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, $B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ and $\Gamma(x+1) = x\Gamma(x)$. Then, we can find the mean,

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

Similarly, we can also find $\mathbb{E}[\theta^2]$,

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

Thus, it follows that

$$
\begin{aligned}
\mathrm{Var}[\theta] &= \mathbb{E}[\theta^2] - \mathbb{E}[\theta]^2 \\
&= \frac{a(a+1)}{(a+b)(a+b+1)} - \left(\frac{a}{a+b} \right)^2 \\
&= \frac{a(a+1)(a+b) - a^2(a+b+1)}{(a+b)^2(a+b+1)} \\
&= \frac{a^3 + a^2b + a^2 + ab - a^3 - a^2b - a^2}{(a+b)^2(a+b+1)} \\
&= \frac{ab}{(a+b)^2(a+b+1)}.
\end{aligned}
$$

Lastly, for mode, we wish to find θ such that $\nabla_\theta P(\theta; a, b) = 0$ on the interval $[0, 1]$.

$$
\begin{aligned}
\nabla_\theta P(\theta; a, b) &= \nabla_\theta \left[\theta^{a-1} (1-\theta)^{b-1} \right] = 0 \\
&= (a-1)\theta^{a-2}(1-\theta)^{b-1} - (b-1)\theta^{a-1}(1-\theta)^{b-2} = 0
\end{aligned}
$$

Solving for θ, we find,

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

which is our mode, as desired. ∎

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 show that $\mathrm{Cat}(x|\mu)$ is in the exponential family by rewriting it in the exponential form.

$$\begin{aligned}
\mathrm{Cat}(x|\mu) &= \prod_{i=1}^{K} \mu_i^{x_i} \\
&= \exp\left[\log\left(\prod_{i=1}^{K} \mu_i^{x_i}\right)\right] \\
&= \exp\left(\sum_{i=1}^{K} \log(\mu_i^{x_i})\right) \\
&= \exp\left(\sum_{i=1}^{K} x_i \log(\mu_i)\right)
\end{aligned}$$

Note that since $\sum_{i=1}^{K} \mu_i = 1$ and $\sum_{i=1}^{K} x_i = 1$, we have $\mu_K = 1 - \sum_{i=1}^{K-1} \mu_i$ and $x_K = 1 - \sum_{i=1}^{K-1} x_i$.

Then, we can rewrite $\mathrm{Cat}(x|\mu)$ using the above information and the property of logarithm.

$$\begin{aligned}
\mathrm{Cat}(x|\mu) &= \exp\left(\sum_{i=1}^{K} x_i \log(\mu_i)\right) \\
&= \exp\left(\sum_{i=1}^{K-1} x_i \log(\mu_i) + x_K \log(\mu_K)\right) \\
&= \exp\left(\sum_{i=1}^{K-1} x_i \log(\mu_i) + \left(1 - \sum_{i=1}^{K-1} x_i\right) \log(\mu_K)\right) \\
&= \exp\left(\sum_{i=1}^{K-1} x_i (\log(\mu_i) - \log(\mu_K)) + \log(\mu_K)\right) \\
&= \exp\left(\sum_{i=1}^{K-1} x_i \log\left(\frac{\mu_i}{\mu_K}\right) + \log(\mu_K)\right)
\end{aligned}$$

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

Note that since $\mu_K = 1 - \sum_{i=1}^{K-1} \mu_i$, we can rewrite μ_K as the following.

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

Thus, we can deduce that $\mu_i = \mu_K e^{\eta_i} = \frac{e^{\eta_i}}{1+\sum_{j=1}^{K-1} e^{\eta_j}}$.

Finally, we can rewrite the distribution in the form of exponential family as $\mathrm{Cat}(x|\mu) = \exp(\eta^\top x - a(\eta))$. Then, we can see that

$$\begin{aligned}
b(\eta) &= 1 \\
T(x) &= x \\
a(\eta) &= -\log(\mu_K) = \log\left(1 + \sum_{i=1}^{K-1} e^{\eta_i}\right)
\end{aligned}$$

To see that the generalized linear model corresponding to this function is the same as softmax regression, we see that $\mu = S(\eta)$, where $S(\eta)$ is exactly the softmax function. This implies that the generalized linear model of this distribution is the same as softmax regression.

■