指数族分布

Exponential Family Distribution

① 背景

② 高斯分布

③ 对数配分函数

④ 极大似然估计

⑤ 最大熵角度

背景（Background）

高斯、伯努利、类别、二项、多项式、Beta、Dirichlet、Gamma

上述的分布都属于指数族分布.

指数族分布的标准形式：

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

其中，$x \in \mathbb{R}^p$

η：参数，是一个向量

$A(\eta)$：log partition function

配分函数.

配分函数：

一般地，$P(x|\theta) = \frac{1}{z}\hat{p}(x|\theta)$

其中 $\hat{p}(x|\theta)$ 是关于 x 的函数，其积分不为1.

所以为了满足概率密度的定义，加上一个归一化因子 z.

此时，$\underbrace{\int p(x|\theta)dx}_{1} = \int \frac{1}{z}\hat{p}(x|\theta)\,dx$

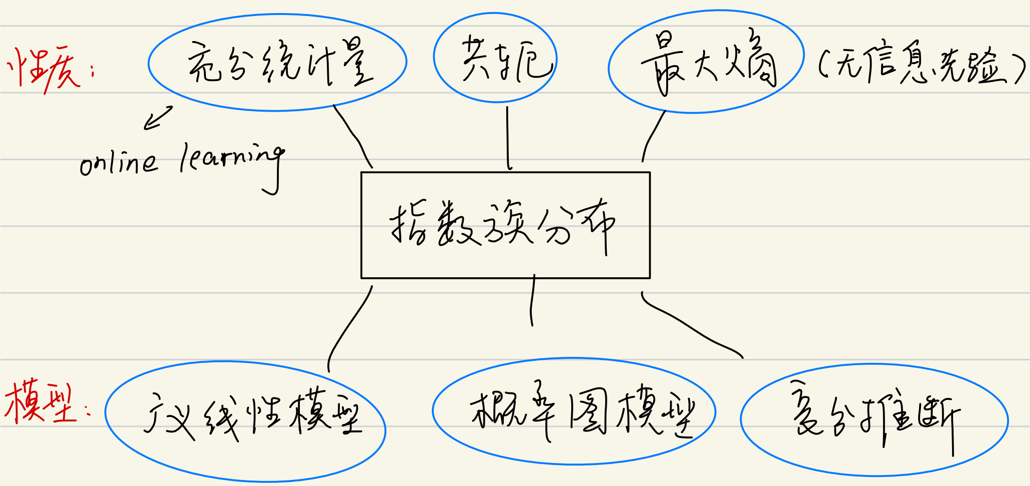
则 $z = \int \hat{p}(x|\theta)\,dx$. 称 z 为配分函数.

（有点类似于能量模型）

则对于指数族的标准形式：

$$p(x|\eta) = h(x)\exp\{\eta^T\phi(x) - A(\eta)\}$$
$$= h(x)\exp\{\eta^T\phi(x)\}\cdot\exp\{-A(\eta)\}$$
$$= \underbrace{\frac{1}{\exp\{A(\eta)\}}}_{Z}\, h(x)\exp\{\eta^T\phi(x)\}.$$

则 $Z = \exp\{A(\eta)\} \Rightarrow A(\eta) = \log Z$. 故称之为 log partition function



充分统计量：指的就是标准形式中的 $\phi(x)$.

充分统计量的作用是压缩数据，不需要储存样本，只需要储存充分统计量就可以表示分布.

共轭： $p(z|x) = \dfrac{p(x|z)p(z)}{\int_z p(x|z)p(z)dz}$.

往往上式这个后验是很难求出的。

原因在于 $\int_z p(x|z)p(z)dz$ 很难求。

主要是由于分布复杂或数据维度过高

所以需要用到近似的方法。

有变分推断、MCMC 等。

但是也可以用共轭的方法来使得后验可计算。

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

如果后验与先验具有相同的分布形式

那么称先验与似然是共轭的

例子

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

Beta 　 二项式 　 Beta.

最大熵（无信息先验）：熵越大，表示不确定性越大。

假设当不知道参数的信息时，所有可能发生的

情况都是等可能的，即熵最大。

确定先验分布的几种方法：

① 共轭 → 计算方便

② 最大熵 → 无信息先验

③ Jeffrey 先验 → 无信息先验

广义线性模型：三个最基本的概念：

① 线性组合 $w^T x$

② Link function 激活函数的反函数.

③ 指数族分布：$y|x$ 满足指数族分布.

例子：线性回归：$y|x \sim \mathcal{N}(\mu, \Sigma)$

分类：$y|x \sim$ 伯努利分布.

泊松回归：$y|x \sim$ 泊松分布

概率图模型：无向图：RBM（限制玻尔兹曼机）→ 需要满足指数族分布

变分推断：指数族分布可以简化变分推断的运算.

高斯分布 (Gaussian Distribution)

标准指数族分布形式：

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

以一维高斯分布举例，将其化为标准指数族分布的形式：

$\theta = (\mu, \sigma^2)$

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

$$= \exp\left\{\log(2\pi\sigma^2)^{-\frac{1}{2}}\right\}\cdot\exp\left\{-\frac{1}{2\sigma^2}(x^2-2\mu x+\mu^2)\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}(x^2-2\mu x) - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\}$$

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

则

$$\left\{\begin{aligned} & h(x) = 1 \\ & \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix} \Rightarrow \begin{cases} \mu = -\frac{\eta_1}{2\eta_2} \\ \sigma^2 = -\frac{1}{2\eta_2} \end{cases} \\ & \phi(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \\ & A(\eta) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) = \frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log\left(-\frac{\pi}{\eta_2}\right) \end{aligned}\right\} \Rightarrow h(x)\exp\{\eta^T\phi(x) - A(\eta)\}$$

对数配分函数
(Log Partition Function)

标准指数族分布形式.

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

两边同时对 x 求积分:

$$\int p(x|\eta)\,dx = \frac{1}{\exp\{A(\eta)\}}\int h(x)\exp\{\eta^T\phi(x)\}\,dx$$

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

两边同时对 η 求导:

$$\frac{\partial}{\partial\eta}\exp\{A(\eta)\} = \frac{\partial}{\partial\eta}\int h(x)\exp\{\eta^T\phi(x)\}\,dx$$

$$\exp\{A(\eta)\}\,A'(\eta) = \int h(x)\exp\{\eta^T\phi(x)\}\,\phi(x)\,dx$$

$$A'(\eta) = \int h(x)\exp\{\eta^T\phi(x) - A(\eta)\}\cdot\phi(x)\,dx.$$

$$A'(\eta) = E_{x\sim p(x|\eta)}[\phi(x)]$$

由上式，得：

$$\exp\{A(\eta)\} A'(\eta) = \int h(x) \exp\{\eta^T \phi(x)\} \phi(x) dx$$

两边对 η 求导：

$$\exp\{A(\eta)\}[A'(\eta)]^2 + \exp\{A(\eta)\} A''(\eta) = \int h(x) \exp\{\eta^T \phi(x)\} \phi^2(x) dx$$

$$[A'(\eta)]^2 + A''(\eta) = \int h(x) \exp\{\eta^T \phi(x) - A(\eta)\} \cdot \phi^2(x) dx$$

$$A''(\eta) = E_{x\sim p(x|\eta)}[\phi^2(x)] - \left(E_{x\sim p(x|\eta)}[\phi(x)]\right)^2$$

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

则

$$\boxed{\begin{aligned} A'(\eta) &= E_{x\sim p(x|\eta)}[\phi(x)] \\ A''(\eta) &= Var_{x\sim p(x|\eta)}[\phi(x)] \end{aligned}}$$

极大似然估计 (MLE)

上一节的推导是直接根据标准形式确定的，本节将使用样本来估计 η。

标准指数族分布的形式：

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

数据 $D = \{x_1, x_2, \dots, x_N\}$

$$\eta_{MLE} = \underset{\eta}{\arg\max} \log P(D|\eta)$$

$$= \underset{\eta}{\arg\max} \log \prod_{i=1}^{N} p(x_i|\eta)$$

$$= \underset{\eta}{\arg\max} \sum_{i=1}^{N} [\log h(x_i) + \eta^T \phi(x_i) - A(\eta)]$$

由于 $h(x_i)$ 与 η 无关，故舍去：

$$\eta_{MLE} = \underset{\eta}{\arg\max} \sum_{i=1}^{N} [\eta^T \phi(x_i) - A(\eta)]$$

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

$\Rightarrow A'(\eta_{MLE}) = \frac{1}{N} \sum_{i=1}^{N} \phi(x_i)$ ，记 $A^{(-1)'}(\cdot)$ 为 $A'(\cdot)$ 的反函数，

则 $\eta_{MLE} = A^{(-1)'}(\eta_{MLE})$

最大熵角度

(Max Entropy Perspective)

对未知概率分布的猜测

最大熵 $\Leftrightarrow$ 等可能.

信息量: $-\log p$.

熵: $E_{x\sim p(x)}[-\log p] = \int -p(x)\cdot\log p(x)\,dx$ 连续

记为 $H[P]$ $= -\sum_{x} p(x)\log p(x)$ 离散.

假设 x 是离散的. 分布列:

x	$1,\ 2,\ \cdots,\ K$
p	$p_1,\ p_2,\ \cdots,\ p_K$.

$\sum_{i=1}^{K} p_i = 1$

则最大熵问题就是一个优化问题

$$\begin{cases} \max\limits_{p} H[P] = \max -\sum_{i=1}^{K} p_i\log p_i \\ s.t.\ \sum_{i=1}^{K} p_i = 1 \end{cases}$$

$$\Rightarrow \begin{cases} \min \sum_{i=1}^{K} p_i\log p_i \\ s.t.\ \sum_{i=1}^{K} p_i = 1 \end{cases}$$

拉格朗日函数: $\mathcal{L}(P,\lambda) = \sum_{i=1}^{K} p_i\log p_i + \lambda(1-\sum_{i=1}^{K} p_i)$

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

可见 $p_1 = p_2 = \cdots = p_K$.

又因 $\sum_{i=1}^{K} p_i = 1$

则 $p_1 = p_2 = \cdots = p_K = \frac{1}{K}$ 均匀分布

在没有任何已知的情况下,
均匀分布的熵最大

$\boxed{\text{最大熵原理}}$

满足已知事实的最大熵，数据服从指数族分布.

已知事实：$Data = \{x_1, x_2, \dots, x_N\}$

$\boxed{\text{通过经验分布量化该已知事实}}$

经验分布：$\hat{p}(X=x) = \hat{p}(x) = \dfrac{\sum \mathbb{1}(x_i = x)}{N}$ $\quad \mathbb{1}(condition) = \begin{cases} 1, & condition = true \\ 0, & condition = false \end{cases}$

empirical distribution

则可根据该分布求得 $E[X]$, $Var[X]$ 等数字特征.

则定义 $E_{\hat{p}}[f(x)] = \Delta \rightarrow$ 已知.

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

x		
p_1	p_2	p_3

则根据最大熵理论，得到优化问题

$$\begin{cases} \min \sum_x p(x) \log p(x) \\ s.t. \quad \sum_x p(x) = 1 \\ \qquad E_p[f(X)] = E_{\hat{p}}[f(X)] = \Delta \end{cases}$$

Lagrange 函数：$\mathcal{L}(p, \lambda_0, \lambda) = \sum_x p(x) \log p(x) + \lambda_0 (1 - \sum_x p(x)) + \lambda^T (\Delta - E_p[f(x)])$

对 $p(x)$ 求偏导

$$\frac{\partial L}{\partial p(x)} = \sum_x (\log p(x)+1) - \sum_x \lambda_0 - \sum_x \lambda^T f(x)$$ 这一步需要仔细思考一下.

$$= \sum_x [\log p(x) + 1 - \lambda_0 - \lambda^T f(x)]$$

$$\triangleq 0$$

这里注意，假设将 $p(x)$ 看作是函数的，则对于 x 的每个取值对应的 p_i，其拉格朗日函数的偏导数都为 0（见该小节第一页），则

蓝字这里还需要具体验证一下，但不知道哪里有理论证明.

$$\log p(x) + 1 + \lambda_0 + \lambda^T f(x) = 0.$$

$$\Rightarrow p(x) = \exp\{\lambda^T f(x) - (\lambda_0+1)\}$$

令 $\eta = \begin{pmatrix}\lambda_0 \\ \lambda\end{pmatrix}$, $\phi(x) = \begin{pmatrix} 0 \\ f(x)\end{pmatrix}$ $A(\eta) = (1, 0)\cdot\eta + 1$, $h(x) = 1$

则 $p(x) = h(x)\exp\{\eta^T\phi(x) - A(\eta)\}$.

∴ 满足已知事实的最大熵所得到的数据分布就是指数族分布.

可以看一下最大熵模型

同时 PRML 中，还用指数族分布引出了 sigmoid 和 softmax，也可以看看