指数族分布

指数族是一类分布，包括高斯分布、伯努利分布、二项分布、泊松分布、Beta 分布、Dirichlet 分布、Gamma 分布等一系列分布。指数族分布可以写为统一的形式：

$$p(x|\eta)=h(x)\exp(\eta^T\phi(x)-A(\eta))=\frac{1}{\exp(A(\eta))}h(x)\exp(\eta^T\phi(x)) \tag{1}$$

其中，η 是参数向量，$A(\eta)$ 是对数配分函数（归一化因子）。

在这个式子中，$\phi(x)$ 叫做充分统计量，包含样本集合所有的信息，例如高斯分布中的均值和方差。充分统计量在在线学习中有应用，对于一个数据集，只需要记录样本的充分统计量即可。

对于一个模型分布假设（似然），那么我们在求解中，常常需要寻找一个共轭先验，使得先验与后验的形式相同，例如选取似然是二项分布，可取先验是 Beta 分布，那么后验也是 Beta 分布。指数族分布常常具有共轭的性质，于是我们在模型选择以及推断具有很大的便利。

共轭先验的性质便于计算，同时，指数族分布满足最大熵的思想（无信息先验），也就是说对于经验分布利用最大熵原理导出的分布就是指数族分布。

观察到指数族分布的表达式类似线性模型，事实上，指数族分布很自然地导出广义线性模型：

$$\begin{gathered}y=f(w^Tx)\\ y|x\sim ExpFamily\end{gathered} \tag{2}$$

在更复杂的概率图模型中，例如在无向图模型中如受限玻尔兹曼机中，指数族分布也扮演着重要作用。

在推断的算法中，例如变分推断中，指数族分布也会大大简化计算。

一维高斯分布

一维高斯分布可以写成：

$$p(x|\theta)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{3}$$

将这个式子改写：

$$\begin{aligned}&\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(x^2-2\mu x+\mu^2))\\ &=\exp(\log(2\pi\sigma^2)^{-1/2})\exp(-\frac{1}{2\sigma^2}\begin{pmatrix}-2\mu & 1\end{pmatrix}\begin{pmatrix}x\\ x^2\end{pmatrix}-\frac{\mu^2}{2\sigma^2})\end{aligned} \tag{4}$$

所以：

$$\eta=\begin{pmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{pmatrix}=\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} \tag{5}$$

于是 $A(\eta)$：

$$A(\eta) = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log(-\frac{\pi}{\eta_2}) \tag{6}$$

充分统计量和对数配分函数的关系

对概率密度函数求积分：

$$\exp(A(\eta)) = \int h(x)\exp(\eta^T\phi(x))dx \tag{7}$$

两边对参数求导：

$$\begin{aligned}\exp(A(\eta))A'(\eta) &= \int h(x)\exp(\eta^T\phi(x))\phi(x)dx \\ \Longrightarrow A'(\eta) &= \mathbb{E}_{p(x|\eta)}[\phi(x)]\end{aligned} \tag{8}$$

类似的：

$$A''(\eta) = Var_{p(x|\eta)}[\phi(x)] \tag{9}$$

由于方差为正，于是 $A(\eta)$ 一定是凸函数。

充分统计量和极大似然估计

对于独立全同采样得到的数据集 $\mathcal{D} = \{x_1, x_2, \cdots, x_N\}$。

$$\begin{aligned}\eta_{MLE} &= \underset{\eta}{argmax}\sum_{i=1}^N \log p(x_i|\eta) \\ &= \underset{\eta}{argmax}\sum_{i=1}^N(\eta^T\phi(x_i) - A(\eta)) \\ &\Longrightarrow A'(\eta_{MLE}) = \frac{1}{N}\sum_{i=1}^N\phi(x_i)\end{aligned} \tag{10}$$

由此可以看到，为了估算参数，只需要知道充分统计量就可以了。

最大熵

信息熵记为：

$$Entropy = \int -p(x)\log(p(x))dx \tag{11}$$

> 一般地，对于完全随机的变量（等可能），信息熵最大。
>
> 我们的假设为最大熵原则，假设数据是离散分布的，k 个特征的概率分别为 p_k，最大熵原理可以表述为：
>
> $$\max\{H(p)\} = \min\{\sum_{k=1}^K p_k\log p_k\}\ s.t.\ \sum_{k=1}^K p_k = 1 \tag{12}$$

> 利用 Lagrange 乘子法：
>
> $$L(p,\lambda)=\sum_{k=1}^{K}p_k\log p_k+\lambda(1-\sum_{k=1}^{K}p_k) \tag{13}$$
>
> 于是可得：
>
> $$p_1=p_2=\cdots=p_K=\frac{1}{K} \tag{14}$$
>
> 因此等可能的情况熵最大。

一个数据集 $\mathcal{D}$，在这个数据集上的经验分布为 $\hat{p}(x)=\frac{Count(x)}{N}$，实际不可能满足所有的经验概率相同，于是在上面的最大熵原理中还需要加入这个经验分布的约束。

对任意一个函数，经验分布的经验期望可以求得为：

$$\mathbb{E}_{\hat{p}}[f(x)]=\Delta \tag{15}$$

于是：

$$\max\{H(p)\}=\min\{\sum_{k=1}^{N}p_k\log p_k\}\ s.t.\ \sum_{k=1}^{N}p_k=1,\mathbb{E}_p[f(x)]=\Delta \tag{16}$$

Lagrange 函数为：

$$L(p,\lambda_0,\lambda)=\sum_{k=1}^{N}p_k\log p_k+\lambda_0(1-\sum_{k=1}^{N}p_k)+\lambda^T(\Delta-\mathbb{E}_p[f(x)]) \tag{17}$$

求导得到：

$$\begin{aligned}\frac{\partial}{\partial p(x)}L&=\sum_{k=1}^{N}(\log p(x)+1)-\sum_{k=1}^{N}\lambda_0-\sum_{k=1}^{N}\lambda^Tf(x)\\&\Longrightarrow\sum_{k=1}^{N}\log p(x)+1-\lambda_0-\lambda^Tf(x)=0\end{aligned} \tag{18}$$

由于数据集是任意的，对数据集求和也意味着求和项里面的每一项都是0：

$$p(x)=\exp(\lambda^Tf(x)+\lambda_0-1) \tag{19}$$

这就是指数族分布。