu) = \prod_{i=1}^{K} \mu_i^{x_i} = \exp\left[\log\left(\prod_{i=1}^{K} \mu_i^{x_i}\right)\right] = \exp\left(\sum_{i=1}^{K} \log(\mu_i^{x_i})\right) = \exp\left(\sum_{i=1}^{K} x_i \log(\mu_i)\right).$$

As $\sum_{i=1}^{K} \mu_i = 1$ and $\sum_{i=1}^{K} x_i = 1$, we just need to specify the first $K-1$ terms, since $\mu_K = 1 - \sum_{i=1}^{K-1} \mu_i$, $x_K = 1 - \sum_{i=1}^{K-1} x_i$.

Therefore, $\text{Cat}(x|\mu) = \exp\left(\sum_{i=1}^{K} x_i \log(\mu_i)\right) = \exp\left(\sum_{i=1}^{K-1} x_i \log(\mu_i) + x_K \log(\mu_K)\right)$

$$= \exp\left[\sum_{i=1}^{K-1} x_i \log(\mu_i) + \left(1 - \sum_{i=1}^{K-1} x_i\right)\log(\mu_K)\right] = \exp\left[\sum_{i=1}^{K-1} x_i(\log(\mu_i) - \log(\mu_K)) + \log(\mu_K)\right]$$

$$= \exp\left[\sum_{i=1}^{K-1} x_i \log\left(\frac{\mu_i}{\mu_K}\right) + \log(\mu_K)\right].$$

Thus, let $\eta = \begin{bmatrix} \log\left(\frac{\mu_1}{\mu_K}\right) \\ \vdots \\ \log\left(\frac{\mu_{K-1}}{\mu_K}\right) \end{bmatrix}$.

Then, $\mu_i = \mu_K e^{\eta_i}$, so $\mu_K = 1 - \sum_{i=1}^{K-1} \mu_i = 1 - \sum_{i=1}^{K-1} \mu_K e^{\eta_i} = 1 - \mu_K \sum_{i=1}^{K-1} e^{\eta_i} = \dfrac{1}{1 + \sum_{i=1}^{K-1} e^{\eta_i}}$.

Thus, $\mu_i = \mu_K e^{\eta_i} = \dfrac{e^{\eta_i}}{1 + \sum_{i=1}^{K-1} e^{\eta_i}}$

To write $\text{Cat}(x|\mu) = \exp(\eta^T x - a(\eta))$, we have $b(\eta) = 1$, $T(x) = x$, $a(\eta) = -\log(\mu_K) = \log\left(1 + \sum_{i=1}^{K-1} e^{\eta_i}\right)$,

so $\text{Cat}(x|\mu)$ is in the exponential family.

Also, $\mu = S(\eta)$, where $S(\eta)$ is the softmax function, which implies the generalized linear model of this distribution is the same as softmax regression.