Exponential Family Distribution 01 Introduction

Chen Gong

23 October 2019

本节主要对指数族分布的概念和性质的一个小小的总结。指数族分布是一个广泛存在于机器学习研究中的分布。包括，Guassian 分布，Bernoulli 分布 (类别分布)，二项分布 (多项式分布)，泊松分布，Beta 分布，Dirichlet 分布，Gamma 分布和 Gibbs 分布等。

1 指数族分布的基本形式

指数族分布的基本形式可以表示为：

$$p(x|y) = h(x)exp\left\{\eta^T \varphi(x) - A(\eta)\right\} \tag{1}$$

η：参数向量，$\eta \in \mathbb{R}^p$。

$A(\eta)$：log partition function (对数配分函数)。

$h(x)$：这个函数只和 x 有关系，所以并不是很重要。

η 和 $h(x)$ 的理解比较简单，但是 log partition function 的理解难度比较大。所以，在这里对此函数做出一定的解释。

1.1 log partition function (对数配分函数)

什么是配分函数呢? 我的理解这是一个归一化的函数因子，用来使概率密度函数的积分值为 1。推导过程如下：

$$\begin{aligned} p(x|\theta) =& \frac{\hat{p}(x|\theta)}{z} \\ \int p(x|\theta)dx =& \int \frac{\hat{p}(x|\theta)}{z}dx = 1 \\ z =& \int \hat{p}(x|\theta)dx \end{aligned} \tag{2}$$

而在指数族函数中有关于 $A(\eta)$ 的配分函数的推导如下：

$$
\begin{aligned}
p(x|\eta) =& h(x)exp\{\eta^T\varphi(x)\}exp\{-A(\eta)\} \\
=& \frac{1}{exp\{A(\eta)\}}h(x)exp\{\eta^T\varphi(x)\} \\
\int p(x|\eta)dx =& \int \frac{1}{exp\{A(\eta)\}}h(x)exp\{\eta^T\varphi(x)\}dx = 1 \\
exp\{A(\eta)\} =& \int h(x)exp\{\eta^T\varphi(x)\}dx \\
A(\eta) =& \log \int h(x)exp\{\eta^T\varphi(x)\}dx
\end{aligned}
\tag{3}
$$

所以，$A(\eta)$ 被称为带有的 log 的 Partition Function。

2 指数族分布的相关知识

和指数族分布的相关知识，可以用下面这张图表来进行概况。

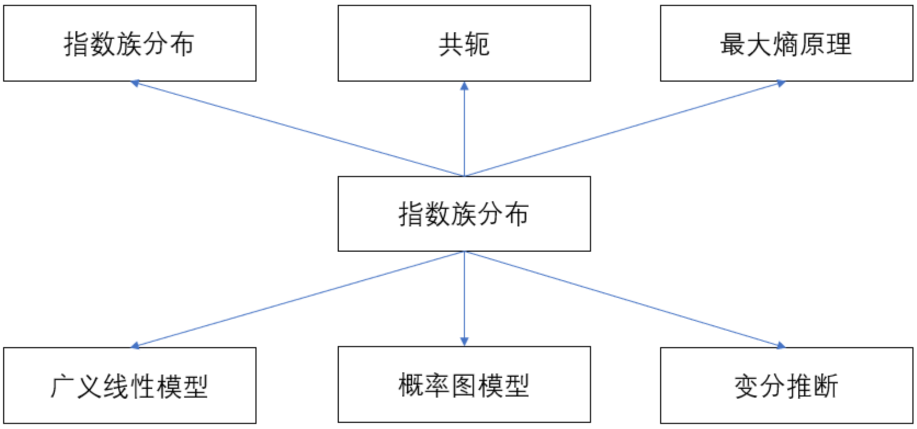

图 1: 指数族分布相关知识表示图

2.1 充分统计量

什么是充分统计量? 我自己的理解，充分统计量是一个有关于样本的函数，有了这个统计量就可以完整的表示出数据集整体的特征。从某种意义上说，我们就可以丢弃样本数据集了。下面对 Guassian Distribution 进行举例，数据集 Data set 为：$\{x_1, x_2, x_3, \cdots, x_N\}$

我们只需要一组充分统计量：

$$
\varphi(x) = \begin{pmatrix} \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i^2 \end{pmatrix} \tag{4}
$$

就可以反映出 Guassian 的所有特征 $\theta = (\mu, \Sigma)$。充分统计量在 online learning 中的使用有很大的作用。这样可以不记录那么多的数据集，只使用少量的数据就可以估计得到数据集整体的特征，可以用来简化计算。

2.2 共轭分布

为什么要使用共轭的概念呢？首先来看看贝叶斯公式：

$$p(z|x) = \frac{p(x|z)p(z)}{\int_z p(x|z)p(z)dz} \tag{5}$$

在这个公式中，$p(z|x)$ 为后验概率分布，$p(x|z)$ 为似然函数，$p(z)$ 为先验分布。在求解 $\int_z p(x|z)p(z)dz$ 时，计算难度是非常大的。或者说很多时候，根本算不出来。而且，换句话说，就算我们求得了 $p(z|x)$，也有可能因为 $p(z|x)$ 的形式过于复杂，导致 $\mathbb{E}_{p(z|x)}[f(x)]$ 根本算不出来。所以，为了解决这个问题，科研人员们想了很多的办法。近似推断的方法，比如，变分和采样。

变分的方法，是用简单的分布来拟合一个很难计算的分布，从而计算得出 $p(z|x)$ 的近似分布形式。而采样的方法，比如蒙特卡罗采样，隐马尔可夫蒙特卡罗采样 (MCMC) 等，是直接来求 $\mathbb{E}_{p(z|x)}[f(x)]$，这样直接跳过了中间那一堆的过程，在强化学习中经常使用。

而共轭是一种很取巧的方法，**它的效果是使先验和后验有着相同的分布形式，只是参数不同**。这样可以大大的简化计算，解决上述的问题。举例，

$$p(z|x) \propto p(x|z)p(z) \tag{6}$$

如果，$p(x|z)$ 为二项分布，$p(z)$ 为 Beta 分布，那么后验分布 $p(z|x)$ 也为 Beta 分布。

2.3 最大熵原理

下面列举几种确定先验 (prior distribution) 的方法，

1. 共轭，主要是为了计算的简单；
2. 最大熵方法，主要是为了解决无信息先验问题；
3. Jerrif。

最大熵原理会在后面的小节做详细的描述，主要思想就是 “等可能”。也就是尽量使所有的结论等可能的出现，来增加不确定性，保证每一项都是公平的。

2.4 广义线性模型

广义线性模型包括：

1. 线性组合，比如，w^Tx；
2. link function，也就是激活函数的反函数；
3. 指数族分布，$y|x \sim$ 指数族分布，包括：
   (a) 线性回归，在我们的线性回归模型中，我们曾定义过假设噪声符合 Guassian Distribution，那么 $y|x \sim \mathcal{N}(\mu, \Sigma)$；
   (b) 二分类问题：
      i. $y|x \sim$Bernoulli 分布；
      ii. $y|x \sim$Possion 分布；

2.5 概率图模型和变分推断

包括无向图等，有玻尔兹曼滤波器等。后续的章节会进行详细的描述。变分推断也在后续的章节有详细的描述。

Exponential Family Distribution 02 Example

Chen Gong

23 October 2019

本节的主要内容是演示 Guassian Distribution 的指数族表达形式，将高斯函数的形式转换为指数族分布的通用表达形式。

指数族分布的基本形式可以表示为：

$$p(x|y) = h(x)exp\left\{\eta^T\varphi(x) - A(\eta)\right\} \tag{1}$$

η：参数向量 parameter，$\eta \in \mathbb{R}^p$。

$A(\eta)$：log partition function (配分函数)。

$\varphi(x)$：充分统计量 sufficient statistics magnitude。

1 思路分析

高斯分布的概率密度函数可表示为：

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

观察指数族分布的表达形式，高斯分布的参数向量是有关于 $\theta = (\mu, \sigma)$ 的。首先观察指数部分的第一部分 $\eta^T\varphi(x)$，只有这个部分和 x 相关。那么把这个部分搞定，系数就是参数矩阵，剩下的就是配分函数了，而且配分函数是一个关于 η 的函数。

2 将 Guassian Distribution 改写为指数族分布的形式

具体推导过程如下所示：

$$\begin{aligned}
p(x|\theta) =& \frac{1}{\sqrt{2\pi}\sigma}exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} & (3)\\
=& \frac{1}{\sqrt{2\pi}\sigma}exp\left\{-\frac{1}{2\sigma^2}(x^2-2\mu x+\mu^2)\right\} & (4)\\
=& \frac{1}{\sqrt{2\pi}\sigma}exp\left\{-\frac{x^2}{2\sigma^2}+\frac{\mu x}{\sigma^2}-\frac{\mu^2}{2\sigma^2}\right\} & (5)\\
=& \frac{1}{\sqrt{2\pi}\sigma}exp\left\{\begin{pmatrix}\frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2}\end{pmatrix}\begin{pmatrix}x & x^2\end{pmatrix}-\frac{\mu^2}{2\sigma^2}\right\} & (6)\\
=& exp\log\frac{1}{\sqrt{2\pi}\sigma}exp\left\{\begin{pmatrix}\frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2}\end{pmatrix}\begin{pmatrix}x & x^2\end{pmatrix}-\frac{\mu^2}{2\sigma^2}\right\} & (7)
\end{aligned}$$

$$=exp\left\{\begin{pmatrix}\frac{\mu}{\sigma^2}\\-\frac{1}{2\sigma^2}\end{pmatrix}\begin{pmatrix}x & x^2\end{pmatrix}-\left(\frac{\mu^2}{2\sigma^2}-\frac{1}{2}\log 2\pi\sigma\right)\right\} \tag{8}$$

令:

$$\eta=\begin{pmatrix}\eta_1\\\eta_2\end{pmatrix}=\begin{pmatrix}\frac{\mu}{\sigma^2}\\-\frac{1}{2\sigma^2}\end{pmatrix}\Longrightarrow\begin{cases}\eta_1=\frac{\mu}{\sigma^2}\\\eta_2=-\frac{1}{2\sigma^2}\end{cases}\Longrightarrow\begin{cases}\mu=-\frac{\eta_1}{2\eta_2}\\\sigma^2=-\frac{1}{2\eta_2}\end{cases} \tag{9}$$

到了现在，我们离最终的胜利只差一步了，

$$\eta=\begin{pmatrix}\eta_1\\\eta_2\end{pmatrix}\quad\varphi(x)=\begin{pmatrix}x\\x^2\end{pmatrix} \tag{10}$$

$$A(\eta)=-\frac{\eta_1^2}{4\eta_2}+\frac{1}{2}\log(2\pi\cdot-\frac{1}{2\eta_2})=-\frac{\eta_1^2}{4\eta_2}+\frac{1}{2}\log(-\frac{\pi}{\eta_2}) \tag{11}$$

于是，Guassian Distribution 成功的被我们化成了指数族分布的形式 $exp\left\{\eta^T\varphi(x)-A(\eta)\right\}$。

Exponential Family Distribution 03 Property

Chen Gong

24 October 2019

本小节主要介绍 Exponential Distribution 中对数配分函数和充分统计量，还有极大似然估计和充分统计量的关系。

指数族分布的基本形式可以表示为：

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

$$p(x|\eta) = \frac{1}{exp\{A(\eta)\}}h(x)exp\left\{\eta^T\varphi(x)\right\} \tag{2}$$

1 对数配分函数和充分统计量

现在有一个问题，那就是我们如何求得对数配分函数 $exp\{A(\eta)\}$，或者说我们可不可以简单的求得对数配分函数。于是，就可以很自然的想到，前面所提到的充分统计量 $\varphi(x)$ 的概念。对数配分函数的目的是为了归一化，那么我们很自然的求出对数配分函数的解析表达式：

$$\begin{aligned}
\int p(x|\eta)dx &= \int \frac{1}{exp\{A(\eta)\}}h(x)exp\left\{\eta^T\varphi(x)\right\}dx \\
\int p(x|\eta)dx &= \frac{\int h(x)exp\left\{\eta^T\varphi(x)\right\}dx}{exp\{A(\eta)\}} = 1 \\
exp\{A(\eta)\} &= \int h(x)exp\left\{\eta^T\varphi(x)\right\}dx
\end{aligned} \tag{3}$$

下一步则是在 $exp\{A(\eta)\}$ 中对 η 进行求导。

$$\begin{aligned}
\frac{\partial exp\{A(\eta)\}}{\partial \eta} &= \nabla_\eta A(\eta) exp\{A(\eta)\} \\
&= \frac{\partial}{\partial \eta}\int h(x)exp\left\{\eta^T\varphi(x)\right\}dx \\
&= \int \frac{\partial}{\partial \eta}h(x)exp\left\{\eta^T\varphi(x)\right\}dx \\
&= \int h(x)exp\left\{\eta^T\varphi(x)\right\}\varphi(x)dx
\end{aligned} \tag{4}$$

将等式的左边的 $exp\{A(\eta)\}$ 移到等式的右边可得，

$$\nabla_\eta A(\eta) = \int h(x)exp\left\{\eta^T\varphi(x) - A(\eta)\right\}\varphi(x)dx \tag{5}$$

$$\nabla_\eta A(\eta) = \int p(x|\eta)\varphi(x)dx \tag{6}$$

$$\nabla_\eta A(\eta) = \mathbb{E}_{x\sim p(x|\eta)}[\varphi(x)] \tag{7}$$

其实通过同样的方法可以证明出，

$$\nabla\eta^2 A(\eta) = Var_{x\sim p(x|\eta)}[\varphi(x)] \tag{8}$$

又因为，协方差矩阵总是正定的矩阵，于是有 $\nabla_\eta^2 A(\eta) \succeq 0$。所以，由此得出 $A(\eta)$ 是一个凸函数。并且，由 $\mathbb{E}_{x\sim p(x|\eta)}[\varphi(x)]$ 和 $Var_{x\sim p(x|\eta)}[\varphi(x)]$ 就可以成功的求解得到 $A(\eta)$ 函数。那么我们做进一步思考，知道了 $\mathbb{E}[x]$ 和 $\mathbb{E}[x^2]$，我们就可以得到所有想要的信息。那么：

$$\mathbb{E}[\varphi(x)] = \begin{pmatrix} \mathbb{E}[x] \\ \mathbb{E}[x^2] \end{pmatrix} \tag{9}$$

2 极大似然估计和充分统计量

假设有一组观察到的数据集：$D = \{x_1, x_2, x_3, \cdots, x_N\}$，那么我们的求解目标为：

$$\begin{aligned}
\eta_{MLE} =& argmax \log \prod_{i=1}^N p(x_i|\eta) \\
=& argmax \sum_{i=1}^N \log p(x_i|\eta) \\
=& argmax \sum_{i=1}^N \log h(x_i) exp\left\{\eta^T\varphi(x_i) - A(\eta)\right\} \\
=& argmax \sum_{i=1}^N \log h(x_i) + \sum_i^N \left(\eta^T\varphi(x_i) - A(\eta)\right)
\end{aligned} \tag{10}$$

$$\frac{\partial}{\partial\eta}\left\{\sum_{i=1}^N \log h(x_i) + \sum_{i=1}^N \left(\eta^T\varphi(x_i) - A(\eta)\right)\right\} = 0 \tag{11}$$

$$\sum_{i=1}^N \varphi(x_i) = N \cdot \nabla_\eta A(\eta) \tag{12}$$

$$\nabla_\eta A(\eta) = \frac{1}{N}\sum_{i=1}^N \varphi(x_i) \tag{13}$$

或者说，我们可以认为是：$\nabla_\eta A(\eta_{MLE}) = \frac{1}{N}\sum_{i=1}^N \varphi(x_i)$。并且，$\nabla_\eta A(\eta_{MLE})$ 是一个关于 η_{MLE} 的函数。那么反解，我们就可以得到 η_{MLE}。所以我们要求 η_{MLE}，我们只需要得到 $\frac{1}{N}\sum_{i=1}^N \varphi(x_i)$ 即可。所以，$\varphi(x)$ 为一个充分统计量。

3 总结

在本小节中，我们使用了极大似然估计和对数配分函数来推导了，充分统计量，这将帮助我们理解 Exponential Distribution 的性质。

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:

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

可以将其改写为:

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

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。**