Thitaree Tanprasert (Mint)
Math189R SP19
Homework 3
Monday, Feb 18, 2019

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.

By definition, $\Gamma(x) = \int_0^\infty \theta^{x-1}e^{-\theta}d\theta$ and $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. Therefore, we have that, if $\theta \sim \text{Beta}(a,b)$, then $B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta$.

- **Mean** θ: this is the same as $\mathbb{E}[\theta]$, so we set the equation as follow:

$$\begin{aligned}
\mathbb{E}[\theta] &= \int_0^1 \theta\mathbb{P}(\theta; a, b)d\theta \\
&= \frac{1}{B(a,b)}\int_0^1 \theta(\theta^{a-1}(1-\theta)^{b-1})d\theta \\
&= \frac{B(a+1,b)}{B(a,b)} \\
&= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+1+b)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}
\end{aligned}$$

Calculate $\Gamma(x+1) = \int_0^\infty \theta^{(x+1)-1}e^{-\theta}d\theta = x\int_0^\infty \theta^{x-1}e^{-\theta}d\theta = x\Gamma(x)$. Therefore,

$$\begin{aligned}
\mathbb{E}[\theta] &= \frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\
&= \frac{a}{a+b}
\end{aligned}$$

- **Variance of** θ: We can use the formula $\text{Var}(\theta) = \mathbb{E}[\theta^2] - (\mathbb{E}[\theta])^2$. We already know $\mathbb{E}[\theta] = \frac{a}{a+b}$, so we know that $(\mathbb{E}[\theta])^2 = \frac{a^2}{(a+b)^2}$. So, we only need to calculate $\mathbb{E}[\theta^2]$ in the same way that we did for the mean:

$$\begin{aligned}
\mathbb{E}[\theta^2] &= \int_0^1 \theta^2 \mathbb{P}(\theta; a, b) d\theta \\
&= \frac{1}{B(a,b)} \int_0^1 \theta^2 (\theta^{a-1}(1-\theta)^{b-1}) d\theta \\
&= \frac{B(a+2,b)}{B(a,b)} \\
&= \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+2+b)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\
&= \frac{(a+1)a\Gamma(a)\Gamma(b)}{(a+b+1)(a+b)\Gamma(a+b)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\
&= \frac{a(a+1)}{(a+b)(a+b+1)}
\end{aligned}$$

Therefore, we get

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

- **Mode** θ: If we visualize the distribution of θ, the mode is the peak of the distribution, so it is the value of θ where the gradient $\nabla_\theta \mathbb{P}(\theta; a, b) = 0$. We call this θ^*.

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

Therefore,

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

■

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

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

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).

First, we will force the equation to be exponential:

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

Based on this, we know that

$$\begin{aligned}
\boldsymbol{\eta} &= \begin{bmatrix} \log(\mu_1/\mu_K) \\ \log(\mu_2/\mu_K) \\ . \\ . \\ \log(\mu_{K-1}/\mu_K) \end{bmatrix} \\
a(\boldsymbol{\eta}) &= -\log(\mu_K) \\
T(x_i) &= x_i \\
b(x) &= 1
\end{aligned}$$

Finally, we need to calculate each μ in terms of $\boldsymbol{\eta}$. Since $\boldsymbol{\eta}_i = \log \mu_i/\mu_K$, we see that

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

And since $\mu_i = \mu_K e^{\boldsymbol{\eta}_i}$, we get

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

All of these are the same parameters and variables as we get for softmax regression, which we derived in class.

■