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.

ANSWER SUMMARY:

$$\begin{aligned} Mean : \mathbb{E}[\theta] &= \frac{a}{a+b} \\ Mode : \theta^* &= \frac{a-1}{b+a-2} \\ Variance : Var[\theta] &= \frac{ab}{(a+b)^2(a+b+1)} \end{aligned}$$

PROOF BELOW:

The mean of a probability distribution is given by $\mathbb{E}[\theta] = \int_0^1 \theta \mathbb{P}(\theta; a, b) d\theta$, therefore we can find:

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

Since the integral of the probability function is one, we have the following:

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

We also know from the given equation:

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

Therefore the mean can be found as:

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

The mode of a probability distribution (peak of the curve) can be described by $\nabla_\theta \mathbb{P}(\theta; a, b) =$

0:

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

Next to find variance, we know that $Var[\theta] = \mathbb{E}[\theta^2] - \mathbb{E}[\theta]^2$:

$$
\begin{aligned}
Var[\theta] &= \mathbb{E}[\theta^2] - \mathbb{E}[\theta]^2 \\
&= \int_0^1 \theta^2 \left(\frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1} \right) d\theta - \left(\frac{a}{a+b} \right)^2 \\
&= \frac{1}{B(a,b)} \int_0^1 \theta^{a+1} (1-\theta)^{b-1} d\theta - \left(\frac{a}{a+b} \right)^2 \\
&= \frac{B(a+2,b)}{B(a,b)} - \left(\frac{a}{a+b} \right)^2 \\
&= \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+2+b)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} - \left(\frac{a}{a+b} \right)^2 \\
&= \frac{a(a+1)\Gamma(a)\Gamma(b)}{(a+b)(a+b+1)\Gamma(a+b)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} - \left(\frac{a}{a+b} \right)^2 \\
&= \frac{a(a+1)}{(a+b)(a+b+1)} - \frac{a^2}{(a+b)^2} \\
&= \frac{a(a+1)(a+b)}{(a+b)^2(a+b+1)} - \frac{a^2(a+b+1)}{(a+b)^2(a+b+1)} \\
&= \frac{(a^2+a)(a+b) - a^3 - a^2 b - a^2}{(a+b)^2(a+b+1)} \\
&= \frac{a^3 + a^2 b + a^2 + ab - a^3 - a^2 b - a^2}{(a+b)^2(a+b+1)} \\
&= \frac{ab}{(a+b)^2(a+b+1)}
\end{aligned}
$$

■

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).

A class of distributions is in the exponential family if:

$$\mathbb{P}(y;\eta) = b(y)exp(\eta^\top T(y)a(\eta))$$

Therefore we need to rearrange the multinoulli distribution to contain exponential:

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

From this form we define:

$$\eta = \begin{bmatrix} log(\frac{\mu_1}{\mu_k}) \\ log(\frac{\mu_2}{\mu_k}) \\ ... \\ log(\frac{\mu_{k-1}}{\mu_k}) \end{bmatrix}$$

$$\eta_i = log(\frac{\mu_i}{\mu_k})$$

So we know that :

$$exp(\eta_i) = \frac{\mu_i}{\mu_k}$$

$$\mu_i = \mu_k exp(\eta_i)$$

Therefore we can redefine μ_k and μ_i as follows:

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

$$\mu_k = 1 - \sum_{i=1}^{K-1} \mu_k \exp(\eta_i)$$

$$\mu_k + \sum_{i=1}^{K-1} \mu_k \exp(\eta_i) = 1$$

$$\mu_k = \frac{1}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)}$$

$$\mu_i = \mu_k exp(\eta_i)$$

$$\mu_i = \frac{1}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)} exp(\eta_i)$$

$$\mu_i = \frac{exp(\eta_i)}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)}$$

Thus we can write he multinoulli distribution in the form of the exponential family:

$$\text{Cat}(\mathbf{x}|\boldsymbol{\mu}) = exp(\eta^\top x - a(\eta))$$

where

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

Therefore we have proven that the distribution $\text{Cat}(\mathbf{x}|\boldsymbol{\mu})$ is in the exponential family.

The generalized linear model of the distribution $\text{Cat}(\mathbf{x}|\boldsymbol{\mu})$ is the same as the soft-max regression since $\mu = S(\eta)$ and $s(\eta)$ is the softmax function.

$\blacksquare$