Cassidy Lê
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.

The mean can be found by calculating the expected value. For Beta distributions, the expected value is defined as

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

Because $\Gamma(x) = (x-1)!$, we know that $B(a+1,b) = \frac{a!(b-1)!}{(a+b)!}$ and $B(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}$, so

$$\begin{aligned}
\mathbb{E}[\theta] &= \frac{a!(b-1)!}{(a+b)!}\cdot\frac{(a+b-1)!}{(a-1)!(b-1)!} \\
&= \frac{a!}{(a-1)!}\cdot\frac{(a+b-1)!}{(a+b)!} \\
\mathbb{E}[\theta] &= \frac{a}{a+b}.
\end{aligned}$$

Therefore, the $\boxed{\text{mean of } \theta \text{ is } \frac{a}{a+b}.}$ In order to compute the mode of θ, we must first solve for the gradient of $\mathbb{P}(\theta; a, b)$, which is

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

Then, set the gradient equal to zero and solve for θ, so

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

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

Therefore, the $\boxed{\text{mode of } \theta \text{ is } \frac{a-1}{a+b-2}.}$ Variance is defined as

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

First, let us determine $\mathbb{E}[\theta^2]$:

$$\mathbb{E}[\theta^2] = \int_0^1 \theta^2 \mathbb{P}(\theta; a, b)$$
$$= \frac{1}{B(a,b)} \int_0^1 \theta^(a+1)(1-\theta)^{b-1} d\theta$$
$$= \frac{B(a+2,b)}{B(a,b)}$$
$$= \frac{(a+1)!(b-1)!}{(a+b+1)!} \cdot \frac{(a+b-1)!}{(a-1)!(b-1)!}$$
$$= \frac{(a+1)!}{(a-1)!} \cdot \frac{(a+b-1)!}{(a+b+1)!}$$
$$\mathbb{E}[\theta^2] = \frac{a(a+1)}{(a+b)(a+b+1)}.$$

Since we know that the mean of θ is $\frac{a}{a+b}$, then $\mathbb{E}[\theta]^2 = \frac{a^2}{(a+b)^2}$. Thus, the variance of θ is

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

Therefore, the $\boxed{\text{variance of } \theta \text{ is } \frac{ab}{(a+b)^2(a+b+1)}.}$ ■

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

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

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).

■

In order to show that $\text{Cat}(\mathbf{x}|\mu)$ is in the exponential family, we must show that it can be rearranged into the form of

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

Consider taking the exponential logarithm of the product:

$$exp\left[\log \prod_{i=1}^{K} \mu_i^{x_i}\right].$$

We know $\log xy = \log x + \log y$, so

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

Since $\log x^y = y \log x$,

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

We know $\sum_{i=1}^{K} x_i = 1$, so $x_K = 1 - \sum_{i=1}^{K-1} x_i$ and $\mu_K = 1 - \sum_{i=1}^{K-1} \mu_i$. Thus,

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

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

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

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

Because $\log a - \log b = \log \frac{a}{b}$,

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

Therefore, we can see that the distribution $\mathrm{Cat}(\mathbf{x}|\mu)$ is in the form of the exponential family, where

$$\eta = \begin{bmatrix} \log \frac{\mu_i}{\mu_K} \\ \vdots \\ \log \frac{\mu_{K-1}}{\mu_K} \end{bmatrix}$$

$$b(\mathbf{y}) = 1$$

$$T(\mathbf{y}) = \mathbf{x}$$

$$a(\eta) = -\log \mu_K.$$

All that is left to show is that the general linear model μ for $\mathrm{Cat}(\mathbf{x}|\mu)$ is the same as the softmax function $h_\theta(x)$. The softmax regression indicates that the softmax function is

$$h_\theta(x) = \begin{bmatrix} \frac{e^{\theta_1^T x}}{1+\sum_{i=1}^{K-1} e^{\theta_i^T x}} \\ \frac{e^{\theta_2^T x}}{1+\sum_{i=1}^{K-1} e^{\theta_i^T x}} \\ \vdots \\ \frac{e^{\theta_{k-1}^T x}}{1+\sum_{i=1}^{K-1} e^{\theta_i^T x}} \end{bmatrix}.$$

From earlier, we determined that $\mu_K = 1 - \sum_{i=1}^{K-1} \mu_i$. And by looking at η, we see that $\mu_i = \mu_k e^{\eta_i}$. Therefore,

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

And so

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

Thus, μ is the same as the softmax function, being of the form

$$\mu = \begin{bmatrix} \frac{e^{\eta_1}}{1+\sum_{i=1}^{K-1} e^{\eta_1}} \\ \frac{e^{\eta_2}}{1+\sum_{i=1}^{K-1} e^{\eta_2}} \\ \vdots \\ \frac{e^{\eta_{k-1}}}{1+\sum_{i=1}^{K-1} e^{\eta_{k-1}}} \end{bmatrix}.$$