Exponential Family of Distribution

2021年6月1日 1:45

P39.40 Background

Features:
- Sufficient Statistic ①
- Conjugate ②
- Maximum Entropy (Noninformative priors)

Application:
- Generalized Linear Model
- Probability Graph
- Variational Inference

Gaussion
Bernoulli → Categorical
Binomial → Multinomial
Poisson
Dirichlet

} All Exponential Family of Distribution

Standard Forms: $y(x|\eta) = h(x)\exp\{\eta^T f(x) - A(\eta)\}$

η: parameter vector, $x \in \mathbb{R}^p$ $A(\eta)$: log partition function (log 配分函数) $f(x)$: Sufficient Statistic

$\because P(x|\eta) = \frac{1}{\exp\{A(\eta)\}} h(x)\exp\{\eta^T f(x)\}$ $\because \int P(x|\eta)dx = 1$

$\therefore \frac{1}{\exp\{A(\eta)\}}\int h(x)\exp\{\eta^T f(x)\} = 1$ $\therefore \exp\{A(\eta)\}$ is patition function (can be simply understood as a normalization factor)

$\because A(\eta) = \log\exp\{A(\eta)\}$ $\therefore A(\eta)$ is log partition fuction

① 包含原样本中全部有用信息，有时可丢弃原样本
如满足 Guassion 分布一组样本，$f(x) = \begin{bmatrix} \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i^2 \end{bmatrix}$ 即为充分统计量

② e.g.: $P(z|x) = P(x|z)P(z)$ If likelihood is EFD, prior and posterior are identically distribution.

③ Initialize prior

P41 Gussion Distribution

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

$P(x_1,x_2|\theta_1,\theta_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\{-\frac{1}{2(1-\rho^2)}[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}]\}$

$P(x_1\cdots x_N|\theta) = 2\pi^{-\frac{N}{2}}|\Sigma|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$

$$\Downarrow$$

$$P(x|\theta) = h(x)\exp\{\eta^T f(x) + A(\eta)\}$$

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

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

$$\eta = \begin{pmatrix}\eta_1\\\eta_2\end{pmatrix} = \begin{pmatrix}\frac{\mu}{\sigma^2}\\-\frac{1}{2\sigma^2}\end{pmatrix} \Rightarrow \begin{cases}\sigma^2 = -\frac{1}{2\eta_2}\\ \mu = -\frac{\eta_1}{2\eta_2}\end{cases} \qquad A(\eta) = -(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log 2\pi\sigma^2) = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log(-\frac{\pi}{\eta_2})$$

原式 $= \exp\{\eta^T f(x) - A(\eta)\}$

P42 Log Partition Function

$P(x|\eta) = \frac{1}{\exp\{A(\eta)\}}h(x)\exp\{\eta^T f(x)\}$ $\because \int P(x|\eta)dx = 1$ $\therefore \exp\{A(\eta)\} = \int h(x)\exp\{\eta^T f(x)\}dx$

Partial derivative of η on both side of equation: $A'(\eta)\exp\{A(\eta)\} = \frac{\partial}{\partial\eta}(\int h(x)\exp\{\eta^T f(x)\}dx) = \int h(x)\exp\{\eta^T f(x)\}f(x)dx$

$\therefore A'(\eta) = \int h(x)\exp\{\eta^T f(x) - A(\eta)\}f(x)dx$ # Derivative and integral are interchangeable

$= \int P(x|\eta)f(x)dx = E[f(x)]$

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

$= \int P(x|\eta)f(x)^2dx - A'(\eta)\int P(x|\eta)f(x)dx$? $\int P(x)x^2dx = E(x^2)$?

$= E[f(x)^2] - E[f(x)]^2 = Var[f(x)]$

P43 MLE & Sufficient Statistics

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

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

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

$\therefore A'(\eta_{MLE}) = \frac{1}{N}\sum_{i=1}^N f(x_i) \Rightarrow \eta_{MLE} = A'^{-1}(\eta_{MLE})$ Inverse Function

P44/45 Maximum Entropy Perspective

$$H(P)\ (\text{Entropy}) = \begin{cases} -\int P(x)\log P(x)dx \\ -\sum P(x_i)\log P(x_i) \end{cases}$$

$$\begin{cases} \max \sum_{i=1}^N -P_i\log P_i = \min \sum_{i=1}^N P_i\log P_i \\ s.t. \sum_{i=1}^N P_i = 1 \end{cases}$$

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

Lagrange Multiplier Method: $\mathcal{L}(p_i, \lambda) = \sum_{i=1}^{N} p_i \log p_i + \lambda(1 - \sum_{i=1}^{N} p_i)$

$\frac{\partial \mathcal{L}}{\partial p_i} = \log p_i + p_i \frac{1}{p_i} - \lambda = 0 \Rightarrow p_i = \exp\{1-\lambda\}$ λ is hyperparameter $\therefore p_i$ is constant

※ $\therefore p_1 = p_2 = \cdots = p_N = \frac{1}{N}$

Maximum Entropy + known facts (constrains) $\Rightarrow$ Priciple of Maximum Entropy

(经验分布)

Data: $\{x_1, x_2, \cdots, x_N\}$ obey Empirical Distribution $\hat{p}(X=c) = \hat{p}(x) = \frac{count(c)}{N}$

$\therefore E_{\hat{p}}[x]$, $Var_{\hat{p}}[x]$ are known # P is the overall distribution ; $\hat{P}$ is the sample distribution

$\therefore E_{\hat{p}}[g(x)]$ is known # $g(x)$ is an arbitrary vector of fuctions of x & $g(x)$ is a sufficient statistic of X

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

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

$\frac{\partial \mathcal{L}}{\partial p(x)} = \sum_x [\log p(x) + 1 - \lambda_0 - \lambda^T g(x)] = 0$ $\therefore \log p(x) + 1 - \lambda_0 - \lambda^T g(x) = 0$?

$\therefore \log p(x) = \lambda^T g(x) + \lambda_0 - 1 \Rightarrow p(x) = \exp\{\underbrace{\lambda^T g(x)}_{\eta^T f(x)} - \underbrace{(1-\lambda_0)}_{A(\eta)}\}$