Henry Trinh

HW3

1. Suppose $\theta \sim \text{Beta}(a,b)$, such that:

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

where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the Beta function and $\Gamma(x)$ is gamma function. Derive the mean, median, and variance of θ.

To derive the mean, we must first show the integral definition of the expected value, $E[\theta]$. Then, we can substitute the prob. density function of a Beta distribution, which is given as:

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

Also, the gamma function has the following property, $\Gamma(n+1) = n\Gamma(n)$, so:

$$E[\theta] = \int_0^1 \theta P(\theta; a,b)\, d\theta = \int_0^1 \theta \left[\frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1}\right] d\theta$$

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

$$= \frac{a\Gamma(a)\cdot\Gamma(b)}{(a+b)\Gamma(a+b)} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \Rightarrow \boxed{\frac{a}{a+b} = E[\theta]}$$

To derive the variance, $\text{var}[\theta]$, we must compute the mean-squared difference, or $E[\theta^2] - E[\theta]^2$. This is equivalent to the following:

$$\text{var}[\theta] = E[\theta^2] - E[\theta]^2 = \frac{a(a+1)}{(a+b)(a+b+1)} - \left[\frac{a}{a+b}\right]^2 = \frac{a(a+1)(a+b) - a^2(a+b+1)}{(a+b)^2(a+b+1)}$$

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

Then, to compute the mode of θ, we must find the maximum of the density function for the beta distribution. Hence, we compute:

$$\text{mode}[\theta] \Rightarrow \frac{1}{B(a,b)} \frac{d}{d\theta} \theta^{a-1}(1-\theta)^{b-1} = \frac{1}{B(a,b)}\left[(a-1)\theta^{a-2}(1-\theta)^{b-1} - \theta^{a-1}(b-1)(1-\theta)^{b-2}\right]$$

$$= \frac{1}{B(a,b)} \theta^{a-2}(1-\theta)^{b-2}\left[(a-1)(1-\theta) - (b-1)\theta\right]$$

As this represents the derivative of our probabalistic function, we now equate the result to 0 to determine the maximum (mode) value:

so, $(a-1)(1-\theta) - (b-1)\theta = 0 \Rightarrow a - 1 - \theta(a-1+b-1) = 0$

$\therefore \boxed{\theta^* = \frac{a-1}{a+b-2}}$ which exists when a,b are on the interval $[0,1]$.

2. Show that the multinoulli distribution $Cat(x,\mu) = \prod_{i=1}^{K} \mu_i^{x_i}$ is in the exponential family. Then, show that the generalized linear model that corresponds to the distribution is the same as multinouli logistic regression (aka softmax regression).

Given a measure η, we define an exponential family of probability distribution to have the following general form:

$$P(x \mid \eta) = h(x)\exp(\eta^T T(x) - A(\eta))$$ * Source: Berkeley.edu

for a parameter vector η and for functions T and h. Now, using our initial givens, we rewrite them to include some logit that is useful to us:

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

$$= \exp(\sum_{i=1}^{K} x_i \cdot \log(\mu_i)).$$

Then, since $\sum_{i=1}^{K} \mu_i = \sum_{i=1}^{K} x_i = 1$, as the integral of a distribution is 1, we will denote the first K-1 terms as follows:

$$\mu_K = 1 - \sum_{i=1}^{K-1} \mu_i \; ; \; x_K = 1 - \sum_{i=1}^{K-1} x_i$$

Hence, we rewrite the multinoulli distribution as:

$$Cat(x|\mu) = \exp(\sum_{K=1}^{K} x_K \log(\mu_K)) = \exp(\sum_{K=1}^{K-1} x_i \log(\mu_i) + x_K \log(\mu_K))$$

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

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

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

Now, suppose the parameter vector η is: $\eta = \begin{bmatrix} \log(\frac{\mu_1}{\mu_K}) \\ \vdots \\ \log(\frac{\mu_{K-1}}{\mu_K}) \end{bmatrix}$, such that $\mu_i = \mu_K e^{\eta_i}$. Then, by substitution, we see that:

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

$$\therefore \mu_i = \mu_K e^{\eta_i} = \frac{e^{\eta_i}}{1 + \sum_{i=1}^{K-1} e^{\eta_i}}$$

We can then write the distribution in the form of an exponential family: Let $h(\eta) = 1$ and $T(x) = x$, as per our known equation. We also have that $A(\eta) = -\log(\mu_K) = \log(1 + \sum_{i=1}^{K-1} e^{\eta_i})$

$\therefore$ $\boxed{Cat(x|\mu) \text{ is in the exponential family}}$ as desired.

Since $\mu = SM(\eta)$, where $SM(\eta)$ is the softmax function, this implies that the general linear model (GLM) of this distribution is the same as softmax regression, or multinoulli logistic regression.