机器学习白板推导系列笔记 (8.1~8.7)

Dexing Huang
dxhuang@bupt.edu.cn
Beijing University of Posts and Telecommunications

日期：2021 年 11 月 15 日

1 指数族分布背景介绍

指数族分布指的是概率分布可以写成以下形式的分布

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

其中 η 是参数向量, $x \in \mathbb{R}^p$, $\phi(x)$ 是充分统计量, $A(\eta)$ 是 log partition function, 具体含义下面进行解释.

一些常见的分布都属于指数族分布, 如 Gauss 分布, Bernoulli 分布, 二项分布, 泊松分布, Beta 分布, Gamma 分布等.

这里解释一下配分函数 (partition function), 可以将其理解为归一化因子, 通常需要求出数据的概率密度函数也就是

$$P(x|\theta) = \frac{1}{z}\hat{P}(x|\theta) \tag{2}$$

注. P 是真实分布, $\hat{P}$ 是估计值

其中 z 是归一化因子, 也称为配分函数.

$$z = \int_x \hat{P}(x|\theta) \tag{3}$$

$$\begin{aligned}
P(x|\eta) &= h(x)\exp(\eta^T\phi(x) - A(\eta)) \\
&= h(x)\exp(\eta^T\phi(x))\exp(-A(\eta)) \\
&= \underbrace{\frac{1}{\exp(A(\eta))}}_{\frac{1}{z}}\underbrace{h(x)\exp(\eta^T\phi(x))}_{\hat{P}(x|\theta)}
\end{aligned} \tag{4}$$

因此 $A(\eta) = \log z$, 所以 $A(\eta)$ 也称为 log 配分函数. 下面介绍一下指数族分布的三个性质和三个应用.

- 三大性质　充分统计量, 共轭, 最大熵 (无信息先验)
- 三大应用　广义线性模型, 概率图模型, 变分推断

充分统计量指的是式子 (1) 中的 $\phi(x)$, 统计量指的是对样本的加工, 即对于一个样本的函数, 比如均值方差等, 充分统计量有什么好处呢, 可以想一下在机器学习中需要有数据集 X, 需要掌握数据集中每个元素的信息才能进行很好的学习, 但也存在这么一种情况, 如果样本的模型已知 (比如服从高斯分布), 那么在学习时不需要掌握每个样本的信息, 而是可以通过统计量就可以对问题实现建模了 (在高斯分布中之需要掌握样本的均值和方差即可).

共轭, 在贝叶斯派解决问题时往往建立后验概率的模型, 根据贝叶斯公式

$$P(z|x) = \frac{P(z)P(x|z)}{\int_x P(x|z)P(z)dz} \tag{5}$$

但是根据之前的描述我们知道该积分 $\int_x P(x|z)P(z)d$ 求解起来十分困难, 共轭就是一种解决积分求解困难的一种方法, 对于似然函数 $P(x|z)$ 来说, 如果能够找到一个先验 $P(z)$ 与其共轭, 那么后验 $P(z|x)$ 会有与先验相同的分布形式, 转而只需要求解后验分布的参数即可.

最大熵原理是指在给定一个限制条件的情况下, 对于未知的部分, 假设所有情况等可能发生, 这样熵越大, 系统的随机性越强.

广义线性模型是为了解决回归和分类问题, 在线性模型的基础上进行拓展有线性组合, link function(激活函数的反函数) 以及指数族分布等几种.

概率图模型无向图:RBM(限制玻尔兹曼机), 在之后的章节会涉及.

变分推断在指数族分布中占据重要的地位.

2 高斯分布的指数族形式

本节以一维高斯分布为例, 将其转化为指数族分布的形式 (式 (1))

$$\begin{aligned}
P(x|\theta) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\} \\
&= \exp(\log(2\pi\sigma^2)^{-\frac{1}{2}})\exp\{-\frac{1}{2\sigma^2}(x^2-2\mu x+\mu^2)\} \\
&= \exp\{-\frac{1}{2\sigma^2}x^2+\frac{1}{\sigma^2}\mu x-\frac{1}{2\sigma^2}\mu^2-\frac{1}{2}\log(2\pi\sigma^2)\} \\
&= \exp\{\underbrace{\begin{pmatrix}\frac{1}{\sigma^2}\mu & -\frac{1}{2\sigma^2}\end{pmatrix}}_{\eta^T}\underbrace{\begin{pmatrix}x \\ x^2\end{pmatrix}}_{\phi(x)}-\underbrace{(\frac{1}{2\sigma^2}\mu^2+\frac{1}{2}\log(2\pi\sigma^2))}_{A(\eta)}\}
\end{aligned} \tag{6}$$

因此

$$\eta = \begin{pmatrix}\eta_1 \\ \eta_2\end{pmatrix} = \begin{pmatrix}\frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2}\end{pmatrix} \tag{7}$$

那么 $\sigma^2 = -\frac{1}{2\eta_2}, \mu = -\frac{\eta_1}{2\eta_2}$, 带入 $A(\eta)$ 得

$$\begin{aligned}
A(\eta) &= \frac{1}{2\sigma^2}\mu^2+\frac{1}{2}\log(2\pi\sigma^2) \\
&= -\frac{\eta_1^2}{4\eta_2^2}+\frac{1}{2}\log(-\frac{\pi}{\eta_2})
\end{aligned} \tag{8}$$

那么就把高斯分布化成了指数族分布的表达形式.

3 log 配分函数 $A(\eta)$ 与充分统计 $\phi(x)$ 的关系

再次回顾一下指数族函数的表达形式

$$P(x|\eta) = h(x)\exp(\eta^T\phi(x)-A(\eta)) \tag{9}$$

进行适当的变换

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

易得

$$\exp(A(\eta)) = \int_x h(x)\exp(\eta^T\phi(x))dx \tag{10}$$

上式两边分别对 η 求导

$$\begin{aligned}\exp(A(\eta))A'(\eta) &= \frac{\partial}{\partial\eta}\int_x h(x)\exp(\eta^T\phi(x))dx \\ &= \int_x h(x)\frac{\partial}{\partial\eta}\exp(\eta^T\phi(x))dx \\ &= \int_x h(x)\exp(\eta^T\phi(x))\phi(x)dx\end{aligned}$$

将 $\exp(A(\eta))$ 移到右边

$$\begin{aligned}A'(\eta) &= \int_x \underbrace{h(x)\exp(\eta^T\phi(x) - A(\eta))}_{P(x|\eta)}\phi(x)dx \\ &= \mathbb{E}_{P(x|\eta)}[\phi(x)]\end{aligned} \tag{11}$$

对上式再求一次导

$$\begin{aligned}A''(\eta) &= \int_x h(x)\exp(\eta^T\phi(x) - A(\eta))\phi(x)(\phi(x) - A'(\eta))dx \\ &= \int_x h(x)\exp(\eta^T\phi(x) - A(\eta))\phi(x)^2dx - \int_x h(x)\exp(\eta^T\phi(x) - A(\eta))\phi(x)dx \cdot A'(\eta) \\ &= \mathbb{E}_{P(x|\eta)}[\phi^2(x)] - [\mathbb{E}_{P(x|\eta)}[\phi(x)]]^2 \\ &= \mathrm{Var}_{P(x|\eta)}[\phi(x)]\end{aligned} \tag{12}$$

于是就得到了 log 配分函数 $A(\eta)$ 与充分统计 $\phi(x)$ 的关系. 从 $A(\eta)$ 的二阶导数等于方差可以得出其二阶导数大等 0, 那么可以得到$A(\eta)$ 是凸函数.

4 极大似然估计与充分统计 $\phi(x)$

使用极大似然估计对指数族分布的参数 η 进行估计, 假设数据集 $D = \{(x_i, y_i)\}_{i=1}^N$

$$\begin{aligned}\eta_{MLE} &= \arg\max \log P(D|\eta) \\ &= \arg\max \log \prod_{i=1}^N p(x_i|\eta) \\ &= \arg\max \sum_{i=1}^N \log p(x_i|\eta) \\ &= \arg\max \sum_{i=1}^N \log[h(x_i)\exp(\eta^T\phi(x_i) - A(\eta))] \\ &= \arg\max \sum_{i=1}^N [\eta^T\phi(x_i) - A(\eta))]\end{aligned}$$

对上述函数进行求导并令导数为 0

$$\frac{\partial}{\partial \eta}\sum_{i=1}^{N}[\eta^T\phi(x_i)-A(\eta))]=\sum_{i=1}^{N}[\phi(x_i)-A'(\eta))]=0$$

解得

$$A'(\eta_{MLE})=\frac{1}{N}\sum_{i=1}^{N}\phi(x_i) \tag{13}$$

那么 η_{MLE} 就可以根据 A' 的反函数求出, 在这里也能再次看出充分统计量 $(\phi(x))$ 的作用, 只要知道充分统计的参数即可求出模型参数 η, 而不用记住原始数据集 D.

5 最大熵原理与指数族分布

可以想一下这个问题, 在之前处理问题时, 一般都是对某一数据分布进行建模, 比如将噪音视为服从高斯分布等等, 相当于加入了先验知识, 但如果先验未知那该如何处理. 信息论为我们提供了一种思路, 那就是熵, 熵是对系统不确定性的度量, 最大熵原理认为在没有信息的情况下, 不确定的东西都认为是等可能的, 熵最大的模型就是最好的.

下面先来介绍一下熵

$$\begin{aligned}H(X)&=\mathbb{E}_{p(x)}[-\log p(x)]\\&=-\sum_{x}p(x)\log p(x)=-\int_{x}p(x)\log p(x)dx\end{aligned} \tag{14}$$

在信息论中学过, 离散随机变量的熵满足

$$0\le H(X)\le \log N \tag{15}$$

注. N 为样本个数

下面来进行简单的证明, N 个离散随机变量的分布 $P(X=x_i)=p_i,(i=1,2,\cdots,N)$, 根据最大熵原理得到约束最优化问题

$$\begin{cases}\arg\min\limits_{p_i}\sum\limits_{x}p_i\log p_i\\ s.t.\sum\limits_{i=1}^{N}p_i=1\end{cases} \tag{16}$$

构造 Lagrange 函数

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

求导并令其等 0, 得

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

得

$$p_i=\exp(\lambda-1) \tag{17}$$

可以看出 p_i 为常数, 因此

$$p_1=p_2=\cdots=p_N=\frac{1}{N} \tag{18}$$

因此, 关于离散变量的无先验分布是均匀分布.

下面讨论一下在先验未知的情况下的概率分布模型, 假设数据集 $D=\{(x_i,y_i)\}_{i=1}^N$, 在使用最大熵原理进行求解模型时, 需要满足一个最基本的约束, 即经验约束, 给定数据集得经验分布为

$$\hat{P}(X=k)=\frac{\sum_{i=1}^N \mathbb{I}(x_i=k)}{N} \tag{19}$$

那么可以根据经验分布得到统计量 $\mathbb{E}_{\hat{p}(x)}[x], \mathrm{Var}_{\hat{p}(x)}[x]$, 假设 $f(x)$ 是任意关于 x 的函数向量, 即

$$f(x)=\begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{pmatrix} \quad \Delta=\begin{pmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \\ \Delta_m \end{pmatrix} \tag{20}$$

设 $\mathbb{E}_{\hat{p}(x)}[f(x)]=\Delta$, 其中 Δ 是已知的, 当 $f(x)$ 是已知时, 假设真实分布为 p, 熵为

$$H(p)=-\sum_{i=1}^N p_i \log p_i \tag{21}$$

根据最大熵构造优化问题

$$\begin{cases} \arg\min\limits_{p_i} \sum\limits_x p_i \log p_i \\ s.t. \sum\limits_{i=1}^N p_i = 1 \\ \qquad \mathbb{E}_{\hat{p}(x)}[f(x)]=\mathbb{E}_{p(x)}[f(x)]=\Delta(\text{无偏估计}) \end{cases} \tag{22}$$

其 Lagrange 函数为

$$\begin{aligned} \mathcal{L}(p,\lambda_0,\lambda) &= \sum_{i=1}^N p_i \log p_i + \lambda_0(1-\sum_{i=1}^N p_i) + \lambda^T(\Delta - \mathbb{E}_{p(x)}[f(x)]) \\ &= \sum_{i=1}^N p_i \log p_i + \lambda_0(1-\sum_{i=1}^N p_i) + \lambda^T(\Delta - \sum_{i=1}^N p_i f(x_i)) \end{aligned} \tag{23}$$

最终目的是求出数据的分布 $p(x)$, 对 Lagrange 函数求偏导并令其为 0

$$\frac{\partial \mathcal{L}}{\partial p(x)} = \log p(x) + 1 - \lambda_0 - \lambda^T f(x) = 0$$

解得

$$p(x)=\exp(\underbrace{\lambda^T}_{\eta^T}\underbrace{f(x)}_{\phi(x)}-\underbrace{(1-\lambda_0)}_{A(\eta)}) \tag{24}$$

可以发现, 得到了一个指数族分布, 一个无先验分布的最大熵是一个指数族分布.