Shanni Lam
Math189R SU20
Homework 3
Thursday, June 16, 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.

Recall the definition of the Beta Function and this property of the Gamma function:

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

$$\Gamma(x+1) = x\Gamma(x)$$

The mean is the expected value: $\mathbb{E}[\theta] = \int_0^1 \theta\mathbb{P}(\theta; a, b)d\theta = \int_0^1 \theta(\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1})d\theta$
$= \frac{1}{B(a,b)}\int_0^1 \theta^a(1-\theta)^{b-1}d\theta = \frac{B(a+1,b)}{B(a,b)}$ using the definition of the Beta Function but letting a be $a+1$.
$= [\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}][\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}] = [\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}][\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}]$
$= [\frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)}][\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}]$ using the property of the Gamma function
$= \frac{a}{a+b}$ after cancelling terms.

Now, we can use the mean to find the variance, which is given by the formula $Var(\theta) = \mathbb{E}[(\theta - \mathbb{E}[\theta])^2] = \mathbb{E}[\theta^2] - \mathbb{E}[\theta]^2$.

First, find $\mathbb{E}[\theta^2]$.

$\mathbb{E}[\theta^2] = \int_0^1 \theta^2\mathbb{P}(\theta; a, b)d\theta = \int_0^1 \theta(\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1})d\theta$
$= \frac{1}{B(a,b)}\int_0^1 \theta^{a+1}(1-\theta)^{b-1}d\theta = \frac{B(a+2,b)}{B(a,b)}$ using the definition of the Beta Function but letting a be $a+2$.
$= [\frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)}][\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}] = [\frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)}][\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}]$
$= [\frac{(a+1)\Gamma(a+1)\Gamma(b)}{(a+b+1)\Gamma(a+b+1)}][\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}] = [\frac{a(a+1)\Gamma(a)\Gamma(b)}{(a+b)(a+b+1)\Gamma(a+b)}][\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}]$ using the property of the Gamma function repeatedly.
$= \frac{a(a+1)}{(a+b)(a+b+1)}$ after cancelling terms.

Now, compute the variance: $Var(\theta) = \mathbb{E}[(\theta - \mathbb{E}[\theta])^2] = \mathbb{E}[\theta^2] - \mathbb{E}[\theta]^2$

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

Next, we will find the mode by finding when $\nabla_\theta \mathbb{P}(\theta; a, b) = 0$ on the interval [0, 1]. We will use the unnormalized distribution (ignoring the $\frac{1}{B(a,b)}$ term because the constant term will be eliminated from simplifying the equation. Then,

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

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

■

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).

Recall the exponential family is in the form $\mathbb{P}(\mathbf{y};\boldsymbol{\eta}) = b(\mathbf{y})exp(\boldsymbol{\eta}^T T(\mathbf{y}) - a(\boldsymbol{\eta}))$.

Rewrite the multinomial distribution to include an exponential and logarithm:

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

Applying logarithms to a product creates a sum of terms, so the above becomes $exp(\sum_{i=1}^{K} log(\mu_i^{x_i})) = exp(\sum_{i=1}^{K} x_i log(\mu_i))$.

Since $\sum_{i=1}^{K} \mu_i = 1$ and $\sum_{i=1}^{K} x_i = 1$, then we only need to specify the first $K-1$ terms because the final terms x_K and μ_K will be automatically determined:

$$\mu_K = 1 - \sum_{i=1}^{K-1} \mu_i$$

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

Working with the result we got above, $Cat(\mathbf{x}|\mu) = exp(\sum_{i=1}^{K} x_i log(\mu_i))$
$= exp(\sum_{i=1}^{K-1} x_i log(\mu_i) + x_K log(\mu_K))$ when breaking up the summation.
$= exp[\sum_{i=1}^{K} x_i log(\mu_i) + (1 - \sum_{i=1}^{K-1} x_i) log(\mu_K)]$ due to how we defined x_K above.
$= exp[\sum_{i=1}^{K} x_i log(\mu_i) + log(\mu_K) - \sum_{i=1}^{K-1} x_i log(\mu_k)]$
$= exp[\sum_{i=1}^{K-1} x_i log(\frac{\mu_i}{\mu_K}) + log(\mu_K)]$

Then, compared to the exponential family form, since our multinomial distribution outputs x, let $b(x) = 1, T(x) = x, a(\eta) = -log(\mu_K)$.

Let the vector $\boldsymbol{\eta}$ be $\boldsymbol{\eta} = \begin{bmatrix} log(\frac{\mu_1}{\mu_K} \\ \cdots \\ log(\frac{\mu_{K-1}}{\mu_K} \end{bmatrix}$

Then, we see $\mu_i = \mu_K e^{\eta_i}$; consider $\boldsymbol{\eta_1} = log(\frac{\mu_1}{\mu_K}) = log(\frac{\mu_K e^{\eta_i}}{\mu_K} = log(e^{\eta_i} = \eta_1$ since $i = 1$.

Then, we can revise μ_K to be $\mu_K = 1 - \sum_{i=1}^{K-1} \mu_i = 1 - \sum_{i=1}^{K-1} \mu_K e^{\eta_i} = 1 - \mu_K \sum_{i=1}^{K-1} e^{\eta_i}$.

Rewrite the equation, so $\mu_K + \mu_K \sum_{i=1}^{K-1} e^{\eta_i} = 1 = \mu_K(1 + \sum_{i=1}^{K-1} e^{\eta_i}$.

Rearranging, we get $\mu_K = \frac{1}{1+\sum_{i=1}^{K-1} e^{\eta_i}}$.

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

We can revise $a(\boldsymbol{\eta})$ to be $a(\boldsymbol{\eta}) = -log(\mu_K) = log(\mu_K^{-1}) = log(1 + \sum_{i=1}^{K-1} e^{\eta_i}$

Thus, by having parameters that fit $b(y), T(y), \eta, a(\eta)$, the multinomial distribution is in the family of exponential dstributions. Recall the softmax function $S(\eta) = \frac{e^{\eta}}{\sum_{i=1}^{\eta}}$. $\mu = \frac{e^{\eta}}{\sum_{i=1}^{K-1} e^{\eta}} = S(\eta)$, so the generalized linear model of the multinomial distribution is the same as multinomial logistic regression or softmax regression.

■