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.

Since $\mu = E[\theta]$, we are going to evaluate $\int_0^1 \mathbb{P}(\theta; a, b)$. This evaluates to:

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

Note that $\Gamma(z+1) = z\Gamma(z)$. The mean is therefore yields:

$$\frac{\Gamma(a+b) \cdot a \cdot \Gamma(a)}{\Gamma(a) \cdot (a+b) \cdot \Gamma(a+b)} = \frac{a}{a+b}$$

The mode is the θ at which $\mathbb{P}(\theta; a, b)$ is maximum. This point will either be at $\theta = 0$ or $\theta = 1$ or at one of the critical points of the PDF. Since $P(0; a, b) = P(1; a, b) = 0$, the maxima of the PDF has to be at one of the critical points.

$$\frac{\partial}{\partial\theta}P(\theta; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}((a-1)\theta^{a-2}(1-\theta)^{b} - 1 + (b-2)\theta^{a-1}(1-\theta)^{b-2}) = 0$$

Therefore,

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

Then,

$$a - a\theta - 1 + \theta = b\theta - \theta$$

, So, the mode is at:

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

The variance is defined as $\sigma^2 = E[\theta^2] - E[\theta]^2$. Evaluating this yields:

$$\frac{1}{B(a,b)}\int_0^1 \theta^{a+1}(1-\theta)^{b-1}d\theta + \frac{a^2}{(a+b)^2}$$

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

∎

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).

First define $x_k = \mathbf{1}(x = k)$ (an indicator function) We can express the above as:

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

Note that since we have K total parametres which are probabilities, our model can be parametrised by $K-1$ parametres where $x_K = 1 - \sum_{i=1}^{K-1} x_i$ and $\mu_K = 1 - \sum_{i=1}^{K-1} \mu_i$. Therefore, the above expression can be written as:

$$\text{Cat}(x|\boldsymbol{\mu}) = \exp[\sum_{i=1}^{K-1} x_i \log(\mu_i) + \log(\mu_K)(1 - \sum_{i=1}^{K-1} x_i)]$$

$$= \exp[\sum_{i=1}^{K-1} x_i \log(\frac{\mu_i}{\mu_k}) + \log(\mu_K)]$$

$$= \exp[\boldsymbol{\theta}^T \mathbf{x} + \log(\mu_K)]$$

Where $\boldsymbol{\theta} = [\log(\frac{\mu_1}{\mu_k}) \dots \log(\frac{\mu_{k-1}}{\mu_k})]^T$ and $\mathbf{x} = [\mathbf{x}_1 \dots \mathbf{x}_{K-1}]^T$.

The multinoulli distribution is therefore in the exponential family!

We know that $\theta_i = \log(\frac{\mu_i}{\mu_K})$, and $\mu_K = \frac{1}{1+\sum_{j=1}^{K-1} e^{\theta_j}}$ so,

$$\mu_i = e^{\theta_i} \mu_K = \frac{e^{\theta_i}}{1 + \sum_{j=1}^{K-1} e^{\theta_j}}$$

From this we have:

$$\mu_K = 1 - \sum_{j=1}^{K-1} \frac{e^{\theta_j}}{1 + \sum_{j=1}^{K-1} e^{\theta_j}} = \frac{1}{1 + \sum_{j=1}^{K-1} e^{\theta_j}}$$

If we expand our $\boldsymbol{\theta}$ to include $\theta_k = \log(\frac{\mu_K}{\mu_K}) = 0$, we can write the following:

$$\mu_i = \frac{e^{\theta_i}}{\sum_{j=1}^{K} e^{\theta_j}}$$

So,

$$\boldsymbol{\mu} = \mathcal{S}(\boldsymbol{\theta})$$

Where $\mathcal{S}$ is the softmax function.

We therefore conclude that the multinoulli distribution lies in the exponential family with $b(\mathbf{x}) = 1$, $\boldsymbol{\theta} = [\log(\frac{\mu_1}{\mu_k}) \ldots \log(\frac{\mu_{k-1}}{\mu_k})]^T$, and
$A(\theta) = 1 + \sum_{j=1}^{K-1} e^{\theta_j}$. We have also shown that the actual parametres of the model, $\boldsymbol{\mu}$, can be derived by taking $\mathcal{S}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of the natural parametres of the model.

■