Marc Laugharn
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.

Given $\Gamma(x+1) = x\Gamma(x)$ the mean is

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

The integrand is equivalent to $\mathbb{P}(\theta; a+1, b) \cdot B(a+1, b)$, so we have

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

The variance is $\mathbb{E}[\theta^2] - \mathbb{E}[\theta]^2$.

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

Therefore the variance is

$$\begin{aligned}\frac{(a+1)a}{(a+b+1)(a+b)} - (\frac{a}{a+b})^2 &= \frac{a^3+a^2b+a^2+ab-a^3-a^2b-a^2}{(a+b+1)(a+b)^2} \\ &= \frac{ab}{(a+b+1)(a+b)^2}\end{aligned}$$

The mode is the most common value of the pmf, i.e. the maximal probability value.

$$\begin{aligned}\arg\max_\theta \mathbb{P}(\theta; a, b) &= \nabla_\theta \mathbb{P}(\theta; a, b) \\ &\approx (a-1)\theta^{a-2} \cdot (1-\theta)^{b-1} + \theta^{a-1} \cdot (1-b)(1-\theta)^{b-2}\end{aligned}$$

Setting this to zero we get

$$\begin{aligned}(b-1)\theta^{a-1}(1-\theta)^{b-2} &= (a-1)\theta^{a-2}(1-\theta)^{b-1} \\ (b-1)\theta \cdot 1 &= (a-1) \cdot 1 \cdot (1-\theta) \\ a\theta + b\theta - \theta - \theta &= a-1 \\ \theta &= \frac{a-1}{a+b-2}\end{aligned}$$

■

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).

Let

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

Because it is a multinoulli distribution $\sum_i^K x_i = 1$ and $\sum_i^K \mu_i = 1$. WLOG let the K'th indices denote the non-free parameters, i.e. $x_k = 1 - \sum_{i=1}^K x_i$ and $\mu_k = 1 - \sum_{i=1}^K \mu_i$.

Then

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

Then $\eta_i = \log \frac{\mu_i}{\mu_k}$ for $i = 1, ..., K-1$ and $\mu_i = \mu_k \exp(\eta_i)$. To find μ_k in terms of η, let

$$\begin{aligned}
\mu_k &= 1 - \sum_{i=1}^{K-1} \mu_i \\
&= 1 - \mu_k \sum_{i=1}^{K-1} \exp(\eta_i) \\
\mu_k (1 + \sum_{i=1}^{K-1} \exp(\eta_i)) &= 1 \\
\mu_k &= \frac{1}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)}
\end{aligned}$$

Plugging this back into our equation for μ_i yields $\mu_i = \frac{\exp(\eta_i)}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)}$

To get this into exponential family form, let $b(\eta) = 1$, $T(\mathbf{x}) = \mathbf{x}$, $a(\eta) = -\log \mu_k = \log(1 + \sum_{i=1}^{K-1} \exp(\eta_i))$.

The softmax function is defined as $S(\eta)_i = \frac{e^{\eta_i}}{\sum_{i=1}^{K-1} e^{\eta_i}}$ for $i = 1...K-1$ and $S(\eta)_k = 1 - \sum_{i=1}^{K-1} S(\eta)_i$; since $\text{Cat}(\mathbf{x}|\boldsymbol{\mu})$ is in this form, and linear in the weights $\boldsymbol{\mu}$, $\text{Cat}(\mathbf{x}|\boldsymbol{\mu})$ is a generalized linear model equivalent to softmax regression over η. ∎