指数族分布

特征

1. 充分统计量 sufficient statistics

   $\phi(x)$ 对样本的函数（加工），譬如：均值，方差

$$\phi(x) = \begin{pmatrix} \sum_i^N x_i \\ \sum_{i=1}^N x_i^2 \end{pmatrix}$$

2. 共轭

$$p(Z|X) = \frac{p(X|Z)p(Z)}{\int_z p(X|Z)p(Z)dZ}$$

   后验积分难求。

$$E_{p(Z|X)}[f(Z)]$$

$$\begin{gathered} p(Z|X) \propto \ p(X|Z)|p(Z) \\ p(X|Z)\text{二项式分布} \\ p(Z)beta \\ p(Z|X)beta \end{gathered}$$

3. 最大熵（无信息先验）

   让熵最大，更加随机
4. 广义线性模型

   线性组合 $w^T x$

   link function 激活函数反函数

   指数族分布：y|x
5. 概率图模型

   无向图模型 RBM
6. 变分模型

   大大简化

> 1. 共轭先验，计算方便
> 2. 最大熵

> 3. Jeffreys prior

分类

1. Bernouli - 类别分布
2. 二项分布 - 多项式分布
3. 泊松分布
4. Beta
5. Dirichilet
6. Gamma
7. Gaussian

形式

$$P(x|\eta) = h(x)\exp\left(\eta^\top \phi(x) - A(\eta)\right)$$

η: parameter参数，向量， $x \in \mathbb{R}^p$

$\phi(x)$: 充分统计量 sufficient statistics

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

> $$\begin{gathered} P(x|\theta) = \frac{1}{Z}\hat{P}(x|\theta) \\ \frac{1}{Z} \Rightarrow \text{归一化因子} \\ Z = \int_x \hat{P}(x|\theta)dx \\ ------------- \\ \int P(x|\theta)dx = \int \frac{1}{Z}\hat{P}(x|\theta)dx \\ 1 = \frac{1}{Z}\int \hat{P}(x|\theta)dx \\ Z = \int \hat{P}(x|\theta)dx \end{gathered}$$

$$
\begin{aligned}
P(x|\eta) &= h(x)\cdot\exp\left(\eta^{\top}\phi(x)\right)\cdot\exp(-A(\eta)) \\
&= \frac{1}{\exp(A(x))}h(x)\cdot\exp\left(\eta^{\top}\phi(x)\right) \\
&= \frac{1}{Z}\hat{P}(x|\eta) \\
\exp(A(x)) &= Z \\
A(\eta) &= \log Z \Rightarrow Z \rightarrow Partition function
\end{aligned}
$$

高斯分布

$$
\begin{aligned}
P(x|\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(x^2-2\mu x+\mu^2\right)\right\} \\
&= \exp\log\left(2\pi\cdot\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(x^2-2\mu x\right)-\frac{\mu^2}{2\sigma^2}\right\} \\
&\left(-2\mu \quad 1\right)\begin{pmatrix} x \\ x^2 \end{pmatrix} \\
&= \exp\log\left(2\pi\cdot\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(-2\mu \quad 1\right)\begin{pmatrix} x \\ x^2 \end{pmatrix}-\frac{\mu^2}{2\sigma^2}\right\} \\
&= \exp\left\{\begin{pmatrix} \frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2} \end{pmatrix}\begin{pmatrix} x \\ x^2 \end{pmatrix}-\frac{\mu^2}{2\sigma^2}-\frac{1}{2}\log 2\pi\sigma^2\right\} \\
&= \exp\left\{\begin{pmatrix} \frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2} \end{pmatrix}\begin{pmatrix} x \\ x^2 \end{pmatrix}-\left(\frac{\mu^2}{2\sigma^2}+\frac{1}{2}\log 2\pi\sigma^2\right)\right\} \\
&\quad \eta^\top \cdot \quad \phi(x) \quad -A(\eta)
\end{aligned}
$$

$$- - - - - - -$$

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

$$\begin{cases} \eta_1 = \frac{\mu}{\sigma^2} \\ \eta_2 = -\frac{1}{2\sigma^2} \end{cases} \Rightarrow \begin{cases} \mu = -\frac{\eta_1}{2\eta_2} \\ \sigma^2 = -\frac{1}{2\eta_2} \end{cases}$$

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

$$
\begin{aligned}
A(n) &= -\frac{n_1^2}{4n_2}+\frac{1}{2}\log\left(2\pi-\frac{1}{2n_2}\right) \\
&= -\frac{n_1^2}{4n_2}+\frac{1}{2}\log\left(-\frac{\pi}{n_2}\right)
\end{aligned}
$$

$$- - - - -$$

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

$$\eta = \eta(\theta)$$

$$A(\eta) = A(\eta(\theta))$$

对数配分函数和充分统计量的关系

$$
\begin{aligned}
P(x|\eta) &= h(x)\exp\left(\eta^{\top}\phi(x)\cdot \exp(-A(\eta)\right) \\
&= \frac{1}{\exp(A(x))} h(x)\cdot \exp\left(\eta^{\top}\phi(x)\right) \\
\exp(A(x)) &= \int h(x)\cdot \exp\left(\eta^{\top}\phi(x)\right)dx \Rightarrow Z = \int \hat{P}(x|\theta)dx \\
\exp(A(\eta))\cdot A'(\eta) &= \frac{\partial}{\partial\eta}\left(\int h(x)\exp\left(\eta^{\top}\phi(x)\right)dx\right) \\
&= \int h(x)\exp\left(\eta^{\top}\phi(x)\right)\cdot\phi(x)dx \\
A'(\eta) &= \frac{\int h(x)\exp\left(\eta^{\top}\phi(x)\right)\cdot\phi(x)dx}{\exp(A(\eta))} \\
&= \int h(x)\exp\left(\eta^{T}\phi(x) - A(\eta)\right)\cdot\phi(x)dx \\
&= \int \quad P(x|\eta)\cdot\phi(x)dx \\
&= E_{P(x|\eta)}[\phi(x)]
\end{aligned}
$$

$$
\begin{aligned}
A'(\eta) &= E_{P(x|\eta)}[\phi(x)] \\
A''(\eta) &= \mathrm{Var}[\phi(x)]
\end{aligned}
$$

$A(\eta)$ is convex function

Take Gaussian distribution as example:

$$
\begin{aligned}
E[\phi(x)] &= \left(\frac{E[x]}{E\left[x^2\right]}\right) \\
E[x] &= \mu \\
A'(n) &= \frac{\partial A(\eta)}{\partial\eta_1}? = \mu \\
A'(\eta) &= -\frac{2\eta_1}{4\eta_2} = -\frac{\eta_1}{2\eta_2} \\
&= \frac{\frac{\mu}{\sigma^2}}{-2\cdot -\frac{1}{2\sigma^2}} = \mu
\end{aligned}
$$

极大似然估计和充分统计量的关系

$$
\begin{aligned}
D =& \{x_1, x_2, \cdots, x_N\} \\
\eta_{MLE} &= \arg\max \log P(D|\eta) \\
&= \arg\max \cdot \log \prod_{i=1}^{N} p\left(x_i|\eta\right) \\
&= \arg\max \sum_{i=1}^{N} \log p\left(x_i|\eta\right) \\
&= \arg\max \sum_{i=1}^{N} \log\left[h\left(x_i\right) \cdot \exp\left(\eta^{\top} \phi\left(x_i\right) - A(\eta)\right)\right] \\
&= \arg\max \sum_{i=1}^{N} \left[\log h\left(x_i\right) + \eta^{T} \phi\left(x_i\right) - A\left(\eta\right)\right] \\
&= \arg\max \sum_{i=1}^{N} \left[\eta^{T} \phi\left(x_i\right) - A\left(\eta\right]\right)
\end{aligned}
$$

$A(\eta)$ is a convex function, while it adds linear function. It is also a convex function.

$$
\begin{aligned}
&\frac{\partial}{\partial \eta}\left(\sum_{i=1}^{N}\left(\eta^{\top} \phi\left(x_i\right) - A(\eta)\right)\right. \\
=& \sum_{i=1}^{N} \frac{\partial}{\partial \eta}\left(\eta^{\top} \phi\left(x_i\right) - A(\eta)\right) \\
=& \sum_{i=1}^{N} \phi\left(x_i\right) - A'(\eta) \\
=& \sum_{i=1}^{N} \phi\left(x_i\right) - N A'(\eta) \\
&= 0 \\
&A'(\eta) = \frac{1}{N} \sum_{i=1}^{N} \phi\left(x_i\right)
\end{aligned}
$$

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

最大熵

等可能发生

信息量: $-\log p(x)$

Entropy:

$$\begin{aligned} E_{p(x)}[-\log p] &= \int -p(x)\cdot\log p(x)dx \\ &= -\sum_x p(x)\cdot\log p(x) \\ \rightarrow H[P] &= -\sum_x p(x)\log p(x) \end{aligned}$$

Assume that x is disrete random variable.

x	1	2	...	K
p	p_1	p_2	...	p_k

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

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

$$\hat{p}_i = \arg\max H[P] = \arg\min \sum_{i=1}^{k} p_i \log p_i$$

拉格朗日乘子法（求导两次判断convex凸函数）：

$$\begin{aligned} \mathcal{L}(P,\lambda) &= \sum_{i=1}^{K} p_i \log p_i + \lambda\left(1-\sum_{i=1}^{k} p_i\right) \\ \frac{\partial\mathcal{L}}{\partial p_i} &= \log p_i + p_i\frac{1}{p_i} - \lambda = 0 \\ \log p_i &+ 1 - \lambda = 0 \\ \hat{p}_i &= \exp(\lambda-1) = constant \\ \hat{p}_1 &= \hat{p}_2 = \cdots = \hat{p}_k = \frac{1}{k} \end{aligned}$$

$p(x)$ 均匀分布

最大熵原理

满足已知事实

检验分布 empirical distribution

$$
\begin{gathered}
\text{Data } =\{x_1, x_2, \ldots, x_N\} \\
\hat{p}(x=x)=\hat{p}(x)=\frac{\operatorname{count}(x)}{N} \\
E_{\hat{p}}[x], \operatorname{Var}_{\hat{p}}(x)
\end{gathered}
$$

是指数族分布