1. $\mathbb{P}(\theta, a, b) = \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1}$

Oscar Scholin
179 HW3

The mean is : $\langle \mathbb{P} \rangle = \int_0^1 \theta \, \mathbb{P}(\theta, a, b) \, d\theta = \int_0^1 \theta \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1} \, d\theta$

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

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

Note $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

$$= \frac{B(a+1, b)}{B(a, b)}$$

$$= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+1+b)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

Note $\Gamma(x+1) = x\Gamma(x)$

$$= \frac{\Gamma(a+1)\Gamma(a+b)}{\Gamma(a+1+b)\Gamma(a)}$$

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

$$= \boxed{\frac{a}{a+b}}$$

$\sigma^2 = \langle \mathbb{P}^2 \rangle - \langle \mathbb{P} \rangle^2$, so we need $\langle \mathbb{P}^2 \rangle$ :

$$\langle \mathbb{P}^2 \rangle = \int_0^1 \theta^2 \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1} \, d\theta$$

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

$$= \frac{B(a+2, b)}{B(a, b)}$$

$$= \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+2+b)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

$$= \frac{\Gamma(a+2)\Gamma(a+b)}{\Gamma(a+2+b)\Gamma(a)}$$

$$= \frac{(a+1)\,\Gamma(a+1)\,\Gamma(a+b)}{(a+1+b)\,\Gamma(a+b+1)\,\Gamma(a)}$$

$$= \frac{(a+1)\,a\,\Gamma(a)\,\Gamma(a+b)}{(a+1+b)(a+b)\,\Gamma(a+b)\,\Gamma(a)}$$

$$= \frac{a(a+1)}{(a+b)(a+b+1)}$$

$$\therefore \sigma^2 = \frac{a(a+1)}{(a+b)(a+b+1)} - \left(\frac{a}{a+b}\right)^2$$

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

$$= \frac{\cancel{a^3} + \cancel{a^2 b} + \cancel{a^2} + ba - \cancel{a^3} - \cancel{ba^2} - \cancel{a^2}}{(a+b)^2(a+b+1)}$$

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

where PDF is maximized

Finally we find the mode, which is when $\nabla_\theta \, \mathbb{P}(\theta, a, b) = 0$ (so we can neglect the normalization term):

$$\nabla_\theta \, \mathbb{P}(\theta, a, b) = \nabla_\theta \left[ \theta^{a-1} (1-\theta)^{b-1} \right] = 0$$

$$= (a-1)\,\theta^{a-2}(1-\theta)^{b-1} - (b-1)\,\theta^{a-1}(1-\theta)^{b-2} = 0 \quad \text{by Chain Rule}$$

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

$$(a-1)(1-\theta) = (b-1)\,\theta$$

$$\therefore (a-1+b-1)\,\theta = a-1$$

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

2. Consider $\overset{\text{categorical}}{\swarrow}$

$$\text{Cat}(\vec{x} \mid \vec{\mu}) = \prod_{i=1}^{K} \mu_i^{k_i}.$$

WTS in exponential family and that corresponding generalized linear model is same as multinomial logistic regression.

$\vec{x}$: 1-hot encoded vector

$\vec{\mu}$: probability vector

Being in the exponential family means:

$$\mathbb{P}(\vec{y}, \vec{\eta}) = h(\vec{y}) \exp(\eta^T T(\vec{y}) - A(\vec{\eta})).$$

Rewrite $\text{Cat}(\vec{x} \mid \vec{\mu})$ as follows:

$$\exp\left[\ln \prod_{i=1}^{K} \mu_i^{k_i}\right] = \exp\left[\sum_{i=1}^{K} \ln(\mu_i^{x_i})\right]$$

$$= \exp\left[\sum_{i=1}^{K} x_i \ln \mu_i\right]$$

Note $\sum_{i=1}^{K} \mu_i = 1$, so we can write $\mu_K = 1 - \sum_{i=1}^{K-1} \mu_i$

$$\sum_{i=1}^{K} x_i = 1 \implies x_K = 1 - \sum_{i=1}^{K-1} x_i$$

$$\therefore \text{Cat}(\vec{x} \mid \vec{\mu}) = \exp\left[\sum_{i=1}^{K} x_i \ln \mu_i\right] = \exp\left[\sum_{i=1}^{K-1} x_i \ln \mu_i + \left(1 - \sum_{i=1}^{K-1} x_i\right) \ln \mu_K\right]$$

$$= \exp\left[\sum_{i=1}^{K-1} x_i (\ln \mu_i - \ln \mu_K) + \ln \mu_K\right]$$

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

$$\left.\begin{aligned} \therefore \text{let } \vec{\eta} &= \left[\ln\left(\frac{\mu_1}{\mu_K}\right), \ln\left(\frac{\mu_2}{\mu_K}\right), \ldots, \ln\left(\frac{\mu_{K-1}}{\mu_K}\right)\right] \\ T(\vec{x}) &= \vec{x} \\ A(\vec{\eta}) &= -\ln \mu_K \\ h(\vec{\eta}) &= 1 \end{aligned}\right\} \therefore \text{Cat}(\vec{x} \mid \vec{\mu}) \text{ is in exponential family.}$$

Now we must show this is same as softmax regression.

Observe

$$\mu_K = 1 - \sum_{i=1}^{K-1} \mu_i = 1 - \sum_{i=1}^{K-1} \mu_K e^{\vec{\eta}_i}$$

$$\therefore \mu_K \left( 1 + \sum_{i=1}^{K-1} e^{\vec{\eta}_i} \right) = 1 \implies \mu_K = \frac{1}{1 + \sum_{i=1}^{K-1} e^{\vec{\eta}_i}}$$

$$\therefore \vec{\mu} = \frac{e^{\vec{\eta}_i}}{1 - \sum_{i=1}^{K-1} e^{\vec{\eta}_i}} = \mathrm{Softmax}(\vec{\eta}). \qquad \square$$