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.

Derive The mean we must do the following: Show the integral definition of the expected value. Indicate substitution of the probability density function of the Beta distribution. Simplify the expression by pulling out the constant 1/B(a, b). Identify the integral as a Beta function. Express the Beta function in terms of Gamma functions. Simplify the Gamma functions using the property Gamma(n+1) = n*Gamma(n) and then
Cancel out like terms. We are able to do this because we remember that the integral of a Beta distribution is :

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

Therefore we can write the above as:

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

Now we can go on to derive the variance: The variance of a random variable θ is given by:

$$\text{Var}[\theta] = \mathbb{E}[(\theta - \mathbb{E}[\theta])^2] = \mathbb{E}[\theta^2] - (\mathbb{E}[\theta])^2 \tag{1}$$

To compute variance, we need to first compute $\mathbb{E}[\theta^2]$ using the same method as above, and we get:

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

Now we can use this result to calculate the variance:

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

Now we can finally move on to deriving the mode of θ.

To compute the mode, we want to find when $\nabla_\theta p(\theta; a, b) = 0$ on the interval $[0, 1]$.

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

where we can now compute the mode

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

$$(a-1)(1-\theta) = (b-1)\theta$$

$$(a+b-2)\theta = a-1$$

$$\theta^* = \frac{a-1}{a+b-2}$$

■

2 (**Murphy 9**) Show that the multinoulli 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 multinoulli logistic regression (softmax regression).

We know that this distribution can be written as:

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

Which when combined with the property of Logarithms leads to our derivation:

$$\log(ab) = \log(a) + \log(b)$$

$$\log(a^b) = b\log(a)$$

To show that the multinomial distribution is in the exponential family,we combine the definitions from above rewriting in terms of them:

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

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

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

Notice that $\sum_{i=1}^{K} \mu_i = 1$ and $\sum_{i=1}^{K} x_i = 1$, as a result we just need to specify the first $K-1$ of the terms, since the term x_K and μ_K will be automatically determined at the end:

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

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

We can therefore split up our summation into the following:

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

Now, let the vector η be

$$\eta = \begin{bmatrix} \log\left(\frac{\mu_1}{\mu_K}\right) \\ \vdots \\ \log\left(\frac{\mu_{K-1}}{\mu_K}\right) \end{bmatrix}$$

such that $\mu_i = \mu_K e^{\eta_i}$, and we make the substitution, so that we have:

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

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

Writing the distribution in the form of exponential family as $\text{Cat}(x|\mu) = \exp(\eta^T x - a(\eta))$:

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

Thus we conclude that the distribution $\text{Cat}(x|\mu)$ is in the exponential family. And $\mu = S(\eta)$, where $S(\eta)$ is the softmax function, which implies the GLM is the same as using softmax regression. ∎