STAT201B Lecture 17 Decision Theory

1 Loss

Logic

Statistical decision theory 研究的是如何在uncertainty 下做出最优决策

我们把这种不确定性用一个 unknown parameter θ 表示, 称为 state of nature

θ 通常表示未知但影响结果的真实参数

Logic

关于 loss 的更多描述, 见 STA3020 Lecture 5

1.1 Definition: action

令:

- 所做的 particular decision 为 **action** a
- collection of all possible actions 为 $\mathcal{A}$

1.2 Definition: loss function

令:

- $\mathcal{A}$ 为 collection of all possible actions
- $a \in A$ 为 a particular decision
- Θ 为 space of state of nature
- $\theta \in \Theta$ 为 a particular (unknown) state of nature

则 **loss function** 被表示为:

$$L(\theta, a) : (\Theta \times \mathcal{A}) \to [0, \infty)$$

表示在 true state of nature 为 θ 时采取 action a 所造成的 "损失"

Example

Example: A drug company has developed a new pain reliever. They are trying to determine how much of the drug to produce, but they are uncertain about the proportion of the market the drug will capture (θ).

Suppose a is an estimate of θ. The company plans to produce an amount proportional to a. One possible loss function is

$$L(\theta, a) = \begin{cases} K(\theta - a) & a - \theta < 0 \\ 2K(a - \theta) & a - \theta \geq 0 \end{cases}$$

for some constant K. This loss function implies that an overestimate of demand (leading to overproduction of the drug) is considered twice as costly as an underestimate. The loss is also taken to be linear, which may be reasonable if the total cost is proportional to the number of units produced.

1.3 常见的 loss functions

常见的 loss functions 包括:

- Squared error loss: $L(\theta, a) = (\theta - a)^2$
- Linear loss: $L(\theta, a) = \begin{cases} K_1(\theta - a), & a - \theta < 0 \\ K_2(a - \theta), & a - \theta \geq 0 \end{cases}$
- Absolute error loss: $L(\theta, a) = |\theta - a|$
- L^p loss: $L(\theta, a) = |\theta - a|^p$
- Zero-one loss: $L(\theta, a) = \begin{cases} 0, & a = \theta \\ 1, & a \neq \theta \end{cases}$

> **Remark**
>
> Absolute error loss 等价于 linear loss with $K_1 = K_2$

2 Risk

> **Logic**
>
> 由于 loss function 通常未知, 且这种未知源于
>
> - true parameter θ 具有随机性 (Bayesian perspective)
> - action 本身具有随机性 (Frequentist & Bayesian perspective)
>
> 因此在选择 optimal decision 的时候, 我们通常考虑 expected loss (剔除随机性), 即 risk
>
> 由于我们可以对不同的随机量 (parameter / action) 求 expectation, 因此有不同类型的 risks:
>
> - posterior risk
> - (frequentist) risk
> - Bayes risk

> **Logic**
>
> 关于 frequentist risk 和对应的 admissibility 的更多描述, 见 STA3020 Lecture 5
>
> 关于 Bayes risk 和对应的 admissibility 的更多描述, 见 STA3020 Lecture 26

> **Logic**
>
> 在接下来的讨论中, 我们仅考虑 estimation problems, 即 action 为 $a = \hat{\theta}(x)$

> **Logic**
>
> 第一种想法是: 利用 posterior distribution 来 take expectation on θ

2.1 posterior risk

2.1.1 Definition: posterior risk

The **posterior risk** 被定义为:

$$
\begin{aligned}
r(\hat{\theta}|x) &= \mathbb{E}_{\theta|X}[L(\theta, \hat{\theta}(x))] \\
&= \int L(\theta, \hat{\theta}(x)) f(\theta|x)\, d\theta
\end{aligned}
$$

> **Remark**
>
> $\mathbb{E}_{\theta|X}[L(\theta, \hat{\theta}(x))]$ 表示对 $\theta|X$ 求 expectation

> **Remark**
>
> - posterior risk 是一个 function of x
> - 对 posterior risk 的理解:
>   - 可以理解为: 在 conditioning on observation x 的前提下, average over uncertainty in θ
>   - 对于一组 observation x, 我们仅会为 estimator $\hat{\theta}$ 求出一个 value

---

> **Logic**
>
> 第二种想法是: 利用 likelihood 来 take expectation on x

2.2 (frequentist) risk

2.2.1 Definition: (frequentist) risk

The **(frequentist) risk** 被定义为:

$$
\begin{aligned}
R(\theta, \hat{\theta}) &= \mathbb{E}_{X|\theta}[L(\theta, \hat{\theta}(x))] \\
&= \int L(\theta, \hat{\theta}(x)) f(x|\theta)\, dx
\end{aligned}
$$

> **Remark**
>
> - (frequentist) risk 是一个 function of θ
> - 对 (frequentist) risk 的理解:
>   - 可以理解为: 在 given the true state of nature 为 θ 的前提下, average over different possible realizations x
>   - 对于一个 parameter θ, 我们仅会为 estimator $\hat{\theta}$ 求出一个 value

2.2.2 Definition: Admissibility

- 一个 estimator $\hat{\theta}$ 被称为 **inadmissible**, 若存在另一个 estimator $\hat{\theta}'$ dominates θ, 即:

$$
\exists\, \hat{\theta}' \quad s.t. \quad \begin{cases} R(\theta, \hat{\theta}') \leq R(\theta, \hat{\theta}) & \text{for all } \theta \in \Theta \\ R(\theta, \hat{\theta}') < R(\theta, \hat{\theta}) & \text{for some } \theta \in \Theta \end{cases}
$$

- 一个 estimator $\hat{\theta}$ 被称为 **admissible**, 若上述 estimator 不存在

> **Remark (补充)**
>
> 对于 convex loss function, 有以下关于 admissibility 的定理/结论:
>
> - Rao-Blackwell Theorem
> - 若 loss function 为 strictly convex, 则 admissible estimator 唯一

> **Example**
>
> ***问题:***

令:

- $X \sim \mathcal{N}(\theta, 1)$
- 使用 squared error loss
- estimator 为 $\hat{\theta}_c(x) = cx$

1. 求出 risk (in terms of c 和 θ)
2. 计算 $c = 1$ 时的 risk
3. 证明: $c > 1$ 时, $\hat{\theta}_c$ 为 inadmissible
4. 作图比较 $c = \frac{1}{2}$ 和 $c = 1$ 时的 risk

解答:

Question 1:

根据定义, risk 为:

$$\begin{aligned} R(\theta, \hat{\theta}_c) &= \mathbb{E}_{X|\theta}[(\theta - cX)^2] \\ &= \mathbb{E}_{X|\theta}[\theta^2 - 2cX\theta + c^2X^2] \\ &= \theta^2 - 2c\theta^2 + c^2(1 + \theta^2) \\ &= (c-1)^2\theta^2 + c^2 \end{aligned}$$

Question 2:

当 $c = 1$ 时, 有:

$$R(\theta, \hat{\theta}_1) = (1-1)^2\theta^2 + 1^2 = 1$$

Question 3:

当 $c > 1$ 时, 对于任意 θ, 我们有:

$$R(\theta, \hat{\theta}_c) = (c-1)^2\theta^2 + c^2 \geq c^2 > 1 = R(\theta, \hat{\theta}_1)$$

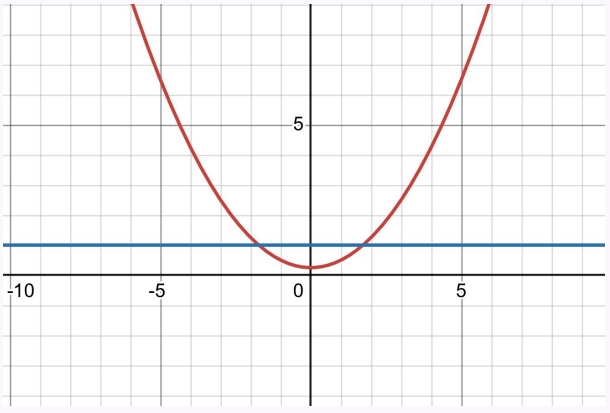
因此 $\hat{\theta}_1$ dominates $\hat{\theta}_c$, $\forall c > 1$,
因此根据定义, $\hat{\theta}_c$ 为 inadmissible

Question 4:

当 $c = \frac{1}{2}$ 时, 有:

$$R\left(\theta, \hat{\theta}_{\frac{1}{2}}\right) = \left(\frac{1}{2} - 1\right)^2\theta^2 + \left(\frac{1}{2}\right)^2 = \frac{\theta^2 + 1}{4}$$

分别作出 $R\left(\theta, \hat{\theta}_{\frac{1}{2}}\right)$ 和 $R(\theta, \hat{\theta}_1)$ 关于 θ 的图像:



Logic

第三种想法是: 利用 joint distribution 来同时 take expectation on x 和 θ

2.3 Bayes risk

2.3.1 Definition: Bayes risk

The **Bayes risk (average risk)** 被定义为:

$$\begin{aligned} r(f,\hat{\theta}) &= \mathbb{E}_{(\theta,X)}[L(\theta,\hat{\theta}(x))] \\ &= \int\int L(\theta,\hat{\theta}(x)) f(x,\theta)\, dx\, d\theta \end{aligned}$$

由于

$$f(x,\theta) = \underbrace{f(x|\theta)}_{\text{likelihood}} \cdot \underbrace{f(\theta)}_{\text{prior}} = \underbrace{f(\theta|x)}_{\text{posterior}} \cdot \underbrace{f(x)}_{\text{难求}}$$

因此 Bayes risk 也可以用 posterior risk 和 (frequentist) risk 来表示:

$$\begin{aligned} r(f,\hat{\theta}) &= \mathbb{E}_{\theta}[R(\theta,\hat{\theta})] \\ &= \mathbb{E}_{\theta}[\mathbb{E}_{X|\theta}[L(\theta,\hat{\theta}(X))]] \end{aligned}$$

$$\begin{aligned} r(f,\hat{\theta}) &= \mathbb{E}_{X}[r(\hat{\theta}|X] \\ &= \mathbb{E}_{X}[\mathbb{E}_{\theta|X}[L(\theta,\hat{\theta}(X))]] \end{aligned}$$

> ⚠ **Remark: STA3020 中 Bayes risk 的定义 (补充)**
>
> 令:
>
> 1. loss function 为 $\mathcal{L}(\theta,\delta)$
> 2. risk function 为 $R(\theta,\delta) = \mathbb{E}_{\theta}[\mathcal{L}(\theta,\delta)]$
> 3. prior distribution 为 $\Lambda(\theta|\lambda)$
>
> 则 **average risk** 被定义为:
>
> $$\begin{aligned} r(\Lambda,\delta) &= \mathbb{E}[R(\theta,\delta)] \\ &= \int_{\Theta} R(\theta,\delta) d\Lambda(\theta|\lambda) \\ &= \int_{\Theta}\int_{\mathcal{X}} \mathcal{L}(\theta,\delta)\, dF(x|\theta)\, d\Lambda(\theta|\lambda) \end{aligned}$$

> ⚠ **Remark**
>
> 对 Bayes risk 的理解:
>
> - average over both θ and X
> - 取决于 $\hat{\theta}$ 的形式

2.3.2 Definition: Bayes rule (Bayes estimator)

一个 decision rule (estimator) $\hat{\theta}$ 被称为 Bayes rule, 若其 **minimizes the Bayes risk**, 即

$$r(f,\hat{\theta}) = \inf_{\tilde{\theta}}\ r(f,\tilde{\theta})$$

> ☰ **Example**
>
> ***问题:***
>
> 令:
>
> - $X \sim \mathcal{N}(\theta,1)$
> - 使用 squared error loss
> - estimator 为 $\hat{\theta}_c(x) = cx$
>
> 1. 若 prior 为 $\theta \sim \mathcal{N}(0,\tau^2)$, 求出 Bayes risk

2. 求出 the Bayes rule among estimators $\hat{\theta}_c$
3. 求出这个 estimator 的 Bayes risk

解答:

Question 1:

在上个例子中, 我们已经求得了 frequentist risk:

$$R(\theta, \hat{\theta}_c) = (c-1)^2\theta^2 + c^2$$

因此 Bayes risk 为:

$$r(f, \hat{\theta}_c) = \mathbb{E}_\theta[R(\theta, \hat{\theta}_c)] = (c-1)^2\mathbb{E}_\theta[\theta^2] + c^2 = (c-1)^2\tau^2 + c^2$$

Question 2:

整理 Bayes risk 的式子:

$$\begin{aligned} r(f, \hat{\theta}_c) &= (\tau^2+1)c^2 - 2\tau^2 c + \tau^2 \\ &= (\tau^2+1)\left(c - \frac{\tau^2}{\tau^2+1}\right)^2 - \frac{\tau^4}{\tau^2+1} + \tau^2 \end{aligned}$$

当 $c = \frac{\tau^2}{\tau^2+1}$ 时, Bayes risk 最小, 因此 the Bayes rule 为:

$$\hat{\theta}_c(X) = \frac{\tau^2}{\tau^2+1}X$$

Question 3:

代入 $c = \frac{\tau^2}{\tau^2+1}$, 有:

$$r(f, \hat{\theta}_c) = \frac{\tau^2}{\tau^2+1}$$

2.3.3 使用 posterior risk 求解 Bayes rule

若我们根据以下方法定义 estimator $\hat{\theta}(x)$:

对于任意 x, 令 $\hat{\theta}(x)$ 为 **the value of $\hat{\theta}$ that posterior risk minimizes $r(\hat{\theta}|x)$** (即 estimator 对于任意 x 均能 minimizes posterior risk)

则 $\hat{\theta}(x)$ 为 Bayes estimator

⚠ **Remark: STA3020 中的表述 (补充)**

3. **Theorem:** 一种求 Bayesian estimator 的方法 (可简化计算)
   令 ① $\theta \in \Theta$ 有 prior distribution $\Lambda(\theta|\lambda)$, 其中 hyperparameter λ 已知.
   ② 当 $\theta = \theta_0$ 时, X 的分布为 $F_{\theta_0} = f(x|\theta_0)$.
   考虑通过非负的 loss function $L(g(\theta), \delta)$ 来估计 $g(\theta)$.
   若 ① 存在 estimator δ_0 with finite risk
   ② 对 almost all x, $\delta_\Lambda(x)$ 能 minimize
   $E[L(g(\theta), \delta(x)) | X = x]$ (先对 θ 取 integral)
   则 $\delta_\Lambda(x)$ 为 Bayesian estimator

⚡ **Proof (补充)**

证明:

若 estimator δ 有 finite risk，则

$$r(\Lambda,\delta) = E^{\theta}E^{X}[L(g(\theta),\delta(x))] < \infty \quad (\text{上标 } X \text{ 表示对 } X \text{ 取期望})$$

$$\Rightarrow r(\Lambda,\delta) = E^{X}E^{\theta}[L(g(\theta),\delta(x))] < \infty$$

$$\Rightarrow E^{\theta}[L(g(\theta),\delta(x))] < \infty \quad a.e. \quad (\text{由于 } L(\cdot) \text{ 非负})$$

根据 $\delta_\Lambda(x)$ 的定义：

$$E^{\theta}[L(g(\theta),\delta(x)) \mid X=x] \geqslant E^{\theta}[L(g(\theta),\delta_\Lambda(x)) \mid X=x] \quad a.e.$$

$$\Rightarrow E^{X}E^{\theta}[L(g(\theta),\delta(x))] \geqslant E^{X}E^{\theta}[L(g(\theta),\delta_\Lambda(x))]$$

$$\Rightarrow r(\Lambda,\delta) \geqslant r(\Lambda,\delta_\Lambda)$$

Example: STA3020 中的例子 (补充)

例 3: (Binary classification)

令 random sample $X_1,\cdots,X_n$ 取自 f_0 或 f_1. 即 $X_i \overset{ind}{\sim} f_{\theta_i}$，$\theta_i \in \Theta=\{0,1\}$，

令每个 θ_i 的 prior 为 $Ber(p)$.

求 0-1 loss $L(\theta_i,\delta_i)=1\{\delta_i \neq \theta_i\}$ 对应的 θ_i 的 Bayesian estimator

---

(Step 1: 求出 posterior distribution)

$$\pi(\theta_i|x) \propto f_{\theta_i}(x_i)\cdot p^{\theta_i}(1-p)^{1-\theta_i}$$

$$= f_0(x_i)\cdot(1-p)\cdot 1\{\theta_i=0\} + f_1(x_i)\cdot p\cdot 1\{\theta_i=1\}$$

(Step 2: 写出 risk，求出 minima)

$$r(\theta_i,\delta_i) = E^{\theta_i}E^{X_i}[1\{\delta_i \neq \theta_i\}]$$

$$\Rightarrow \text{根据 theorem，仅需 minimize } E[1\{\delta_i \neq \theta_i\}|x]$$

$$E[1\{\delta_i \neq \theta_i\}|x] = f_0(x_i)\cdot(1-p)\cdot 1\{\delta_i \neq 0\} + f_1(x_i)\cdot p\cdot 1\{\delta_i \neq 1\}$$

$$\Rightarrow \begin{cases} \text{若 } f_0(x_i)\cdot(1-p) \geqslant f_1(x_i)\cdot p, \text{ 选取 } \delta_i=0 \\ \text{若 } f_0(x_i)\cdot(1-p) < f_1(x_i)\cdot p, \text{ 选取 } \delta_i=1 \end{cases} \quad (\delta_i \text{ 也可取其他值，但为了有意义，令 } \delta_i \in \{0,1\})$$

$$\Rightarrow \delta_i = 1\left\{\frac{f_1(x_i)\cdot p}{f_0(x_i)\cdot(1-p)} \geqslant 1\right\}$$