STAT201B Lecture 5 Sufficiency

1 Sufficiency

Logic

在对 θ 进行 inference 的时候, 我们希望仅将 data 中相关的 information 分离出来, 即在不损失信息的情况下将 data 压缩成 $T(X)$, 这样的好处在于:

1. 提升 computational efficiency
2. 降低 storage requirements
3. 包含 irrelevant information 可能会增加 estimator 的 risk (见 Rao-Blackwell Theorem)
4. 提升数据的 scientific interpretability

1.1 Definition: Sufficient Statistic

Logic

关于 Sufficient Statistics 的更多论述, 见笔记 STA3020 Notes for Lecture 4

令

1. X 的分布来自于 $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ (一个与 θ 相关的分布族),
2. Statistic T 的 range 为 $\mathcal{T}$

则 statistic T 被称为 **sufficient**,

若 **对于任意 $t \in \mathcal{T}$, conditional distribution $P_\theta(X|T(X)=t)$ 与 θ independent**

Example

令 $X_i \overset{i.i.d.}{\sim} Ber(\theta), i = 1, \dots, n$, 则 $T = \sum_{i=1}^n X_i$ 是 θ 的 sufficient statistic:

由于

$$
\begin{aligned}
P_\theta(X|T(X)=t) &= \frac{P_\theta(X_1, \dots, X_n, T(X)=t)}{P_\theta(T(X)=t)} \\
&= \begin{cases} 0 & \text{if } t \neq \sum_{i=1}^n X_i \\ \frac{P_\theta(X_1=x_1,\dots,X_n=x_n)}{P_\theta(\sum_{i=1}^n x_i = t)} = \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{x_i}}{\binom{n}{t}\theta^t(1-\theta)^t} = \frac{\theta^t(1-\theta)^t}{\binom{n}{t}\theta^t(1-\theta)^t} = \frac{1}{\binom{n}{t}} & \text{if } t = \sum_{i=1}^n X_i \end{cases}
\end{aligned}
$$

与 θ independent, $\forall t \in \mathcal{T}$, T 为 θ 的 sufficient statistic

1.2 Theorem: Neyman Factorization Theorem

Logic

使用定义求解 sufficient statistics 较为繁琐, 可以使用 Neyman Factorization Theorem 快速求解

令 distribution family $\{P_\theta : \theta \in \Omega\}$ 有 joint mass / density $\{p(x;\theta) : \theta \in \Omega\}$

则

$$
T \text{ is sufficient for } \theta \iff \exists \text{ functions } h \text{ and } g, \text{ such that } p(x;\theta) = h(x) \cdot g(T(x), \theta)
$$

Proof (仅考虑 discrete case)

Proof. (of Theorem.1.10: Neyman-Fisher Factorization Theorem) We prove the FactorizationTheorem for discrete random variables as an illustration, the general proof follow the same line.

- Suppose T is sufficient. We want to prove the right hand side of (1.2). Let $t = T(x)$. The joint pmf. of X is

X=x 时的 realization

X=x 时一定有T=t

$$f(x|\theta) = \mathbb{P}_\theta(X = x) = \mathbb{P}_\theta(\{X = x\} \bigcap \{T = t\})$$
$$= \mathbb{P}_\theta(X = x|T = t)\mathbb{P}_\theta(T = t) =: h(x)g(t, \theta).$$

由 sufficiency: ⊥ θ

- Suppose the right hand side of (1.2) holds. We want to prove T is sufficient. Apparently, when $T(x) \neq t$, we have $\mathbb{P}_\theta(X = x|T = t) = 0$ by definition, which is invariant over θ. Meanwhile, when $T(x) = t$. Let $S = \{x' \in \mathcal{X}_n : T(x') = t\}$. Then

$$\mathbb{P}(X = x|T = t) = \frac{\mathbb{P}_\theta(\{X = x\} \bigcap \{T = t\})}{\mathbb{P}_\theta(T = t)} = \frac{\mathbb{P}_\theta(\{X = x\})}{\mathbb{P}_\theta(T = t)}$$
$$= \boxed{\frac{\mathbb{P}_\theta(\{X = x\})}{\{\sum_{x' \in S} \mathbb{P}_\theta(X = x')\}}} = \frac{g(t, \theta)h(x)}{\{\sum_{x' \in S} g(t, \theta)h(x')\}} = \frac{h(x)}{\{\sum_{x' \in S} h(x')\}}$$

which is invariant over θ, and conclude that T is sufficient by definition.

Example

令 $Y_i \overset{i.i.d.}{\sim} Uniform(0, \theta), i = 1, \ldots, n$
证明: $T = Y_{(n)}$ 为 θ 的 sufficient statistic

$$\begin{aligned} P_\theta(Y) &= \prod_{i=1}^n \left(\frac{1}{\theta}\right) \mathbf{1}(0 < Y_i < \theta) \\ &= \left(\frac{1}{\theta}\right)^n \left[\prod_{i=1}^n \mathbf{1}(0 < Y_i < \theta)\right] \\ &= \left(\frac{1}{\theta}\right)^n \cdot \mathbf{1}(Y_{(1)} > 0) \cdot \mathbf{1}(Y_{(n)} < \theta) \\ &= \mathbf{1}(Y_{(1)} > 0) \cdot \left[\left(\frac{1}{\theta}\right)^n \cdot \mathbf{1}(Y_{(n)} < \theta)\right] \end{aligned}$$

其中 $\mathbf{1}(Y_{(1)} > 0)$ 可以被视作 $h(y)$, $(\frac{1}{\theta})^n \cdot \mathbf{1}(Y_{(n)} < \theta)$ 可以被视作 $g(T(y), \theta)$

Example

例 1: $X_1, \cdots, X_n \overset{i.i.d}{\sim} Poisson(\lambda)$, 求 λ 的 sufficient statistic

$$f(x|\lambda) = \prod_{i=1}^n \left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \cdot 1\{x_i \in N\}\right)$$
$$= \underbrace{\left(\prod_{i=1}^n \frac{1}{x_i!} \cdot 1\{x_i \in N\}\right)}_{h(x)} \cdot \underbrace{\left(e^{-n\lambda} \cdot \lambda^{n\bar{x}}\right)}_{g(t,\lambda)}$$

$t = \bar{x}$, i.e. $T(X) = \bar{X}$

Example

例2: $X_1,\cdots,X_n \overset{iid}{\sim} \text{Bernoulli}(p)$，其中 $p\in(0,1)$，求入的 sufficient statistic

$$f(x|p)=\prod_{i=1}^{n}\left(p^{x_i}(1-p)^{1-x_i}\cdot 1_{\{x_i\in\{0,1\}\}}\right)$$

$$=\underbrace{\left(\frac{p}{1-p}\right)^{n\bar{x}}\cdot(1-p)^n}_{g(t,\theta)}\cdot\underbrace{\left[\prod_{i=1}^{n}1_{\{x_i\in\{0,1\}\}}\right]}_{h(x)}$$

$t=\bar{x}$, i.e. $T(X)=\bar{X}$

Example

例3: $X_1,\cdots,X_n \overset{iid}{\sim} N(\mu,\sigma^2)$，其中 $\mu\in\mathbb{R}$, $\sigma^2\in\mathbb{R}^+$. 求下述情况下 θ 的 sufficient statistic

① μ 未知, σ^2 已知, $\theta=\mu$

$$f(x|\theta)=\prod_{i=1}^{n}\left\{(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right\}$$

$$=\underbrace{\left[\frac{\exp\left(-\frac{\sum x_i^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{n/2}}\right]}_{h(x)}\cdot\underbrace{\exp\left\{-\frac{n\mu^2-2n\mu\bar{x}}{2\sigma^2}\right\}}_{g(t,\theta)}$$

$t=\bar{x}$, i.e. $T(X)=\bar{X}$

② μ 已知, σ^2 未知, $\theta=\sigma^2$

$$f(x|\theta)=\prod_{i=1}^{n}\left\{(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right\}$$

$$=\underbrace{\frac{1}{(2\pi\sigma^2)^{n/2}}\cdot\exp\left\{-\frac{n\cdot\left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}\right)}{2\sigma^2}\right\}}_{g(t,\theta)}$$

$t=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$, i.e. $T(X)=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$

③ μ, σ^2 未知, $\theta=(\mu,\sigma^2)$

$$f(x|\theta)=\prod_{i=1}^{n}\left\{(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right\}$$

$$=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2\right]\right\}$$

$$=\underbrace{(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)\cdot s^2+n(\bar{x}-\mu)^2\right]\right\}}_{g(t,\theta)}\quad\left(s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}\right)$$

$T(X)=(S^2,\bar{X})$

1.3 Theorem: Rao-Blackwell Theorem

Logic

关于 loss, risk, admissibility, Rao-Blackwell Theorem 的更多描述, 见 STA3020 Notes for Lecture 5

令:

- X 为分布为 $P_\theta\in\mathcal{P}=\{P_\theta,\theta\in\mathbb{R}\}$ 的随机变量
- $\delta(X)$ 为 θ 的 (任意) 一个 estimator

若:

- $T(X)$ 为 θ 的一个 **sufficient** statistic
- Loss function $\mathcal{L}(\theta,\delta(X))$ 为关于 θ 的 **strictly convex** function (如 L_2 loss)
- $\delta(X)$ 有 finite expectation 与 risk，即 $R(\theta,\delta(X))=\mathbb{E}[\mathcal{L}(\theta,\delta(X))]<\infty$

则:

- 若定义 $\eta(t) = E_\theta[\delta(X)|T=t], \forall t$ (即 $\delta(X)$ 在 T 下的条件期望，是一个关于 T 的函数)
- 则 estimator $\eta(T) = E_\theta[\delta(X)|T(X)]$ 满足:

$$R(\theta, \eta) < R(\theta, \delta)$$

  除非 $\delta(X) = \eta(T)$ with probability 1

Remark

1. 若 strictly convex 被替换为 convex, 则 $<$ 被替换为 $\leq$, 但是若去除 convexity assumption, 则定理不成立
2. 此处 T 被要求为 sufficient statistic, 这主要是为了**确保 $\eta(T)$ independent with θ** (因此可以被视作一个 estimator)
3. Rao-Blackwell theorem 的实际意义在于: 我们**可以通过 conditioning on sufficient statistics 来优化现有的 estimator**

Proof

由 Jensen's inequality, 有

$$\begin{aligned}\mathcal{L}(\theta, \eta(t)) &= \mathcal{L}(\theta, \mathbb{E}[\delta(X)|T(X)=t]) \\ &\leq \mathbb{E}[\mathcal{L}(\theta, \delta|T(X)=t)]\end{aligned}$$

当且仅当 $\theta = \eta(t)$ with probability 1 时取等;
对两侧取期望, 有:

$$R(\theta, \eta) = \mathbb{E}[\mathcal{L}(\theta, \eta)] \leq \mathbb{E}[\mathbb{E}[\mathcal{L}(\theta, \delta(X)|T)]] = R(\theta, \delta)$$

Example: 对 MSE 的优化

考虑 estimator δ 和 L2 loss $\mathcal{L}(\theta, \delta) = (\theta - \delta)^2$, 则经过 Rao-Blackwell 优化过的 estimator $\eta = \mathbb{E}_\theta[\delta|T]$ 满足:

$$\begin{aligned}R(\theta, \eta) &= \mathbb{E}_\theta[\mathcal{L}(\theta, \eta)] \\ &= \mathbb{E}_\theta[(\theta - \mathbb{E}_\theta[\delta|T])^2] \\ &= \mathbb{E}_\theta[(\mathbb{E}_\theta[\theta - \delta|T])^2] \quad (\text{condition on } \theta, \text{可以将 } \theta \text{ 视作常数放入期望}) \\ &\leq \mathbb{E}_\theta[\mathbb{E}_\theta[(\theta - \delta)^2|T]] \quad (\text{Jensen's inequality}) \\ &= \mathbb{E}_\theta[(\theta - \delta)^2] \\ &= R(\theta, \delta)\end{aligned}$$

1.4 Jensen's Inequality

令:

1. $(\Omega, \mathcal{F}, \mathbb{P})$ 为一个 probability space
2. $X: \Omega \to R$ 为一个 integrable random variable (即 $\mathbb{E}[|X|] < \infty$)

若:

1. $\varphi: \mathbb{R} \to \mathbb{R}$ 为 **convex** function
2. $\varphi: \mathbb{R} \to \mathbb{R}$ 为 integrable

则:

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$$

Remark

若 φ 为 strictly convex, 则当且仅当 X is almost surely constant 时取等

Example

$(\mathbb{E}[X])^2 \leq \mathbb{E}[X^2]$