

Exponential Family Distribution 01

Chen Gong

22 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：

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

η : 参数向量, $\eta \in \mathbb{R}^p$ 。

$A(\eta)$: log partition function (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 (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:

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

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。

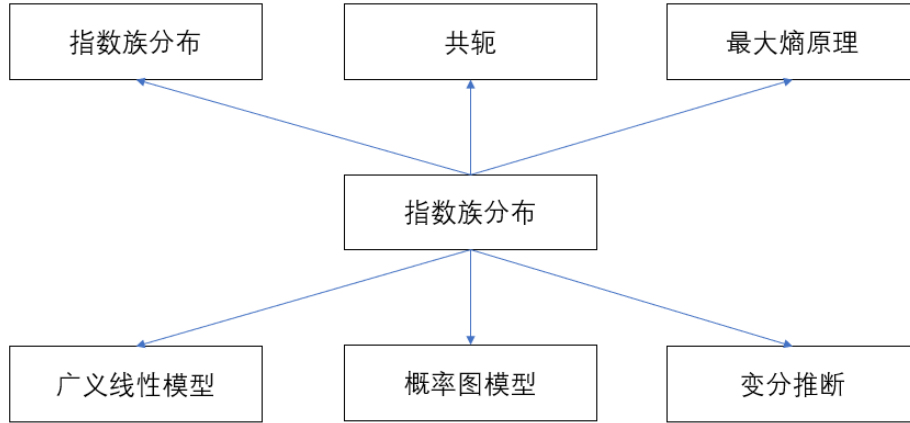


图 1: 指数族分布相关知识表示图

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 Gaussian Distribution 进行举例, 数据集 Data set 为: $\{x_1, x_2, x_3, \dots, x_N\}$

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 Gaussian 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

为什么要使用共轭的概念呢？首先来看看贝叶斯公式：

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

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。这样可以大大的简化计算, 解决上述的问题。举例,

$$p(z|x) \propto p(x|z)p(z) \quad (6)$$

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, $w^T x$;
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. $y|x \sim \text{Bernoulli}$ 分布;
 - ii. $y|x \sim \text{Poisson}$ 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。