Exponential Family Distribution 04 Maximum Entropy

Chen Gong

26 October 2019

从这节开始，我们将从最大熵的角度来解析指数族分布。首先，我们需要定义一下什么是熵？所谓熵，就是用来衡量信息反映的信息量的多少的单位。这里我们首先介绍一下，什么是熵？

1 最大熵原理

假设 p 是一个分布，所谓信息量就是分布的对数的相反数 (p 是小于 1 的，为了使信息量的值大于 0)，即为 $-\log p$。而熵则被我们定义为:

$$\begin{aligned} \mathbb{E}_{x\sim p(x)}[-\log p(x)] &= \int_x -p(x)\log p(x)dx \\ &= -\sum_x p(x)\log p(x) \end{aligned} \tag{1}$$

而最大熵原理实际上就可以定义为等可能。这是一种确定无信息先验分布的方法，它的原理就是是所有的可能都尽可能的出现，而不会出现类似于偏见的情况。接下来，我们令

$$H(x) = -\sum_x p(x)\log p(x) \tag{2}$$

假设 x 是离散的，

x	1	2	$\cdots$	k
p	p_1	p_2	$\cdots$	p_k

表 1: 随机变量 x 的概率密度分布情况

并且，需要满足约束条件，

$$s.t. \qquad \sum_{i=1}^N p_i = 1 \tag{3}$$

那么，总结一下上述的描述，优化问题可以写为:

$$\left\{ \begin{array}{ll} \mathrm{argmax} - \sum_x p(x)\log p(x) \\ s.t. \qquad \sum_{i=1}^N p_i = 1 \end{array} \right. \tag{4}$$

可以将其改写为:

$$\left\{ \begin{array}{ll} \mathrm{argmin} \sum_x p(x)\log p(x) \\ s.t. \qquad \sum_{i=1}^N p_i = 1 \end{array} \right. \tag{5}$$

实际上也就是求 $\hat{p}_i = \text{argmin} -H(p(x))$，其中 $p = \begin{pmatrix} p_1 & p_2 & \cdots & p_k \end{pmatrix}^T$。我们使用拉格朗日乘子法来求带约束的方程的极值。定义损失函数为：

$$\mathcal{L}(p,\lambda) = \sum_{i=1}^{N} p(x_i)\log p(x_i) + \lambda(1 - \sum_{i=1}^{k} p_i) \tag{6}$$

下面是对 $\hat{p}_i$ 的求解过程，

$$\frac{\partial \mathcal{L}}{\partial p_i} = \log p_i + p_i \frac{1}{p_i} - \lambda = 0 \tag{7}$$

解得：

$$p_i = exp(\lambda - 1) \tag{8}$$

又因为 λ 是一个常数，所以 $\hat{p}_i$ 是一个常数，那么我们可以轻易得到

$$\hat{p}_1 = \hat{p}_2 = \hat{p}_3 = \cdots = \hat{p}_k = \frac{1}{k} \tag{9}$$

很显然 $p(x)$ 是一个均匀分布，**那么关于离散变量的无信息先验的最大熵分布就是均匀分布**。

2 指数族分布的最大熵原理

我们首先写出指数族分布的形式：

$$p(x|\eta) = h(x)exp\left\{\eta^T \varphi(x) - A(\eta)\right\} \tag{10}$$

我们可以换一种形式来定义，为了方便之后的计算：

$$p(x|\eta) = \frac{1}{Z(\eta)} h(x)exp\left\{\eta^T \varphi(x)\right\} \tag{11}$$

但是，我们用最大熵原理来求指数族分布的时候，还差一个很重要的东西，也就是经验约束。也就是我们的分布要满足既定的事实上基础上进行运算。那么，我们需要怎么找到这个既定事实的分布呢？假设我们有一个数据集 $Data = \{x_1, x_2, x_3, \cdots, x_N\}$。那么，我们定义分布为，

$$\hat{p}(X = x) = \hat{p}(x) = \frac{Count(x)}{N} \tag{12}$$

那么我们可以得到一系列的统计量 $\mathbb{E}_{\hat{p}}(x)$，$Var_{\hat{p}}(x)$，$\cdots$。那么假设，$f(x)$ 是关于任意 x 的函数向量。那么我们定义 $f(x)$ 为：

$$f(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_Q(x) \end{pmatrix} \qquad \Delta = \begin{pmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \\ \Delta_Q \end{pmatrix} \tag{13}$$

其中，假设 $\mathbb{E}_{\hat{p}}[f(x)] = \Delta$(已知)。同样，我们将熵表达出来，

$$H[p] = -\sum_{x} p(x)\log p(x) \tag{14}$$

那么，这个优化问题，可以被我们定义为:

$$\begin{cases} \text{argmin} \sum_x p(x) \log p(x) \\ s.t. \qquad \sum_{i=1}^N p_i = 1 \\ \qquad\qquad \mathbb{E}_p[f(x)] = \mathbb{E}_{\hat{p}}[f(x)] = \Delta \end{cases} \tag{15}$$

其中，我们期望在总体数据上的特征和在给定数据上的特征一致。同样，我们使用拉格朗日乘子法来求带约束的方程的极值。定义损失函数为：

$$\mathcal{L}(p, \lambda_0, \lambda) = \sum_{i=1}^N p(x_i) \log p(x_i) + \lambda_0 (1 - \sum_x p) + \lambda^T (\Delta - \mathbb{E}_p[f(x)]) \tag{16}$$

将 $\mathbb{E}_p[f(x)])$ 进行改写为：

$$\mathcal{L}(p, \lambda_0, \lambda) = \sum_{i=1}^N p(x_i) \log p(x_i) + \lambda_0 (1 - \sum_x p) + \lambda^T (\Delta - \sum_x p(x) f(x)) \tag{17}$$

我们的目的是求一个 $\hat{p}(x)$，那么使用求偏导的方法（关于一个给定的 x，对于 $p(x)$ 求偏导）：

$$\frac{\mathcal{L}(p, \lambda_0, \lambda)}{p(x)} = \left(\log p(x) + p(x) \frac{1}{p(x)} \right) - \lambda_0 + \lambda^T f(x) = 0 \tag{18}$$

$$\log p(x) + 1 - \lambda_0 - \lambda^T f(x) = 0 \tag{19}$$

$$\log p(x) = \lambda_0 - 1 + \lambda^T f(x) \tag{20}$$

$$p(x) = exp \left\{ \lambda_0 - 1 + \lambda^T f(x) \right\} \tag{21}$$

整理一下即可得到 $p(x) = exp \left\{ \lambda^T f(x) - (1 - \lambda_0) \right\}$，那么我们可以将 $\eta = \begin{pmatrix} \lambda_0 \\ \lambda \end{pmatrix}$，$f(x) = \varphi(x)$，$(1 - \lambda_0) = A(\eta)$。很显然，$p(x)$ 是一个指数族分布。**那么我们可以得到一个结论，一个无先验信息先验的分布的最大熵分布是一个指数族分布**。