Nico Espinosa Dice
Math189R SP19
Homework 3
Monday, Feb 18, 2019

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.

The mean of θ is equivalent to its expected value. The expected value of θ is given in Murphy 2.2.7. Note that the bounds of the integral are from 0 to 1 because θ is distributed according to the Beta distribution, which goes from 0 to 1.

$$\begin{aligned}
&\mathbb{E}(\theta) \triangleq \int_0^1 \theta \mathbb{P}(\theta; a, b) d\theta \\
&= \int_0^1 \theta \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1} d\theta \\
&= \int_0^1 \frac{1}{B(a,b)}\theta^{a}(1-\theta)^{b-1} d\theta \\
&= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a}(1-\theta)^{b-1} d\theta \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 \theta^{a}(1-\theta)^{b-1} d\theta \quad \text{(since } B(a,b) \text{ does not depend on } \theta\text{)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 \theta^{(a+1)-1}(1-\theta)^{b-1} d\theta \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} B(a+1, b) \quad \text{(since } B(x,y) = \int_0^1 u^{x-1}(1-u)^{y-1} du\text{)}^1 \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)} \frac{a\Gamma(a)}{(a+b)\Gamma(a+b)} \qquad \text{(since a Gamma function property is: } \Gamma(n+1) = n\Gamma(n)\text{)}^2 \\
&= \frac{a}{a+b}.
\end{aligned}$$

[1] This equation for the Beta function is not given in Murphy. I found this equation on Wikipedia.
[2] I also found this property online.

We know by Murphy 2.24-2.25 that the variance of θ is given by:

$$
\begin{aligned}
&var[\theta] = \mathbb{E}[\theta^2] - \mu^2 \\
&= \mathbb{E}[\theta^2] - (\mathbb{E}[\theta])^2 \\
&= \int_0^1 \theta^2 \mathbb{P}(\theta; a, b) d\theta - (\frac{a}{a+b})^2 \quad \text{(using the result for the mean found above)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^{a+1}(1-\theta)^{b-1} d\theta - (\frac{a}{a+b})^2 \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} B(a+2, b) \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)} \frac{(a+1)\Gamma(a+1)}{(a+b+1)\Gamma(a+b+1)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)} \frac{(a+1)(a)\Gamma(a)}{(a+b+1)(a+b)\Gamma(a+b)} \\
&= \frac{(a+1)(a)}{(a+b+1)(a+b)}.
\end{aligned}
$$

$\mathbb{P}(\theta)$ describes the probability of θ. The mode of θ is the value of θ that has the highest probability. To find this value of θ, we will find the critical points of $\mathbb{P}$. To do this, we take the gradient of the probability function, set the expression equal to zero, and solve for the values of θ:

$$
\begin{aligned}
&0 = \nabla_\theta(\mathbb{P}(\theta)) \\
&0 = \nabla_\theta[\frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1}] \\
&0 = \frac{1}{B(a,b)} \nabla_\theta[\theta^{a-1}(1-\theta)^{b-1}] \\
&0 = \nabla_\theta[\theta^{a-1}(1-\theta)^{b-1}] \\
&0 = (1-\theta)^{b-1} \nabla_\theta[\theta^{a-1}] + \theta^{a-1} \nabla_\theta[(1-\theta)^{b-1}] \quad \text{(product rule)} \\
&0 = (1-\theta)^{b-1}(a-1)\theta^{a-2} - \theta^{a-1}(b-1)(1-\theta)^{b-2} \\
&\theta^{a-1}(b-1)(1-\theta)^{b-2} = (1-\theta)^{b-1}(a-1)\theta^{a-2} \\
&\theta(b-1) = (1-\theta)(a-1) \\
&\theta b - \theta = a - 1 - a\theta + \theta \\
&\theta(b-1+a-1) = a-1 \\
&\theta = \frac{a-1}{b+a-2}.
\end{aligned}
$$

Thus, the mode of $\theta = \frac{a-1}{b+a-2}$. Since there is only one critical point, and we are asked to find a local maximum, we can assume that this value of θ is the maximum. If we wanted to prove that it is in fact a maximum, we could use the second derivative test. ∎

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, let us consider the multinoulli distribution:

$$\begin{aligned}
Cat(\mathbf{x}|\boldsymbol{\mu}) &= \prod_{i=1}^{K} \mu_i^{x_i} \\
&= \exp[\sum_{k=1}^{K} x_k \log \mu_k] \quad \text{(given in Murphy 9.2.2)} \\
&= \exp[\sum_{k=1}^{K-1} x_k \log \mu_k + x_K \log \mu_K] \\
&= \exp[\sum_{k=1}^{K-1} x_k \log \mu_k + (1 - \sum_{k=1}^{K-1} x_k) \log \mu_K \quad \text{(since } \sum_{k=1}^{K} \mu_k = 1 = \sum_{k=1}^{K} x_k)^3 \\
&= \exp[\sum_{k=1}^{K-1} x_k \log \mu_k - \sum_{k=1}^{K-1} x_k \log \mu_K + \log \mu_K] \\
&= \exp[\sum_{k=1}^{K-1} x_k \log \frac{\mu_k}{\mu_K} + \log \mu_K].
\end{aligned}$$

We know that the exponential family form is:

$$Cat(\mathbf{x}|\boldsymbol{\theta}) = \exp(\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})).$$

If we let $\boldsymbol{\theta} = \begin{bmatrix} \log \frac{\mu_1}{\mu_K} & \log \frac{\mu_2}{\mu_K} & \dots & \log \frac{\mu_{K-1}}{\mu_K} \end{bmatrix}$, then we see that $\mu_k = \mu_K e^{\theta_k}$. Consequently, we have:

$$\begin{aligned}
\mu_K &= 1 - \sum_{k=1}^{K-1} \mu_k = 1 - \sum_{k=1}^{K-1} \mu_K e^{\theta_k}, \text{ so} \\
\mu_K &= \frac{1}{1 + \sum_{k=1}^{K-1} e^{\theta_k}}.
\end{aligned}$$

Furthermore, let $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$, and let $A(\boldsymbol{\theta}) = -\log \mu_K = -\log \frac{1}{1+\sum_{k=1}^{K-1} e^{\theta_k}} = \log(1 + \sum_{k=1}^{K-1} e^{\theta_k})$.

[3] We know that $\sum_{k=1}^{K} \mu_k = 1$ since μ represents probability, so the probability of all of the possible outcomes must be 1. We know that $\sum_{k=1}^{K} x_k = 1$ since the multinoulli distribution is defined as having $Cat(\mathbf{x}, \boldsymbol{\mu}) \stackrel{\Delta}{=} Mu(\mathbf{x}|1, \boldsymbol{\theta})$, where $n = 1$ is the size of the set that will be divided up into subsets with sizes x_1 up to x_K.

Thus, we see that we can express the multinoulli distribution in exponential family form, so it is in the exponential family.

Next, we will consider $S(\boldsymbol{\theta})$, so:

$$S(\boldsymbol{\theta}_c) = \frac{e^{\theta_c}}{\sum_{c'=1}^{C} e^{n_{c'}}} \quad \text{(by definition of the softmax function)}$$
$$= \frac{e^{\log \frac{\mu_c}{\mu_K}}}{\sum_{c'=1}^{C} e^{\log \frac{\mu_c'}{\mu_K}}}$$
$$= \frac{\frac{\mu_c}{\mu_K}}{\frac{1}{\mu_K}\sum_{c'=1}^{C} \mu_c'} \quad \text{(since } \mu_K \text{ does not depend on } c'\text{)}$$
$$= \frac{\frac{\mu_c}{\mu_K}}{\frac{1}{\mu_K}} \quad \text{(since } \sum_{c'=1}^{C} \mu_c' = 1\text{, shown on the previous page)}$$
$$= \mu_c.$$

Thus, we see that $S(\boldsymbol{\theta}) = \boldsymbol{\mu}$. Therefore, since S is the softmax function, we know that the generalized linear model corresponding to the multinoulli distribution is the same as multinoulli logistic regression or softmax regression. ■