Marcelo Almora Rios
Math189R SU2020
Homework 3
June 18, 2020

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 θ.

Start by computing the mean of θ with the given probability distribution $\mathbb{P}(\theta; a, b)$.

$$\mathbb{E}[\theta] = \int_0^1 \theta \mathbb{P}(\theta; a, b)\, d\theta \tag{1}$$

$$= \int_0^1 \theta \left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\right] d\theta \tag{2}$$

$$\tag{3}$$

Recall that

$$\beta(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1} d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{4}$$

Hence, equation (2) becomes

$$\mathbb{E}[\theta] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^{a}(1-\theta)^{b-1} d\theta = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+1+b)}. \tag{5}$$

Finally, since

$$\Gamma(a+b+1) = (a+b)\Gamma(a+b), \tag{6}$$

equation (5) becomes

$$\mathbb{E}[\theta] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+1+b)} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{a\,\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)} = \frac{a}{a+b}. \tag{7}$$

Now we compute the mode of θ. This is equivalent to find the solution to the equation

$$\nabla_\theta \mathbb{P}(\theta; a, b) = 0. \tag{8}$$

In other words, finding the solution to

$$\nabla_\theta \left[\theta^{a-1}(1-\theta)^{b-1}\right] = 0. \tag{9}$$

This is a simple computation. The left hand side evaluates to

$$(a-1)\theta^{a-2}(1-\theta)^{b-1} - (b-1)\theta^{a-1}(1-\theta)^{b-2} = 0 \tag{10}$$

which makes equation (9) become

$$(a-1)\theta^{a-2}(1-\theta)^{b-1} = (b-1)\theta^{a-1}(1-\theta)^{b-2} \tag{11}$$

$$\implies (a+b-2)\theta = a-1 \tag{12}$$

$$\implies \theta = \frac{a-1}{a+b-2} \tag{13}$$

after cancelling a factor of θ^{a-2} and $(1-\theta)^{b-2}$ from both sides. This is the mode.

Next we go to the variance! Recall that the variance is defined as

$$\text{Var}[\theta] = \mathbb{E}[\theta^2] - \mathbb{E}[\theta]^2. \tag{14}$$

We automatically have our second term from the first part of this problem. Hence, we focus on computing the first term. But this is simply:

$$\mathbb{E}[\theta^2] = \int_0^1 \theta^2 \mathbb{P}(\theta; a, b)\, d\theta \tag{15}$$

$$= \int_0^1 \theta^2 \left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\right] d\theta \tag{16}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^{a+1}(1-\theta)^{b-1} d\theta \tag{17}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma\big((a+1)+1\big)\Gamma(b)}{\Gamma\big((a+b+1)+1\big)} \tag{18}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \cdot \frac{(a+1)a}{(a+b+1)(a+b)} \tag{19}$$

$$= \frac{(a+1)a}{(a+b+1)(a+b)}. \tag{20}$$

Therefore,

$$\begin{aligned}
\text{Var}[\theta] &= \mathbb{E}[\theta^2] - \mathbb{E}[\theta]^2 && (21)\\
&= \frac{(a+1)a}{(a+b+1)(a+b)} - \frac{a^2}{(a+b)^2} && (22)\\
&= \frac{(a+b)(a+1)a}{(a+b+1)(a+b)^2} - \frac{a^2(a+b+1)}{(a+b)^2(a+b+1)} && (23)\\
&= \frac{(a^2+a+ba+b)a - a^3 - a^2b - a^2}{(a+b)^2(a+b+1)} && (24)\\
&= \frac{a^3+a^2+ba^2+ba - a^3 - a^2b - a^2}{(a+b)^2(a+b+1)} && (25)\\
&= \frac{ba}{(a+b)^2(a+b+1)}, && (26)
\end{aligned}$$

gives us the variance of θ, as desired.

■

2 (**Murphy 9**) Show that the multinoulli 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 multinoulli logistic regression (softmax regression).

We know that a probability distribution is in the exponential family if it can be expressed as

$$\mathbb{P}(\mathbf{y};\boldsymbol{\eta}) = b(\mathbf{y}) \exp\left(\eta^T T(\mathbf{y}) - a(\boldsymbol{\eta})\right). \tag{27}$$

So, by taking the logarithm and exponentiating both sides (in that order), we have

$$\text{Cat}(\mathbf{x}|\boldsymbol{\mu}) = \exp\left[\log \prod_{i=1}^{K} \mu_i^{x_i}\right] \tag{28}$$

$$= \exp\left[\sum_{i=1}^{K} x_i \log \mu_i\right] \tag{29}$$

Here I took a hint from the solutions because I had no idea how to go forward. Apparently, by acknowledging that

$$\sum_{i=1}^{K} x_i = \sum_{i=1}^{K-1} x_i + x_K = 1 \tag{30}$$

we can make the following substitution from the third to fourth line:

$$\text{Cat}(\mathbf{x}|\boldsymbol{\mu}) = \exp\left[\log \prod_{i=1}^{K} \mu_i^{x_i}\right] \tag{31}$$

$$= \exp\left[\sum_{i=1}^{K} x_i \log \mu_i\right] \tag{32}$$

$$= \exp\left[\sum_{i=1}^{K-1} x_i \log \mu_i + x_K \log \mu_K\right] \tag{33}$$

$$= \exp\left[\sum_{i=1}^{K-1} x_i \log \mu_i + \left(1 - \sum_{i=1}^{K-1} x_i\right) \log \mu_K\right] \tag{34}$$

$$= \exp\left[\sum_{i=1}^{K-1} x_i \big(\log \mu_i - \log \mu_K\big) + \log \mu_K\right] \tag{35}$$

$$= \exp\left[\sum_{i=1}^{K-1} x_i \log\left(\frac{\mu_i}{\mu_K}\right) + \log \mu_K\right]. \tag{36}$$

Therefore, we can set our canonical parameter to be

$$\boldsymbol{\eta} = \begin{bmatrix} \log\left(\frac{\mu_1}{\mu_K}\right) \\ \vdots \\ \log\left(\frac{\mu_{K-1}}{\mu_K}\right) \end{bmatrix} \tag{37}$$

where we see that $\mu_i = \mu_K e^{\boldsymbol{\eta}_i}$. By substituting this into equation (30),

$$\mu_K = 1 - \mu_K \sum_{i=1}^{K-1} e^{\boldsymbol{\eta}_i} \tag{38}$$

$$\implies \mu_K = \frac{1}{1 + \sum_{i=1}^{K-1} e^{\boldsymbol{\eta}_i}} \tag{39}$$

Hence,

$$\mu_i = \frac{1}{1 + \sum_{i=1}^{K-1} e^{\boldsymbol{\eta}_i}} \cdot e^{\boldsymbol{\eta}_i}. \tag{40}$$

Which finally allows us to write:

$$\text{Cat}(\mathbf{x}|\boldsymbol{\mu}) = \exp(\boldsymbol{\eta}^T \mathbf{x} + \log \mu_K) \tag{41}$$

proving that $\text{Cat}(\mathbf{x}|\boldsymbol{\mu})$ is in the exponential family. Observe that the Softmax function is (Wikipedia)

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{42}$$

for $i = 1, \ldots, K$, which means that the generalized linear model for Cat is the same as the softmax regression. ∎