An Introduction to Exponential Family

中南财经政法大学　应用统计 1701　曹奕涵

指数族分布通常有很多很好的性质，包括在贝叶斯推断、统计计算、信息论中，都有很广泛的应用。本文将从指数族开始介绍。

1　Exponential Family

我们假设空间 χ 中存在某个 base measure μ，并且存在一个有效统计量：$\phi : \chi \to R^d$。对于这样一个函数 ϕ，令 $\theta \in R^d$ 为规范参数的相关向量，那么：

Definition 1 对于一个函数 ϕ 和一个基础度量 μ，指数族为一系列的具有以下密度函数 p_θ 的分布：

$$p_\theta(x) = \exp(< \theta, \phi(x) > -A(\theta))$$

并且，函数 A 是对数配分函数（也可以称为累积量生成函数）：

$$A(\theta) := \log \int_\chi \exp(< \theta, \phi(x) >) d\mu(x)$$

且 A 是有限的。

一般来说，以下定义会更加方便一些：

Definition 2 对于一个基函数 $h : \chi \to R_+$，指数族分布表示为：

$$p_\theta(x) = h(x) \exp(< \theta, \phi(x) > -A(\theta))$$

以上就是指数族分布的一个较为基础的定义，为了更加正式的定义指数族分布，我们先介绍指数倾斜 (exponential tilting) 的构造：

Definition 3 设 (χ, F_χ, v) 上的分布 P 具有概率密度: $\frac{dP}{dv} =: h$，由 $Borel$ 函数可以导出矩母函数：

$$M(\xi) = E_{X \sim P}[e^{\xi^T T(X)}] = \int_\chi e^{\xi^T T(X)} h(x) dv(x)$$

对于 $\Xi \in \{\xi \in R^m : M(\xi \leq \infty)\}$ 记累计量生成函数：$K(\xi) = \log M(\xi)$，那么：

$$f_\xi(x) := \frac{1}{M(\xi)} e^{\xi^T T(x)} h(x) = exp\{\xi^T T(x) - K(\xi)\} h(x)$$

定义了一族分布 $\{f_\xi dv : \xi \in \Xi\}$，该分布为自然指数族 (natural exponential family)。

上述定义就为指数倾斜，可以查看UCB stat210A的代码，了解如何在 R 中构造一个指数倾斜。

下面我们来定义一般的指数族分布：

Definition 4 一个参数族 $P = \{P_\theta : \theta \in \Theta\}$ 定义在空间 (χ, F_χ, v) 上，如果概率密度有：

$$\frac{dP_\theta}{dv}(x) = \exp\{\xi(\theta)^T T(x) - B(\theta)\} h(x)$$

，称 $\Xi = \{\xi(\theta) : \theta \in \Theta\}$ 为自然参数空间，重新参数化，得到：

$$\{\exp\{\xi^T T(x) - B(\theta(\xi))\} h(x) dv(x) : \xi \in \Xi\}$$

为指数族的规范形式。

2 Some Examples on Exponential Family

2.1 Bernoulli Distribution:

令 μ 为 $\{0, 1\}$ 上的计数测度，$\theta = \log \frac{p}{1-p}$，那么，伯努利可以写成：

$$\begin{aligned} P(X = x) = p(x) = p^x (1-p)^{1-x} &= \exp(x \log p - x \log(1-p)) \\ &= exp(x \log \frac{p}{1-p} + \log(1-p)) \\ &= \exp(x\theta - \log(1 + e^\theta)) \end{aligned} \tag{1}$$

它的对数配分为：$A(\theta) = \log(1 + e^\theta)$.

2.2 Poisson Distribution:

泊松分布的参数 $\lambda > 0$。令 μ 为 $\{0, 1, ...\}$ 上的计数测度，$\theta = \log\lambda$，那么：

$$p(x) = \frac{1}{x!}\lambda^x e^{-\lambda} = \exp(x\log\lambda - \lambda)\frac{1}{x!} = \exp(x\theta - e^\theta)\frac{1}{x!} \tag{2}$$

其中，$h(x) = \frac{1}{x!}$，对数配分 $A(\theta) = e^\theta$

2.3 Normal Distribution:

正态分布中，令 μ 为 $(-\infty, \infty)$ 上的勒贝格测度，并且，重参数化 $N(\mu, \Sigma)$，有 $\Theta = \Sigma^{-1}$，$\theta = \Sigma^{-1}$，得到可以用欧式内积 $<, >$ 表达的：

$$p_{\theta,\Theta}(x) \propto \exp(<\theta, x> + \frac{1}{2} < xx^T, \Theta >)$$

为了表达的更加易懂，也可以写成：

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{(2\pi)}}\exp\{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{\log(\sigma^2)}{2}\}$$

同样的，前面三项可以看作是内积，对数配分为后面的 $\frac{\log(\sigma^2)}{2}$.

2.4 Geometric Distribution:

几何分布中，参数 $p \in (0, 1)$，对于 N^* 上的计数测度有：

$$f_p(x) = (1-p)^{x-1}p = \exp\{(x-1)\log(1-p) + \log p\}$$

对数配分 $A(\theta) = \log p$

2.5 Binomial Distribution:

二项分布中，参数 $p \in (0, 1)$，对于 $\{0, 1, ..., n\}$ 上的计数测度，有：

$$f_p(x) = C_n^x p^x (1-p)^{n-x} = exp\{xlogit(p) + n\log(1-p)\}C_n^x$$

其中，对数配分为 $A(\theta) = -n\log(1-p)$, $h(x) = C_n^x$, logit 代表 logit(p)。

2.6 Negative Binomial Distribution:

负二项同上，计数测度为 N，$h(x) = C_{r+x-1}^{r-1}$

2.7 Multinomial Distribution:

多项分布服从 $Multi(n, p_0, p_1, \cdots, p_k) : p_j \in (0,1), \sum_{j=0}^{k}$, 其离散测度：

$$\{(x_0, x_1, \cdots, x_k) : x_j \in N, \sum_{j=0}^{k} x_j = n\}$$

上有：

$$f_p(x) = \exp\{\sum_{j=1}^{k} x_j \log(\frac{p_j}{p_0}) + n\log(p_0)\}\frac{n!}{x_0!x_1!\cdots x_k!}$$

2.8 Gamma Distribution:

伽马分布 $Gamma(k, \theta) : k > 0, \theta > 0$，对于 R_+ 上的勒贝格测度有：

$$f_{k,\theta}(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta} = \exp\{(k-1)\log(x) - x/\theta - k\log(\theta) - \log\Gamma(k)\}$$

2.9 Exponential Distribution:

指数分布 $\exp(\lambda) = Gamma(1, \lambda^{-1})$：

$$f_\lambda = \exp\{\log\lambda - \lambda x\}$$

2.10 χ^2 Distribution:

$\chi_d^2 = Gamma(d/2, 2)$:

$$\begin{aligned} \chi_d^2 &\propto x^{d/2-1}\exp\{-x/2\} \\ &= \exp\{(\frac{d}{2} - 1)\log x - \frac{d}{2}\} \end{aligned} \tag{3}$$

3 Some Properties of Exponential Families

定义指数族分布的一个原因，就是指数族分布常常有很好的性质。从前面的分析可以看出来，一些平常我们使用的分布，都可以表示成指数族分布

的形式。对于优化的 Bayes Proceudre(我不知道怎么翻译，可能是贝叶斯决策吧，详见MIT stat655，大概讲的就是贝叶斯决策问题跟平方损失)，有着很好的鲁棒性。更多的，他们的分析性质也很好。比如，指数族分布中对数配分的图形常常使得指数族分布的估计更加简单。对于样本来说，假设我们要估计参数 θ，使用极大似然法（也就是最小化对数损失），那么根据指数族分布的表达式，我们可以写成：

$$\begin{aligned} &\min_{\theta} \quad \sum_{i=1}^{n} \log \frac{1}{p_{\theta}(X_i)} \\ &= \min \quad \sum_{i=1}^{n} [- < \theta, \phi(X_i) > + A(\theta)] \end{aligned} \tag{4}$$

这说明，在这个最小化中是不存在局部最小的，也就是在进行梯度运算的时候，不会陷入局部最小出不来。

当然，指数族分布也有很好的统计性质。对于来自指数族分布的样本，希望估计参数，比如 θ，此时 $\sum_{i=1}^{n} T(X_i)$ 就是充分统计量，其中 $T(x)$ 来自 $f_{\theta}(x) = exp(\xi(\theta)^T T(x) - B(\theta))h(x)$。这个看起来有点像因子分解定理，实际上的确用到了因子分解定理来证明。

除此以外，广义线性模型与指数族分布联系也非常密切。指数族分布就常常可以用来“建造”GLM。

4 参考资料

- Stanford Lecture Notes for stat311/ee377, John Duchi, March 13, 2019
- 知乎用户 @ 纯粹的专栏文章：指数族分布
- Stanford Lecture Notes for CS229, Part5 Generalized Linear Models,Andrew NG