STAT201B Lecture 17-18 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

2.1.2 Definition: posterior rule (posterior estimator)

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

$$
R(\hat{\theta}|x) = \inf_{\tilde{\theta}} R(\tilde{\theta}|x)
$$

> ☰ **Example**
>
> ***问题:***
>
> 令:
>
> - $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$
> - prior distribution 为 $\theta \sim \mathcal{N}(0, \tau^2)$
>
> 求 quadratic loss 对应的 posterior estimator
>
> ***解答:***
>
> 首先计算 posterior distribution:
>
> $$
> \begin{aligned}
> f(\theta|X_1, \dots, X_n) &\propto \exp\left\{-\frac{\sum_{i=1}^n (X_i - \theta)^2}{2}\right\} \cdot \exp\left\{-\frac{\theta^2}{2\tau^2}\right\} \\
> &\propto \exp\left\{-\frac{(n\tau^2+1)\left(\theta - \frac{n\tau^2}{n\tau^2+1}\bar{X}_n\right)^2}{2\tau^2}\right\}
> \end{aligned}
> $$
>
> 上式为 $\mathcal{N}\left(\frac{n\tau^2}{n\tau^2+1}\bar{X}_n, \frac{\tau^2}{n\tau^2+1}\right)$ 的 kernel, 因此
>
> $$
> \theta|X_1, \dots, X_n \sim \mathcal{N}\left(\frac{n\tau^2}{n\tau^2+1}\bar{X}_n, \frac{\tau^2}{n\tau^2+1}\right)
> $$
>
> 因此 quadratic loss 对应的 posterior risk 为:
>
> $$
> \begin{aligned}
> r(\hat{\theta}|X) &= \mathbb{E}_{\theta|X}[(\theta - \hat{\theta})^2] \\
> &= \mathbb{E}_{\theta|X}[\theta^2 - 2\theta\hat{\theta} + \hat{\theta}^2] \\
> &= \mathbb{E}_{\theta|X}[\theta^2] - 2\hat{\theta}\mathbb{E}_{\theta|X}[\theta] + \hat{\theta}^2 \\
> &= \left(\frac{n\tau^2}{n\tau^2+1}\bar{X}_n\right)^2 + \frac{\tau^2}{n\tau^2+1} - 2\hat{\theta} \cdot \frac{n\tau^2}{n\tau^2+1}\bar{X}_n + \hat{\theta}^2 \\
> &= \left(\hat{\theta} - \frac{n\tau^2}{n\tau^2+1}\bar{X}_n\right)^2 + \frac{\tau^2}{n\tau^2+1}
> \end{aligned}
> $$
>
> 因此 posterior estimator 为
>
> $$
> \hat{\theta} = \frac{n\tau^2}{n\tau^2+1}\bar{X}_n
> $$

Remark

对于 quadratic loss, posterior estimator 等于 posterior mean, 这同时也是 Bayes estimator (类似的结论适用于大部分 loss)

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 posterior risk
2. 求出 posterior rule (estimator) 及其对应的 posterior risk

解答:

Question 1:

在上个例子中, 我们已经求出了 posterior distribution:

$$\theta|X \sim \mathcal{N}\left(\frac{n\tau^2}{n\tau^2+1}X, \frac{\tau^2}{n\tau^2+1}\right)$$

因此 posterior risk 为:

$$\begin{aligned} r(\hat{\theta}_c|x) &= \mathbb{E}_{\theta|x}[(\theta - cx)^2] \\ &= \mathbb{E}_{\theta|x}[\theta^2] - 2cx\mathbb{E}_{\theta|x}[\theta] + c^2x^2 \\ &= \left(\frac{n\tau^2}{n\tau^2+1}x\right)^2 + \frac{\tau^2}{n\tau^2+1} - 2cx \cdot \frac{n\tau^2}{n\tau^2+1}x + c^2x^2 \\ &= \frac{\tau^2}{\tau^2+1} + x^2 \cdot \left(c - \frac{\tau^2}{\tau^2+1}\right) \end{aligned}$$

Question 2:

Minimize posterior risk, 则 posterior rule (estimator) 为:

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

其对应的 posterior risk 为 $\frac{\tau^2}{\tau^2+1}$

Remark

由于 $\hat{\theta}_c$ 对于任意 x 均能 minimize posterior risk, 因此其也为 Bayes rule (会在后面展开)

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

> ☰ **Example: square loss 对应的 risk 为 MSE**

若使用 square loss, 则

$$
\begin{aligned}
R(\theta, \hat{\theta}) &= \mathbb{E}_{X|\theta}[L(\theta, \hat{\theta})] \\
&= \mathbb{E}_{X|\theta}[(\theta - \hat{\theta})^2] \\
&= \mathbb{E}_{X|\theta}[\hat{\theta}^2] - 2\theta \mathbb{E}_{X|\theta}[\hat{\theta}] + \theta^2 \\
&= Var_{X|theta}[\hat{\theta}] + (\mathbb{E}_{X|\theta}[\hat{\theta}])^2 - 2\theta \mathbb{E}_{X|\theta}[\hat{\theta}] + \theta^2 \\
&= Var_{X|theta}[\hat{\theta}] + (\mathbb{E}_{X|\theta}[\hat{\theta}] - \theta)^2 \\
&= MSE(\hat{\theta})
\end{aligned}
$$

在这种情况下, 一个很好的 estimator 是 MLE, 因为 asymptotically, MLE 为 unbiased 且方差最小 (achieve Cramer-Rao lower bound), 因此 minimize MSE

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:

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

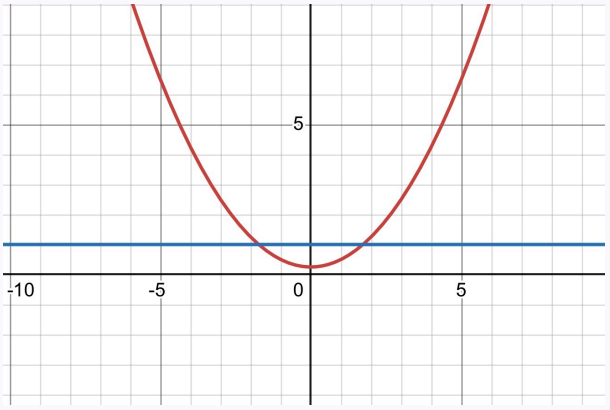
因此 $\hat{\theta}_1$ dominates $\hat{\theta}_c$, $\forall c > 1$,
因此根据定义, $\hat{\theta}_c$ 为 inadmissible

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:

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

Example: Bayes rule 的现实场景应用

问题:

The owner of a ski shop must order skis for the upcoming season. Orders must be placed in quantities of 25 pairs of skis. The cost per pair of skis is \$50 if 25 are ordered, \$45 if 50 are ordered, and \$40 if 75 are ordered. The skis will be sold at \$75 per pair. Any skis left over at the end of the year can be sold (for sure) at \$25 per pair. If the owner runs out of skis during the season, she will suffer a loss of "goodwill" among unsatisfied customers. She rates this loss at \$5 per unsatisfied customer. For simplicity, suppose the owner feels that demand for the skis will be 30, 40, 50, or 60 pairs of skis, with probabilities 0.2, 0.4, 0.2, and 0.2, respectively.

(a) Describe the parameter space Θ and the space of possible actions $\mathcal{A}$.

(b) What is the prior distribution?

(c) For each possible $\theta \in \Theta$ and $a \in \mathcal{A}$, compute the loss. (The loss in this case may be negative, representing a good outcome for the shop owner.) Display these possibilities in a matrix.

(d) What is the Bayes rule? That is, what action minimizes the Bayes risk? Note that in this example, there is no data, so the frequentist risk is the same as the loss.

解答:

Question 1:

(a) The parameter space is

$\Theta = \{30, 40, 50, 60\}$

where $\theta \in \Theta$ represents the demand for skis.

The possible actions can be represented as

$\mathcal{A} = \{25, 50, 75\}$

where $a \in \mathcal{A}$ represents the number of skis to order.

Question 2:

(b) The prior distribution is

$$f(\theta) = \begin{cases} 0.2 & \text{if } \theta = 30 \\ 0.4 & \text{if } \theta = 40 \\ 0.2 & \text{if } \theta = 50 \\ 0.2 & \text{if } \theta = 60 \\ 0 & \text{otherwise} \end{cases}$$

Question 3:

(c) For each possible θ and a, the loss can be calculated by

$$L(\theta,a) = c(a)\cdot a + 5(\theta-a)_+ - 75\cdot\min(\theta,a) - 25\cdot(a-\theta)_+$$

where $(x)_+ = \max\{x, 0\}$, and

$$c(a) = \begin{cases} 50 & \text{if } a=25 \\ 45 & \text{if } a=50 \\ 40 & \text{if } a=75 \end{cases}$$

Therefore, the loss can be calculated and summarized by

	$\theta=30$	40	50	60
$a=25$	-600	-550	-500	-450
50	-500	-1000	-1500	-1450
75	-375	-875	-1375	-1875

Question 4:

(d) Since there's no data, the frequentist risk is the loss, therefore, we can compute the Bayes risk for each action:

When $a=25$,

Bayes risk $= 0.2\times(-600) + 0.4\times(-550) + 0.2\times(-500) + 0.2\times(-450) = -530$

When $a=50$,

Bayes risk $= 0.2\times(-500) + 0.4\times(-1000) + 0.2\times(-1500) + 0.2\times(-1450) = -1090$

When $a=75$,

Bayes risk $= 0.2\times(-375) + 0.4\times(-875) + 0.2\times(-1375) + 0.2\times(-1875) = -1075$

Since $a=50$ minimizes the Bayes risk, the Bayes rule is $\hat{a}=50$

> ⚠ **Remark**
>
> 注意, 在这道题中, action 和 data 无关, 因此 risk 即为 loss:
>
> $$\begin{aligned} R(\theta, a(X)) &= \mathbb{E}_{X|\theta}[\mathcal{L}(\theta, a(X))] \\ &= \int \mathcal{L}(\theta,a) f(x|\theta)\, dx \\ &= \mathcal{L}(\theta,a)\int f(x|\theta)\, dx \\ &= \mathcal{L}(\theta,a) \end{aligned}$$

2.3.3 使用 posterior risk 求解 Bayes rule

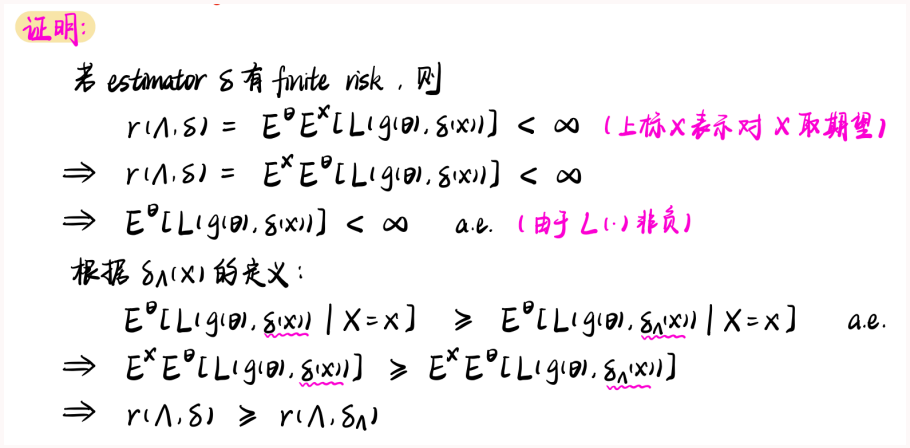
若我们根据以下方法定义 estimator $\hat{\theta}(x)$:

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)

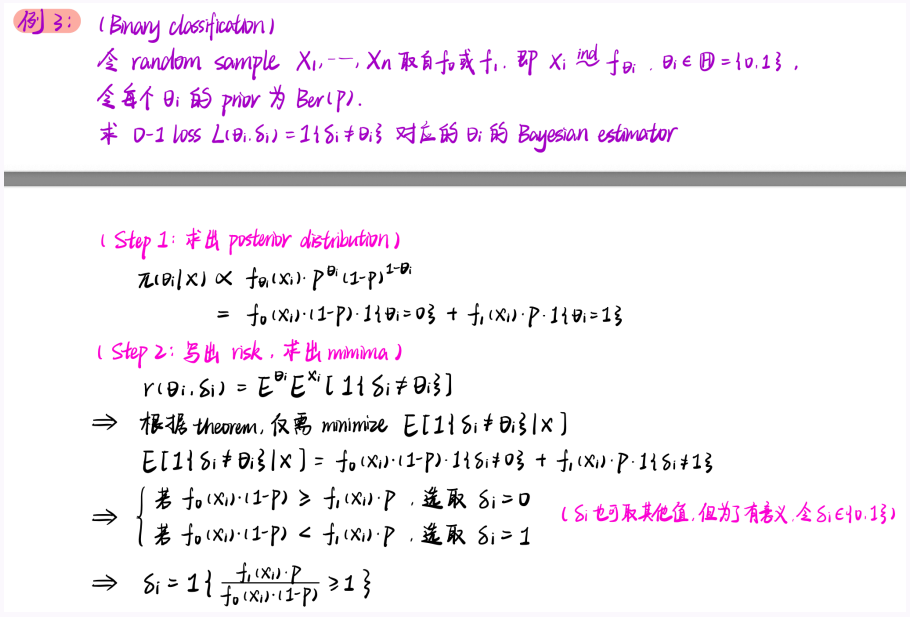
则 $\hat{\theta}(x)$ 为 Bayes estimator

> ⚠ **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 $L(g(\theta), s)$ 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
>
> $E[L(g(\theta), \delta(x)) \mid X=x]$ (先对 θ 取 integral)
>
> 则 $\delta_\Lambda(x)$ 为 Bayesian estimator

> **Proof (补充)**



Example: STA3020 中的例子 (补充)



2.3.4 特定 loss function 对应的 Bayes rule

在特定 conditions 下, 以下几种 standard loss functions 显式地对应了几种 Bayes rules:

1. 令 $\mathcal{L}(g(\theta),\delta)=(\delta-g(\theta))^2$ **(quadratic loss)**,
   则 $\delta_\Lambda(X)=\mathbb{E}[g(\theta)|X]$ **(posterior mean)**
2. 令 $\mathcal{L}(g(\theta),\delta)=\omega(\theta)(\delta-g(\theta))^2$ **(weighted quadratic loss)**,
   则 $\delta_\Lambda(X)=\frac{\int\omega(\theta)g(\theta)\cdot\pi(\theta|x)d\theta}{\int\omega(\theta)\cdot\pi(\theta|x)d\theta}$ ($\frac{\text{posterior weighted mean}}{\text{weight mean}}$)
3. 令 $\mathcal{L}(g(\theta),\delta)=|\delta-g(\theta)|$ **(absolute loss)**,
   则 $\delta_\Lambda(X)=\text{median}[g(\theta)|X]$ **(posterior median)**
4. 令 $\mathcal{L}(g(\theta),\delta)=\mathbf{1}(\delta\neq g(\theta))$ **(zero-one loss)**,
   则 $\delta_\Lambda(X)=\underset{a\in\mathcal{A}}{arg\,max}\,\mathbb{P}(g(\theta)=a|X)$ **(posterior mode)**

Proof: quadratic loss 对应的 Bayes rule

我们考虑使用 posterior risk 来求解 Bayes rule:

对于 quadratic loss, posterior risk 为:

$$
\begin{aligned}
r(\hat{\theta}, X) &= \mathbb{E}_{\theta|X}[\mathcal{L}(\theta, \hat{\theta})] \\
&= \mathbb{E}_{\theta|X}[(\theta - \hat{\theta})^2] \\
&= \mathbb{E}_{\theta|X}[\theta^2 - 2\theta \cdot \hat{\theta} + \hat{\theta}^2] \\
&= \mathbb{E}_{\theta|X}[\theta^2] - 2\mathbb{E}_{\theta|X}[\theta] \cdot \hat{\theta} + \hat{\theta}^2 \\
&= \mathrm{Var}(\theta \mid X) + \left(\hat{\theta}^2 - 2\mathbb{E}_{\theta|X}[\theta] \cdot \hat{\theta} + \{\mathbb{E}_{\theta|X}[\theta]\}^2\right) \\
&= \mathrm{Var}(\theta \mid X) + \left(\hat{\theta} - \mathbb{E}_{\theta|X}[\theta]\right)^2.
\end{aligned}
$$

因此 Bayes rule 为 $\mathbb{E}_{\theta|X}[\theta]$

Proof: weighted quadratic loss 对应的 Bayes rule (STA3020 中的证明) (补充)

证明: ① 和 ②

由上一个 theorem，对于 weighted quadratic loss，我们仅需研究 $\delta_\Lambda(X)$ who minimizes

$E^\theta[w(\theta)(\delta_\Lambda(X) - g(\theta))^2] = \delta_\Lambda(X)^2 \cdot E(w(\theta)|X) - 2\delta_\Lambda(X) E(w(\theta) \cdot g(\theta)|X) + E(w(\theta) \cdot g^2(\theta)|X)$

注意到为关于 $\delta_\Lambda(X)$ 的 quadratic function，因此

$$\delta_\Lambda(X) = \frac{E[w(\theta) g(\theta|x)]}{E[w(\theta)|X]}$$

若 $w(\theta) \equiv 1$，则

$$\delta_\Lambda(X) = E(g(\theta)|X)$$

Proof: absolute loss 对应的 Bayes rule (STA3020 中的证明) (补充)

证明: ③

思路：证明在 minimize $E[|\delta_\Lambda(X) - g(\theta)| \mid X = X]$ 层面，posterior median ξ 要优于其他 $\forall a \neq \xi$

Meanwhile, for absolute loss, we focus on $\delta_\Lambda(x)$ who minimizes

$$\mathbb{E}\left(|\delta_\Lambda(x) - g(\theta)| \, |x\right). \tag{2.5}$$

Define ξ to be any median of the conditional distribution of $g(\theta)$ given $X = x$, in other words, we have

$$\mathbb{P}(g(\theta) > \xi|x) = 1 - \mathbb{P}(g(\theta) \leq \xi|x) \leq \frac{1}{2}. \quad \text{(median 的定义)}$$

Notice that for $\forall a \leq \xi$, （先考虑 $a \leq \xi$ 的情况）

$$\begin{aligned}
\frac{1}{2}\mathbb{E}\left(|g(\theta) - a||x\right) &= \frac{1}{2}\left[\mathbb{E}\left((g(\theta) - a) \cdot \mathbb{1}(g(\theta) > a)|x\right)\right. \\
&\quad \left. - \mathbb{E}\left((g(\theta) - a) \cdot \mathbb{1}(g(\theta) \leq a)|x\right)\right] \\
&= \mathbb{E}\left((g(\theta) - a) \cdot \left[\mathbb{1}(g(\theta) > a) - \frac{1}{2}\right]|x\right). \qquad (2.6)
\end{aligned}$$

Since $a \leq \xi$, utilize form (2.6), we have

$$\begin{aligned}
\frac{1}{2}\mathbb{E}\left(|g(\theta) - \xi||x\right) &= \mathbb{E}\left((g(\theta) - a) \cdot \left[\mathbb{1}(g(\theta) > \xi) - \frac{1}{2}\right]|x\right) \\
&\quad + (a - \xi) \cdot \left[\mathbb{P}(g(\theta) > \xi|x) - \frac{1}{2}\right] \\
&\leq \mathbb{E}\left((g(\theta) - a) \cdot \left[\mathbb{1}(g(\theta) > \xi) - \frac{1}{2}\right]|x\right).
\end{aligned}$$

therefore,

$$\begin{aligned}
&\frac{1}{2}\left[\mathbb{E}\left(|g(\theta) - a||x\right) - \mathbb{E}\left(|g(\theta) - \xi||x\right)\right] \\
\geq &\mathbb{E}\left((g(\theta) - a) \cdot \left[\mathbb{1}(g(\theta) > a) - \mathbb{1}(g(\theta) > \xi)\right]|x\right) \geq 0.
\end{aligned}$$

Similarly, for $\forall a \geq \xi$, we have （再考虑 $a \geq \xi$ 的情况）

$$\mathbb{P}(g(\theta) < \xi|x) = 1 - \mathbb{P}(g(\theta) \geq \xi|x) \leq \frac{1}{2},$$

and we have

$$\begin{aligned}
\frac{1}{2}\mathbb{E}\left(|g(\theta) - a||x\right) &= \mathbb{E}\left((g(\theta) - a) \cdot \left[\frac{1}{2} - \mathbb{1}(g(\theta) < a)\right]|x\right), \\
\frac{1}{2}\mathbb{E}\left(|g(\theta) - \xi||x\right) &\leq \mathbb{E}\left((g(\theta) - a) \cdot \left[\frac{1}{2} - \mathbb{1}(g(\theta) < \xi)\right]|x\right),
\end{aligned}$$

which leads to

$$\begin{aligned}
&\frac{1}{2}\left[\mathbb{E}\left(|g(\theta) - a||x\right) - \mathbb{E}\left(|g(\theta) - \xi||x\right)\right] \\
\geq &\mathbb{E}\left((g(\theta) - a) \cdot \left[\mathbb{1}(g(\theta) < \xi) - \mathbb{1}(g(\theta) < a)\right]|x\right) \geq 0.
\end{aligned}$$

As a summary, we have

$$\mathbb{E}\left(|g(\theta) - a||x\right) - \mathbb{E}\left(|g(\theta) - \xi||x\right) \geq 0$$

for arbitrary $a \in \mathbb{R}$, meaning that (2.5) is minimized by $\delta_\Lambda(x) = \xi$, i.e., the bayes estimator $\delta_\Lambda(X)$ should be arbitrary median of the conditional distribution of $g(\theta)$ given X. □

Proof: zero-one loss 对应的 Bayes rule

我们考虑使用 posterior risk 来求解 Bayes rule:

对于 zero-one loss, posterior risk 为:

$$
\begin{aligned}
r(\hat{\theta}, X) &= \mathbb{E}_{\theta|X}[\mathcal{L}(\theta, \hat{\theta})] \\
&= \mathbb{E}_{\theta|X}[0 \cdot \mathbf{1}(\theta = \hat{\theta}) + 1 \cdot \mathbf{1}(\theta \neq \hat{\theta})] \\
&= \mathbb{P}_{\theta|X}(\theta \neq \hat{\theta}) \\
&= 1 - \mathbb{P}_{\theta|X}(\theta = \hat{\theta})
\end{aligned}
$$

因此, 当 $\hat{\theta}$ 为 posterior mode 时, $r(\hat{\theta}, X)$ 被 minimized

Example: $\Theta = \{0, 1\}$ 时的情况

考虑 $\Theta = \{0, 1\}$ (binary classification) 时的情况, 此时 $\hat{\theta}(X)$ 的取值范围为: $\hat{\theta}(X) \in \{0, 1\}$, 若我们发现 $\mathbb{P}_{\theta|X}(\theta = 0) < \mathbb{P}_{\theta|X}(\theta = 1)$, 则我们选择 Bayes rule $\hat{\theta}(X) = 1$

Example

问题:

令:

- $X_1, \dots, X_n | \sigma^2 \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma^2)$, 其中 θ 已知
- σ^2 的 prior distribution 为 inverse gamma with parameter a and b, 即

$$
f(\sigma^2; a, b) = \frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} \exp\{-b/\sigma^2\}
$$

求:

1. σ^2 的 posterior distribution
2. σ^2 的 Bayes estimator under squared error loss
3. σ^2 的 Bayes estimator under absolute error loss
4. σ^2 的 Bayes estimator under zero-one loss

Remark: inverse gamma distribution 的性质

对于 distribution $InverseGamma(a, b)$, 有以下性质:

- mean 为 $b/(a-1)$ when $a > 1$
- mode 为 $b/(a+1)$
- median 没有 closed form

解答:

Question 1:

注意到 posterior distribution:

$$
\begin{aligned}
f(\sigma^2 | X_1, \dots, X_n) &\propto \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \cdot \exp\left\{-\frac{\sum_{i=1}^n (X_i - \theta)^2}{2\sigma^2}\right\} (\sigma^2)^{-a-1} \exp\left\{-\frac{b}{\sigma^2}\right\} \\
&\propto (\sigma^2)^{-\frac{n}{2}} \cdot (\sigma^2)^{-a-1} \cdot \exp\left\{-\frac{\frac{\sum_{i=1}^n (X_i - \theta)^2}{2} + b}{\sigma^2}\right\} \\
&= (\sigma^2)^{-(\frac{n}{2}+a)-1} \cdot \exp\left\{-\frac{\frac{\sum_{i=1}^n (X_i - \theta)^2}{2} + b}{\sigma^2}\right\}
\end{aligned}
$$

为 $Inverse - Gamma(a', b')$ 的 kernel, 其中

$$
\begin{aligned}
a' &= \frac{n}{2} + a \\
b' &= \frac{\sum_{i=1}^n (X_i - \theta)^2}{2} + b
\end{aligned}
$$

因此 $\sigma | X_1, \dots, X_n \sim Inverse - Gamma(a', b')$

Question 2:

Under squared error loss, Bayes rule 为 posterior mean, 即

$$\hat{\sigma^2} = \frac{b'}{a'-1}$$

Question 3:

Under absolute error loss, Bayes rule 为 posterior median, 但由于 Inverse-Gamma 分布的 median 没有 closed form, 此时需要使用 sampling method

Question 4:

Under zero-one loss, Bayes rule 为 posterior mode, 即

$$\hat{\sigma^2} = \frac{b'}{a'+1}$$

2.3.5 Bayes rule 的 admissibility

若以下 weak conditions 满足:

- $\Theta \subset \mathbb{R}$
- 对于任意 $\hat{\theta}$, $R(\theta, \hat{\theta})$ 均为关于 θ 的 continuous function
- prior density f 对 Θ 的任意 open subset 都 assigns positive probability
- Bayes rule $\hat{\theta}^f$ 有 finite Bayes risk

则 Bayes rule $\hat{\theta}^f$ 为 admissible

Proof

Remark

证明的大致思路为: 使用 argue by contradiction, 假设 Bayes rule 不 admissible, 那么另一个 rule 会有 lower risk, 那么这个 rule 也会有更小的 Bayes risk, 形成矛盾 (不可能比 Bayes rule 的 Bayes risk 更小)

假设 Bayes rule $\hat{\theta}^f$ 不为 admissible, 则存在另一个 estimator $\tilde{\theta}$, 满足:

$$\begin{aligned} R(\theta, \tilde{\theta}) &\leq R(\theta, \hat{\theta}^f) \quad \forall \theta \in \Theta, \\ R(\theta, \tilde{\theta}) &< R(\theta, \hat{\theta}^f) \quad \text{for at least one } \theta^* \in \Theta \end{aligned}$$

由于 $R(\theta, \hat{\theta})$ 关于 θ continuous, 则存在一个 open set Ω ($\theta^* \in \Omega$), 使得以上 inequality 为 strict
因此有

$$\begin{aligned} r(f, \tilde{\theta}) &= \mathbb{E}_\theta[R(\theta, \tilde{\theta})] \\ &< \mathbb{E}_\theta[R(\theta, \hat{\theta}^f)] = r(f, \hat{\theta}^f) \end{aligned}$$

因此 $\hat{\theta}^f$ 不为 Bayes rule (contradiction)

故 Bayes rule $\hat{\theta}^f$ 为 admissible

Remark: STA3020 中的表述 (补充)

若一个 Bayes estimator 为 unique,
则该 estimator 为 admissible

Proof

Proof. Denote δ_Λ to be the bayes estimator corresponding to prior Λ. If there exists an estimator δ' such that $R(\theta, \delta') \leq R(\theta, \delta_\Lambda)$ for all $\theta \in \Theta$. Then,

$$r(\Lambda, \delta') = \int_\Theta R(\theta, \delta') d\Lambda(\theta) \leq \int_\Theta R(\theta, \delta_\Lambda) d\Lambda(\theta) = r(\Lambda, \delta_\Lambda)$$

But $\delta_\Lambda = \arg\min_\delta r(\Lambda, \delta)$ according to the definition of bayes estimator. Thus we have $r(\Lambda, \delta') = r(\Lambda, \delta_\Lambda)$, and since we require bayes estimator to be unique, therefore, we have $\delta' = \delta_\Lambda$ with probability 1, meaning that this bayes estimator δ_Λ is admissible. □

⚠ **Remark: Bayes estimator 的 uniqueness**

*** Theorem 3.6 (♣ Uniqueness of the Bayes Estimator).** *Denote Q to be the marginal distribution of X, i.e.,*

$$Q(x) = \int f(x|\theta) d\Lambda(\theta).$$

if the average risk $r(\Lambda, \delta)$ defined with respect to a strictly convex loss function is finite for some δ, and $f(\cdot|\theta)$ is absolutely continuous with respect to Q for all θ. Then the bayes estimator δ_Λ who minimize the average risk is unique.

A proof of Thorem.3.6 is provided in Lehmann and Casella (2006).