Math189R SP19
Homework 3
Monday, Feb 18, 2019

Consulted Solution for **checking** 1 and 2.

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.

We will first find the mean of the relevant distribution. It follows that

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

We can simplify by using $B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta$, giving

$$\begin{aligned}&= \frac{B(a+1,b)}{B(a,b)} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\ \mathbb{E}[\theta] &= \boxed{\frac{a}{a+b}}.\end{aligned}$$

Now, we will find the variance of the relevant distribution.

$$\begin{aligned}Var[\theta] &= \mathbb{E}[(\theta - \mathbb{E}[\theta])^2] \\ &= \mathbb{E}[\theta^2] - \mathbb{E}[\theta]^2.\end{aligned}$$

We know what $\mathbb{E}[\theta]$ is and therefore know what $\mathbb{E}[\theta]^2$ is. We must then find $\mathbb{E}[\theta^2]$. Note

that the calculation is very similar to the mean and we can thus skip a handful of steps

$$\begin{aligned}\mathbb{E}[\theta^2] &= \frac{B(a+2,b)}{B(a,b)} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} \\ &= \frac{a(a+1)}{(a+b)(a+b+1)}.\end{aligned}$$

It then follows that

$$\begin{aligned}Var[\theta] &= \frac{a(a+1)}{(a+b)(a+b+1)} - \frac{a^2}{(a+b)^2} \\ Var[\theta] &= \frac{a}{a+b}(\frac{a+1}{a+b+1} - \frac{a}{a+b}) \\ &= \frac{a}{a+b}\frac{(a+b)(a+1) - a(a+b+1)}{(a+b)(a+b+1)} \\ &= \frac{a}{a+b}\frac{b}{(a+b)(a+b+1)} \\ Var[\theta] &= \boxed{\frac{ab}{(a+b)^2(a+b+1)}.}\end{aligned}$$

Now, we will find the mode of the relevant distribution. Since the mode is the most common value to occur and since we are working with a continuous PDF (not a discrete one) it follows that the mode is simply the roots of $\frac{d}{d\theta}(\mathbb{P}(\theta;a,b))$ where $\frac{d^2}{d\theta^2}(\mathbb{P}(\theta;a,b)) < 0$ or the maximum of the PDF.

Let us calculate the two derivatives

$$\frac{d}{d\theta}(\mathbb{P}(\theta;a,b)) = \frac{1}{B(a,b)}\theta^{a-2}(1-\theta)^{b-2}((a-1)(1-\theta) - (b-1)\theta)$$

Notice that the inner term of "$(a-1)(1-\theta) - (b-1)\theta$" is the only term which can be zero for $\theta \neq 0, 1$ and thus the roots are

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

Consider $\frac{a-1}{a+b-2} + \delta$ where $\delta > 0$ is a small quantity. If we plug this value into the term we used to find the roots we get

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

Since $a + b \geq 2$ it follows that this quantity is necessarily negative. We can then approximate

$$\frac{d^2}{d\theta^2}(\mathbb{P}(\theta; a, b)) \approx \frac{-(a+b-2)\delta - 0}{\delta} = -(a+b-2),$$

showing that $\frac{d^2}{d\theta^2}(\mathbb{P}(\theta; a, b)) < 0$ or that we really found the maximum and not some local minimum. Therefore the mode is

$$\boxed{\frac{a-1}{a+b-2}}.$$

■

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).

A distribution of the exponential family can be expressed as

$$f_X(x|\theta) = h(x)\exp[\eta(\theta)\cdot T(x) - A(\theta)].$$

Consider, then writing $\text{Cat}(\mathbf{x}, \boldsymbol{\mu})$ with log and exp.

$$\begin{aligned}
\text{Cat}(\mathbf{x}, \boldsymbol{\mu}) &= \prod_{i=1}^{K} \mu_i^{x_i} \\
&= \exp(\log(\prod_{i=1}^{K} \mu_i^{x_i})) \\
&= \exp(\sum_{i=1}^{K} \log(\mu_i^{x_i})) \\
&= \exp(\sum_{i=1}^{K} x_i \log(\mu_i))
\end{aligned}$$

By the definition of a probability distribution, it follows that $\sum_{i=1}^{K} \mu_i = 1$. Furthermore, Since $\mathbf{x}$ is a vector where a single entry is 1 - it also follows that $\sum_{i=1}^{K} x_i = 1$.

Since the multinouilli logistic regression uses μ_K as a pivot we need to separate μ_K. Then

$$\begin{aligned}
&= \exp(\sum_{i=1}^{K-1} x_i \log(\mu_i) + x_K \log(\mu_K)) \\
&= \exp(\sum_{i=1}^{K-1} x_i \log(\mu_i) + (1 - \sum_{i=1}^{K-1} x_i) \log(\mu_K)) \\
&= \exp(\sum_{i=1}^{K-1} x_i (\log(\mu_i) - \log(\mu_K)) + \log(\mu_K)) \\
&= \exp(\sum_{i=1}^{K-1} x_i \log(\mu_i/\mu_K) + \log(\mu_K)).
\end{aligned}$$

It then follows that $\boxed{h(x) = 1}$,

$$\boxed{\eta(\theta) = \begin{pmatrix} \log(\mu_1/\mu_K) \\ \dots \\ \log(\mu_{K-1}/\mu_K) \end{pmatrix}},$$

$$\boxed{T(x) = \begin{pmatrix} x_1 \\ \dots \\ x_{K-1} \end{pmatrix}},$$

and $A(\theta) = -\log(\mu_K)$. Showing that the multinoulli distribution is in the exponential family.

If we wish to write the exponential family in canonical form we can simply express $A(\theta)$ as

$$\boxed{-\log(1 - \mu_K \sum_{i=1}^{K-1} e^{\eta_i(\theta)})}.$$

This form showcases that $\mu_i = \mu_K e^{\eta_i(\theta)}$ which implies that

$$\mu_i = \frac{1}{\sum_{i=1}^{K} \mu_i / \mu_K} e^{\eta_i(\theta)}$$
$$\mu_i = \frac{1}{1 + \sum_{i=1}^{K-1} e^{\eta_i(\theta)}} e^{\eta_i(\theta)}$$

or that

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

For μ_K we can follow a similar approach - but do not need the initial term of $e^{\eta_i(\theta)}$, meaning that

$$\boxed{\mu_K = \frac{1}{1 + \sum_{i=1}^{K-1} e^{\eta_i(\theta)}}}.$$

We have then showed that $\boldsymbol{\mu}$ precisely follows a multinoulli logistic regression. ∎