Andy Liu
Math189R SU19
Homework 3
June 18, 2020

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.

(Mean) Note that $\mathbb{E}[\theta] = \int_0^1 \theta\mathbb{P}(\theta; a, b)\mathrm{d}\theta = \int_0^1 \theta \cdot \frac{1}{B(a,b)} \cdot \theta^{a-1} \cdot (1-\theta)^{b-1}\mathrm{d}\theta = \frac{1}{B(a,b)}\int_0^1 \theta^a \cdot (1-\theta)^{b-1}\mathrm{d}\theta$. However, from the definition of $B(a,b)$, note that our integral is just equal to $B(a+1,b)$. Thus, our expected value is equal to $\boxed{\frac{B(a+1,b)}{B(a,b)}}$.

By the definition of $B(a,b)$, this can be rewritten as $\frac{\Gamma(a+1)\Gamma(b)\Gamma(a+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+1)}$. Since $\Gamma(x+1) = x\Gamma(x)$, this reduces to $\mathbb{E}[\theta] = \frac{a}{a+b}$, as desired.

(Mode) To find the mode, it suffices to find the value of θ that maximizes the expression $\mathbb{P}(\theta; a, b) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}$. This is equivalent to maximizing $\theta^{a-1}(1-\theta)^{b-1}$, as $B(a,b)$ is a constant. Using the product rule, we compute that the derivative of $\theta^{a-1}(1-\theta)^{b-1}$ is equal to $(a-1)\theta^{a-2}(1-\theta)^{b-1} + (1-b)\theta^{a-1}(1-\theta)^{b-2} = \theta^{a-2}(1-\theta)^{b-2}((a-1)(1-\theta) + (1-b)(\theta))$. Setting the derivative equal to zero, note that this implies that $\theta = 0, \theta = 1$, or $(a-1)(1-\theta) + (1-b)(\theta) = 0 \Rightarrow \theta = \frac{a-1}{a+b-2}$. Since this final critical point is the only one for which $\mathbb{P} > 0$, the mode must be $\boxed{\frac{a-1}{a+b-2}}$.

(Variance) Note that the variance is equal to $\mathbb{E}[\theta^2] - \mathbb{E}[\theta]^2$. Since we have shown in our calcuations to find the mean that $\mathbb{E}[\theta] = \frac{a}{a+b}$, it suffices to find $\mathbb{E}[\theta^2] = \int_0^1 \theta^2\mathbb{P}(\theta; a, b)\mathrm{d}\theta = \int_0^1 \theta^2 \cdot \frac{1}{B(a,b)} \cdot \theta^{a-1} \cdot (1-\theta)^{b-1}\mathrm{d}\theta$. Again, we factor the constant $\frac{1}{B(a,b)}$ out and rewrite the integral in terms of a B expression to find that our expression is equal to $\frac{B(a+2,b)}{B(a,b)}$.

Rewriting in our Beta functions in terms of Gamma functions, this becomes $\dfrac{\Gamma(a+2)\Gamma(b)\Gamma(a+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+2)}$.

Since $\Gamma(x+2) = x(x+1)\Gamma(x)$, this simplifies to $\dfrac{a(a+1)}{(a+b)(a+b+1)}$.

Thus, the variance is equal to $\mathbb{E}[\theta^2] - \mathbb{E}[\theta]^2 = \dfrac{a(a+1)}{(a+b)(a+b+1)} - \dfrac{a^2}{(a+b)^2}$

$$= \frac{a(a+1)(a+b) - a^2(a+b+1)}{(a+b)^2(a+b+1)} = \boxed{\frac{ab}{(a+b)^2(a+b+1)}}.$$

■

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).

Note that $\text{Cat}(\mathbf{x}|\boldsymbol{\mu}) = \exp(\log(\prod_{i=1}^{K} \mu_i^{x_i})) = \exp(\sum_{i=1}^{K} x_i \log(\mu_i))$. We can rewrite this as $\exp(\sum_{i=1}^{K-1} x_i \log(\mu_i) + x_K \log(\mu_K))$. Additionally, since $\sum x_i = \sum \mu_i = 1$ (by the definition of the distribution), $x_K = 1 - \sum_{i=1}^{K-1} x_i$; substituting this into our expression, we find that it is equal to

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

If we set η to equal the vector $\begin{bmatrix} \log(\frac{\mu_1}{\mu_K}) \\ \log(\frac{\mu_2}{\mu_K}) \\ \cdots \\ \log(\frac{\mu_{K-1}}{\mu_K}) \end{bmatrix}$, then our distribution function becomes $\exp(\mathbf{x}\eta + \log(\mu_k)$. For this to be in the exponential family, it suffices to write $\log(\mu_k)$ in terms of η.

However, note that $\mu_K = 1 - \sum_{i=1}^{K-1} \mu_i = \sum_{i=1}^{K-1} \mu_K e_i^{\eta}$, so we can rewrite $\log(\mu_K)$ as $\log(\frac{1}{1 + \sum_{i=1}^{K-1} e_i^{\eta}})$. Thus, our distribution function becomes

$$\exp(\mathbf{x}\eta + \log(\frac{1}{1 + \sum_{i=1}^{K-1} e_i^{\eta}}))$$

Where $b(\eta) = 1$, $T(\mathbf{x}) = \mathbf{x}$, and $a(\eta) = -\log(\frac{1}{1 + \sum_{i=1}^{K-1} e_i^{\eta}}) = \log(1 + \sum_{i=1}^{K-1} e_i^{\eta})$. Thus, our distribution is in the exponential family.

Additionally, since $\mu_i = \mu_i = \mu_K e_i^{\eta} = \frac{e_i^{\eta}}{1 + \sum_{i=1}^{K-1} e_i^{\eta}}$, which is the softmax function, the generalized linear model of the distribution is also the softmax regression. ∎