STAT201B Lecture 6 Minimal Sufficiency

1 Minimal Sufficiency

Logic

关于 minimal sufficiency 的更多论述, 见 STA3020 Lecture 4, 包括:

- minimal sufficient 的定义
- Lehmann-Scheffé theorem
- Bahadur's theorem
- Exponential family 的 minimal sufficient statistic

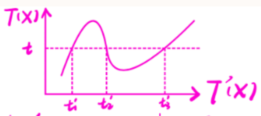
1.1 Definition: Minimal Sufficiency

令 $T(X)$ 为一个 sufficient statistic
若对于**任意**其他 sufficient statistic $S(X)$, **$T(X)$ 为 a function of $S(X)$**, 即 $T = f(S)$ for some f
则 $T(X)$ 为 **minimal sufficient**

Remark

- $T = f(S)$ 表示了两件事:
  - 关于 S 的 knowledge implies 关于 T 的 knowledge
  - T 提供了 greater reduction of data, 除非 f 为 one-to-one

    注: 换言之, $T'(x)=T'(y) \Rightarrow T(x)=T(y)$
    $T(x)=T(y) \not\Rightarrow T'(x)=T'(y)$

    也就是说, 不同的 $\underbrace{\text{values}}_{(t_1',t_2',t_3')}$ of $T'(x)$ 所包含的信息, 可以被 $T(x)$ 的一个 $\underbrace{\text{value}}_{t}$ 所包含.
- Minimal sufficient statistic 仍然不是 unique 的 (通过 one-to-one mapping preserve)

1.2 Theorem: Lehmann-Scheffé theorem (补充)

令:

- $X_1, \dots, X_n \overset{i.i.d.}{\sim} f(\cdot|\theta)$, 其中 $\theta \in \Theta$
- $T = T(X)$ 为一个 statistics

若:

$$T(x) = T(x') \iff \frac{f(x|\theta)}{f(x'|\theta)} \text{ is invariant over } \theta, \forall x, x' \in \Omega$$

则 T 为 **minimal sufficient** for θ

Remark

换言之, 当且仅当 $T(X)$ 的 value 变化时, likelihood function 中关于 θ 的信息才会变

2 Exponential Family 的 Minimal Sufficiency

2.1 Definition: Exponential family

Logic

关于 exponential family 的更多论述, 见 STA3020 Lecture 3, 包括:

- Exponential family, parameter space, canonical form, curved exponential family 的定义
- Exponential family 的例子
- Exponential family 的性质

d-parameter exponential family:

一个 **d-parameter exponential family** 的 pmf/pdf 满足以下形式:

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{\theta}) &= h(\boldsymbol{x}) \cdot c(\boldsymbol{\theta}) \cdot exp\left[\sum_{i=1}^{d} \eta_i(\boldsymbol{\theta})\, T_i(\boldsymbol{x})\right] \\ &= h(\boldsymbol{x}) \cdot exp\left[\sum_{i=1}^{d} \eta_i(\boldsymbol{\theta}) T_i(\boldsymbol{x}) \;-A(\boldsymbol{\theta})\right] \end{aligned}$$

其中,

- $h(\boldsymbol{x}) \geq 0$
- $c(\boldsymbol{\theta}) \geq 0$
- $T_1(\boldsymbol{x}), \ldots, T_d(\boldsymbol{x})$ 为关于 $\boldsymbol{x} = (x_1, \ldots, x_n)$ 的 real value functions, 且不取决于 $\boldsymbol{\theta}$
- $\eta_1(\boldsymbol{\theta}), \ldots, \eta_d(\boldsymbol{\theta})$ 为关于 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ 的 real value functions, 且不取决于 $\boldsymbol{x}$

full rank exponential family:

若:

- $\boldsymbol{\eta}(\Theta) = \{\eta_1(\boldsymbol{\theta}), \ldots, \eta_d(\boldsymbol{\theta})\}$ 在 $\mathbb{R}^d$ 中有 non-empty interior
- $T_1(\boldsymbol{x}), \ldots, T_d(\boldsymbol{x})$ 为 linearly independent

则该 exponential family 被称为 **full rank**

Remark

若 $\eta(\Theta)$ 仅包含 $\mathbb{R}^s$ $(s < d)$ 中的 open set, 则该 exponential family 被称为 curved exponential family with dimension s

Example

1. $\mathcal{N}(\mu, \mu), \mu > 0$ 构成一个 curved exponential family
2. $\mathcal{N}(\mu, \mu^3), \mu > 0$ 构成一个 curved exponential family

一个 curved exponential family 的例子

若 random sample $X_1, \cdots, X_n$ i.i.d. 取自 normal distribution $N(\theta, \theta^3)$, 其中 $\Theta = \{\theta\} \in R^+$ 未知, 则 sample 构成一个 curved exponential family

证明:

$$f(x|\theta) = \underbrace{\left(\frac{1}{\sqrt{2\pi}}\right)^n}_{h(x)} \underbrace{\left(\frac{1\{\theta>0\}}{\theta^{3n/2}} exp\left\{-\frac{n}{2\theta}\right\}\right)}_{c(\theta)} exp\left\{\underbrace{-\frac{1}{2\theta^3}}_{w_1} \underbrace{\sum_{i=1}^{n} X_i^2}_{T_1} + \underbrace{\frac{1}{\theta^2}}_{w_2} \underbrace{\sum_{i=1}^{n} X_i}_{T_2}\right\}$$

$$\Theta = \{\theta > 0\}, \quad dim(\Theta) = 1 < 2$$

Example

问题:
证明 $X \sim Exponential(\lambda)$ 为指数族

解答:

$$f(x, \lambda) = \lambda \cdot e^{-\lambda x} \cdot \mathbf{1}(x \geq 0)$$

因此有

$$h(x) = \mathbf{1}(x \geq 0), \quad c(\theta) = \lambda, \quad \eta(\theta) = -\lambda, \quad T(x) = x$$

Example

问题:
证明 $X \sim Binomial(n, p)$ 为指数族

解答:

$$\begin{aligned}
f(x,p) &= \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} \\
&= \binom{n}{x} \cdot exp\{x\ logp + (n-x) \cdot log(1-p)\} \\
&= \binom{n}{x} \cdot exp\left\{\left(log\frac{p}{1-p}\right)x + n \cdot log(1-p)\right\} \\
&= \binom{n}{x} \cdot (1-p)^n \cdot exp\left\{\left(log\frac{p}{1-p}\right) \cdot x\right\}
\end{aligned}$$

因此有

$$h(x) = \binom{n}{x}, \quad c(p) = (1-p)^n, \quad \eta(p) = log\frac{p}{1-p}, \quad T(x) = x$$

Remark

严谨来说, density function 应该还要包括 (与 support 相关的) indicator function

Example

问题:
证明 $X \sim \mathcal{N}(\mu, \sigma^2)$ (μ 和 σ^2 均未知) 为指数族

解答:

$$\begin{aligned}
f(x;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{x^2}{2\sigma^2} + \frac{2\mu x}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right\} \\
&= \frac{1}{\sqrt{2\pi}} \cdot \left(\frac{1}{\sigma} \cdot exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}\right) \cdot exp\left\{-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}\right\}
\end{aligned}$$

因此有

$$h(x) = \frac{1}{\sqrt{2\pi}}, \quad c(\theta) = \frac{1}{\sigma} \cdot exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}, \quad \eta(\theta)^T = \left(-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}\right), \quad T(x)^T = (x^2, x)$$

Example

Poisson exponential family

若 random sample $X_1, \cdots, X_n$ i.i.d. 取自 Poisson distribution $Poi(\lambda)$, 其中 $\theta = \{\lambda\}$ 未知,
则 sample 构成一个 exponential family

证明:

加上定义域

$$\begin{aligned}
f(x|\theta) &= \prod_{i=1}^{n} \frac{1}{x_i!} \lambda^{x_i} e^{-\lambda} \cdot 1\{x_i \in N\} \\
&= exp\left\{ log\left[\left(\prod_{i=1}^{n} \frac{1}{x_i!}\right) \lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}\right]\right\} \cdot 1\{x_i \in N\} \\
&= exp\left\{-\sum_{i=1}^{n} log(x_i!) + log(\lambda) \cdot \left(\sum_{i=1}^{n} x_i\right) - n\lambda\right\} \cdot 1\{x_i \in N\} \\
&= \underbrace{exp\left\{-\sum_{i=1}^{n} log(x_i!)\right\} \cdot 1\{x_i \in N\}}_{h(x)} \cdot \underbrace{exp\{-n\lambda\}}_{c(\theta)} \cdot exp\left\{\underbrace{log(\lambda)}_{w_1} \cdot \underbrace{\left(\sum_{i=1}^{n} x_i\right)}_{T_1}\right\}
\end{aligned}$$

☰ **Example** ∨

Gamma exponential family

若 random sample $X_1,\cdots,X_n$ i.i.d. 取自 Gamma distribution $\Gamma(\alpha,\beta)$，其中 $\theta=\{\alpha,\beta\}$ 未知，则 sample 构成一个 exponential family

☰ **Example** ∨

Student t distribution with degree of freedom θ (一个反例)

若 random sample $X_1,\cdots,X_n$ i.i.d. 取自 Student t distribution t_θ，其中 $\theta=\{\theta\}$ 未知，则 sample 不构成一个 exponential family

证明:

$$f(x|\theta)=\prod_{i=1}^{n}\frac{\Gamma(\frac{\theta+1}{2})(1+\frac{x_i^2}{\theta})^{-\frac{\theta+1}{2}}}{\sqrt{\theta\pi}\,\Gamma(\frac{\theta}{2})}\cdot 1\{\theta>0\}$$

$$=\left(\frac{\Gamma(\frac{\theta+1}{2})}{\sqrt{\theta\pi}\,\Gamma(\frac{\theta}{2})}\right)^{n}\cdot 1\{\theta>0\}\cdot\exp\left\{-\frac{\theta+1}{2}\sum_{i=1}^{n}\log\left(1+\frac{X_i^2}{\theta}\right)\right\}$$

没法分开 θ 和 X_i

☰ **Example** ∨

Uniform distribution (一个反例)

若 random sample $X_1,\cdots,X_n$ i.i.d. 取自 Uniform distribution $Unif(0,\theta)$，其中 $\theta=\{\theta\}$ 未知，则 sample 不构成一个 exponential family

证明:

$$f(x|\theta)=\prod_{i=1}^{n}\frac{1}{\theta}1\{0\le X_i\le\theta\}$$

$$=\left(\frac{1}{\theta}\right)^{n}1\{\min X_i\ge 0\}\cdot 1\{\max X_i\le\theta\}$$

没法分开 θ 和 X_i

2.2 Theorem: Exponential family 的 minimal sufficient statistic

若: $X=\{X_1,\dots,X_n\}$ 的分布来自 **full rank exponential family**
则: $T=(T_1,\dots,T_d)$ 为 minimal sufficient statistics

☰ **Example** ∨

问题:

令 $X_1,\dots,X_n\overset{i.i.d.}{\sim}\mathcal{N}(\mu,\sigma^2)$, 求 μ 和 σ^2 的 minimal sufficient statistic

解答:

对于 Normal random variables, 我们可以做以下变形:

$$\begin{aligned}f_{\mu,\sigma^2}(x_1,\dots,x_n)&=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\prod_{i=1}^{n}exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}\\&=\left(\frac{1}{\sqrt{2\pi}}\right)^n exp\left\{-\frac{\sum_{i=1}^{n}x_i^2-2\mu\sum_{i=1}^{n}x_i+\mu^2}{2\sigma^2}-n\,ln\sigma\right\}\\&=\left(\frac{1}{\sqrt{2\pi}}\right)^n exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2-\frac{\mu}{2\sigma^2}\sum_{i=1}^{n}x_i+\frac{n\mu^2}{2\sigma^2}-n\,ln\sigma\right\}\end{aligned}$$

其中:

- $h(\boldsymbol{x})=\left(\frac{1}{\sqrt{2\pi}}\right)^n$
- $\eta_1(\boldsymbol{\theta})=-\frac{1}{2\sigma^2}$
- $T_1(\boldsymbol{x})=\sum_{i=1}^{n}x_i^2$
- $\eta_2(\boldsymbol{\theta})=-\frac{\mu}{2\sigma^2}$
- $T_2(\boldsymbol{x})=\sum_{i=1}^{n}x_i$

- $A(\boldsymbol{\theta}) = \frac{n\mu^2}{2\sigma^2} - n\ ln\sigma$

因此 $(T_1(\boldsymbol{x}), T_2(\boldsymbol{x})) = (\sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} x_i)$ 为 minimal sufficient statistics