指数族分布

1. Introduction.

- 指数族分布是一类分布. 如

  Gauss Distribution. $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

  Bernoulli $f(x|p) = \begin{cases} p^x q^{1-x} & x=0,1 \\ 0 & x \neq 0,1 \end{cases}$

  ....

- 指数族分布可以写成统一的形式:

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

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

其中, η 是参数向量, $x \in \mathbb{R}^p$

$A(\eta)$: log partition function (可看作归一化因子)

$\phi(x)$: 充分统计量 (包含样本所有信息, 有它可代替样本点, 可理解为"人大代表")

Explanation:

$p(x|\theta) = \frac{1}{2}\hat{p}(x|\theta)$ → 归一化因子

时左边式子. $\exp(A(\eta))$ 看作 2, $h(x)\exp(\eta^T\phi(x))$ 看作 $\hat{p}(x|\theta)$.

则 $A(\eta) = \log(2)$

- Example

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

Gauss Distribution 的指数族分布形式

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

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

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

$e^a \cdot e^b = e^{a+b}$

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

η^T $\phi(x)$ $A(\eta)$

$\because \eta^T = \frac{\mu}{\sigma^2} - \frac{1}{2\sigma^2}$, 即 $\eta = \begin{bmatrix}\frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2}\end{bmatrix} \Rightarrow \eta_1 = \frac{\mu}{\sigma^2}, \eta_2 = -\frac{1}{2\sigma^2}$

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

$A(\eta)$ 可进一步化简为：$-\frac{\mu^2}{2\sigma^2}+\frac{1}{2}\cdot\log 2\pi\sigma^2$

$$= -\frac{\frac{\eta_1^2}{4\eta_2^2}}{-\frac{1}{\eta_2}} - \frac{1}{2}\log\left(2\pi\cdot\frac{1}{2\eta_2}\right)$$

$$= \frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log\left(\frac{\pi}{\eta_2}\right)$$

∴ Gauss Distribution 的指数族分布可写成

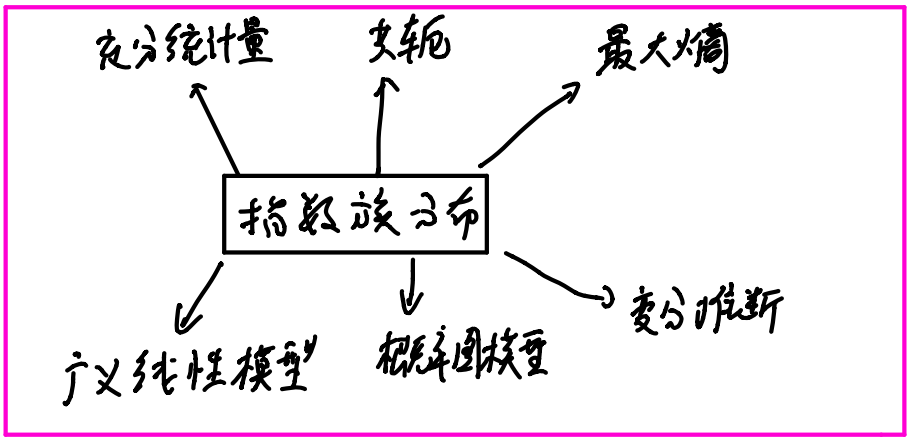
$$\eta=\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix},\quad \phi(x)=\begin{pmatrix}x\\ x^2\end{pmatrix}$$

$$A(\eta)=\frac{\eta_1^2}{4\eta_2}-\frac{1}{2}\log\left(\frac{\pi}{\eta_2}\right)$$

$$\Rightarrow \exp(\eta^T\phi(x)-A(\eta))$$

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

· 指数族分布的性质和应用



2. $\boxed{p(x|\eta)=h(x)\exp(\eta^T\phi(x)-A(\eta))}$ 中，$\phi(x)$ 与 $A(\eta)$ 的关系

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

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

explanation：$\int p(x|\eta)=1$

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

对两边 η 进行求导：

$$\exp(A(\eta)) \cdot A'(\eta) = \frac{\partial}{\partial \eta} \left(\int h(x) \exp(\eta^T \phi(x)\, dx \right)$$

对划线部分的 η 进行求导！

$$= \int h(x) \exp(\eta^T \phi(x)) \cdot \phi(x)\, dx$$

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

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

$$= \int p(x|\eta) \cdot \phi(x)\, dx$$

$$= E[\phi(x)]$$

连续型变量的期望定义：

$E(x) = \int p(x) f(x)\, dx$，其中 $p(x)$ 是密度函数，$f(x)$ 是分布函数

即有

$$A'(\eta) = E[\phi(x)]$$
$$A''(\eta) = Var[\phi(x)]$$

3. 极大似然估计与充分统计量

$$D = \{x_1, x_2, \cdots, x_n\}$$

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

对 $p(x|\eta)$ 进行极大似然估计：

$$\eta_{MLE} = \arg\max_{\eta} \log p(x|\eta)$$

$$= \arg\max_{\eta} \sum_{i=1}^{N} \log p(x_i|\eta)$$

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

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

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

$$\frac{\partial p(x|\eta)}{\partial \eta} = 0 \Rightarrow \frac{\partial}{\partial \eta} \sum_{i=1}^{N} [\eta^T \phi(x_i) - A(\eta)]$$

$$= \sum_{i=1}^{N} \phi(x_i) - A'(\eta)$$

$$= \sum_{i=1}^{N} \phi(x_i) - N A'(\eta) = 0$$

$$\Rightarrow \boxed{A'(\eta)_{MLE} = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)}$$

结论：为了估计参数，只需要知道充分统计量就可以了.

4. 最大熵角度

参考书：《统计学习方法》P94

《深度学习》花书指导 P_6 30+分钟

- 几个概念

① 信息量：$-\log p$

② 熵：$E_{p(x)}[-\log p] = \int -p(x)\cdot \log p(x)\, dx$ （连续型）

$$= -\sum_{x} p(x)\log p(x) \quad \text{（离散型）}$$

$$H[p] = -\sum_{x} p(x)\log p(x)$$

③ **最大熵 $\Leftrightarrow$ 等可能**

- Example（最大熵）

假设 x 是离散的

x	1	2	⋯	k
p	p_1	p_2	⋯	p_k

，$\sum_{i=1}^{k} p_i = 1$

求最大熵，则有如下最优化问题

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

写成拉格朗日形式

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

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

$$\Rightarrow \hat{p_i} = e^{\lambda - 1} \quad \text{（常数）}$$

$$\therefore p_1 = p_2 = \cdots p_k = \frac{1}{k}$$

∴ $p(x)$是均匀分布 → 等可能

最大熵原理：满足约束条件的模型集合中选取熵最大的模型。

④ 在实际问题中，求最大熵需满足已知事实(数据)

$Data = \{x_1, x_2, \cdots, x_n\}$

满足：经验分布 $\hat{p}(X=x) = \hat{p}(x) = \frac{count(x)}{N}$

则易求得 $E_{\hat{p}}[x]$, $Var_{\hat{p}}[x]$. 特征函数 $f(x)$ 关于经验分布 $\hat{p}$ 的期望值

设 $f(x)$ 是任意关于 x 的函数. $E_{\hat{p}}[f(x)] = \Delta$ → 已知

$f(x) = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, $\Delta = \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \\ \Delta_n \end{bmatrix}$

加上约束后，最大熵可写成

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

解决 → $p(x)$是指数族分布

真实分布 = 训练集样本分布

写成拉格朗日形式：

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

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

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

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

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

指数族分布