STA3000F Advanced theory of stats.

Wenlong Mou

mouwenlong.github.io/teaching/sta3000f24

Grading: 3 hw × 20% + 40% take-home final.

10:10 am – 1pm.

Decision theory (classical theory).

Asymptotic theory (n large, or $n \to +\infty$)

(a non-asymptotic approach).

nonparametric estimation.

Probability Recap.

$(\mathcal{X}, \mathcal{F}, \mu)$ $\qquad \mu : \mathcal{F} \to \mathbb{R}$.

space $\quad$ σ-field $\quad$ prob msr.

Def (Radon-Nikodym derivate)

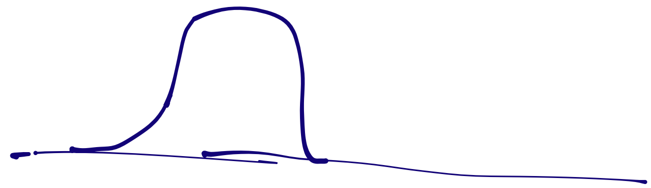
$\mu \ll \lambda$. (μ abs. cts w.r.t λ)

$\forall A \in \mathcal{F}$ $\qquad \lambda(A) = 0 \Longrightarrow \mu(A) = 0$.

$p(x) := \frac{d\mu}{d\lambda}(x)$ $\exists$.

$\forall A \in \mathcal{F}$. $\mu(A) = \int_A p(x)\, d\lambda(x)$.

λ : Leb



μ : density.

Markov ineq.

$X \geq 0$ $\cancel{\mathbb{E}[X] < +\infty}$

then $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$.

Proof:

$\mathbb{P}(X \geq a) = \mathbb{E}[1_{X \geq a}] \leq \mathbb{E}\left[\frac{X}{a} \cdot 1_{X \geq a}\right]$

$\leq \frac{\mathbb{E}[X]}{a}$.

Chebyshev ineq

Assume $\mathbb{E}_X[X^2] < +\infty$.

$Y = (X - \mathbb{E}[X])^2 \geq 0$

$$\mathbb{P}(|X-\mathbb{E}[X]| \geq a) = \mathbb{P}(Y \geq a^2) \leq \frac{\mathrm{var}(X)}{a^2}$$

Extension: $\mathbb{E}[|X|^p] < +\infty \quad (p \geq 2)$

$$\mathbb{P}(|X-\mathbb{E}(X)| \geq a) \leq \frac{\mathbb{E}[|X-\mathbb{E}X|^p]}{a^p}.$$

$$\mathbb{E}[e^{\lambda X}] =: m_X(\lambda).$$

$$m_X(\lambda) = \sum_{n=0}^{+\infty} \frac{\lambda^n}{n!} \mathbb{E}[X^n].$$

e.g. $X \sim \mathcal{N}(\mu, \sigma^2)$ $\quad \underline{m_X(\lambda) = \exp\left(\lambda\mu + \frac{\sigma^2\lambda^2}{2}\right)}$

e.g. $X \sim \mathrm{Ber}(p)$. $\quad m_X(\lambda) = 1-p+p\cdot e^{\lambda}.$

$\forall \lambda > 0.$

$$\mathbb{P}(X \geq \mathbb{E}(X)+a) \leq \exp(-(a+\mathbb{E}X)\cdot\lambda)\cdot m_X(\lambda).$$

e.g. $X \sim \mathcal{N}(\mu, \sigma^2)$

$a>0 \quad \mathbb{P}(X \geq \mu+a) \leq \exp(-(a+\mu)\cdot\lambda)\cdot\exp\left(\lambda\mu + \frac{\lambda^2\sigma^2}{2}\right)$

$$= \exp\left(-a\lambda + \frac{\lambda^2\sigma^2}{2}\right)$$

$$\left(\lambda = \frac{a}{\sigma^2}\right)$$

$$= \exp\left(-\frac{a^2}{2\sigma^2}\right).$$

$$\mathbb{P}(|X-\mu| \geq a) \leq 2\exp\left(-\frac{a^2}{2\sigma^2}\right).$$

e.g. $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Rade $\begin{cases} +1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2. \end{cases}$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \geq a\right).$$

$$= 2\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq a\right)$$

$$\left(Y = \sum_{i=1}^n X_i\right)$$

$$\leq 2\exp(-\lambda a n) \cdot m_Y(\lambda).$$

$$m_Y(\lambda) = \mathbb{E}\left[\exp\left(\lambda(X_1 + X_2 + \cdots X_n)\right)\right].$$

$$= \mathbb{E}\left[\exp(\lambda X_1)\right]^n$$

$$= \left(\frac{1}{2}e^{\lambda} + \frac{1}{2}e^{-\lambda}\right)^n.$$

$$\frac{1}{2}(e^{\lambda} + e^{-\lambda}) = 1 + \frac{\lambda^2}{2!} + \frac{\lambda^4}{4!} + \frac{\lambda^6}{6!} + \cdots$$

$$\exp\left(\frac{1}{2}\lambda^2\right) = 1 + \frac{\lambda^2}{2} + \frac{1}{2!}\cdot\left(\frac{\lambda^2}{2}\right)^2 + \frac{1}{3!}\cdot\left(\frac{\lambda^2}{2}\right)^3 + \cdots$$

$$\frac{\lambda^{2n}}{(2n)!} \le \frac{\lambda^{2n}}{n! \cdot 2^n}.$$

So $\frac{1}{2}(e^{\lambda} + e^{-\lambda}) \le e^{\frac{1}{2}\lambda^2}$

$$m_Y(\lambda) \le \exp\left(\frac{n\lambda^2}{2}\right).$$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \ge a\right)$$

$$\le 2 \cdot \exp(-\lambda a n) \cdot \exp\left(\frac{n\lambda^2}{2}\right)$$

$(\lambda = a)$

$$= 2 \cdot \exp(-na^2/2).$$

" Union bound ".

$$\mathbb{P}\left(\bigcup_{i \in I} A_i\right) \le \sum_{i \in I} \mathbb{P}(A_i).$$

e.g. — $X_1, X_2, \dots, X_n \sim N(0,1)$ marginally.

$$Y = \max_{1 \le i \le n} X_i$$

$$\mathbb{P}(Y \geq a) = \mathbb{P}(\exists i,\ X_i \geq a)$$
$$\leq \sum_{i=1}^{n} \mathbb{P}(X_i \geq a)$$
$$\leq n \cdot \exp\left(-\frac{a^2}{2}\right), \text{ (want } \leq \delta\text{).}$$

$$a = \sqrt{2\log(n/\delta)}.$$

$$\mathbb{P}\left(Y \geq \sqrt{2\log(n/\delta)}\right) \leq \delta.$$

Stats model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$

Observe $X \sim \mathbb{P}_{\theta^*}$

— Decision rule δ maps from X to $a \in \mathcal{A}$.

— Loss function

$$L(\theta, \delta(X)) \in \mathbb{R}.$$

— Risk. $R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta(X))]$.

$\mathbb{E}_\theta[\cdots\cdots]$: Expectation under $\mathbb{P}_\theta$.

e.g. $g: \Theta \to \mathbb{R}$. $\delta: \mathcal{X} \to \mathbb{R}$.

(Estimation) $L(\theta, a) = (g(\theta) - a)^2$.

$R(\theta, \delta) = \mathbb{E}_\theta (g(\theta) - \delta(X))^2$.

e.g. Testing. $\mathcal{A} = \{0, 1\}$

$\Theta_0 \subseteq \Theta$.

$$L(\theta, a) := \begin{cases} 1, & \theta \in \Theta_0, a = 1. \\ 1, & \theta \notin \Theta_0, a = 0. \\ 0, & \text{otherwise}. \end{cases}$$

$$R(\theta, \delta) = \begin{cases} \text{Type-I err}, & \theta \in \Theta_0 \\ \text{Type-II err}, & \theta \notin \Theta_0 \end{cases}$$

e.g. Statistical learning.

$X = \{(Z_i, Y_i)\}_{i=1}^n$.

$\mathcal{A}$: set of functions

from $\mathcal{Z}$ to $\{0, 1\}$.

$L(\theta, a) = \mathbb{P}_\theta(a(Z) \neq Y)$

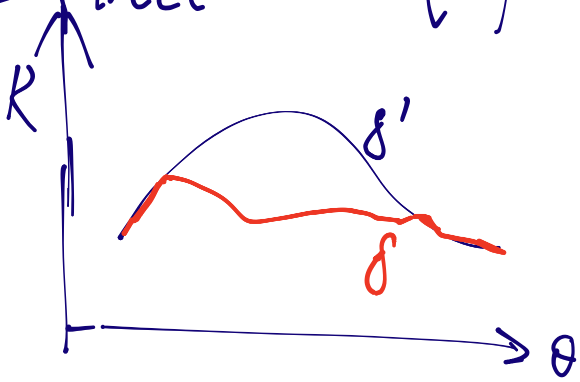
$R(\theta, \delta) = \mathbb{E}_\theta [L(\theta, \delta(X))]$

Criteria.

— Admissibility.

If $R(\theta; \delta) \leq R(\theta; \delta')$ $\forall \theta \in \Theta$

with strict ineq for some θ



e.g. $X \sim N(\theta, 1)$ $\delta(X) = 0$.

— Bayes risk. π : prior on Θ

$$r_\pi(\delta) = \int_\Theta R(\theta, \delta)\, \pi(d\theta)$$

$$\delta_{\text{Bayes}, \pi} = \arg\min \{ r_\pi(\delta) \}.$$

Suppose, $p_\theta(x) = \frac{d\mathbb{P}_\theta}{d\lambda}(x)$.

$$r_\pi(\delta) = \int_\Theta \int_{\mathcal{X}} L(\theta, \delta(x)) \cdot p_\theta(x)\, \lambda(dx) \cdot \pi(d\theta)$$

$$= \int_{\mathcal{X}} \left(\underbrace{\int_{\Theta} L(\theta, \delta(x)) \cdot \cancel{p_\theta(x)\, \pi(d\theta)}} \right) \lambda(dx)$$

Pointwise minimization.

$$\delta_{Bayes,\pi}(x) = \underset{a \in A}{\operatorname{argmin}} \left\{ \frac{\int_{\Theta} L(\theta, a)\, \cancel{p_\theta(x)\, \pi(d\theta)}}{\int_{\Theta} p_\theta(x)\, \pi(d\theta)} \right\}$$

$$\Pi(\cdot \mid x) := \frac{\pi(\cdot)\, p_{\cdot}(x)}{\int_{\Theta} p_\theta(x) \cdot \pi(d\theta)}.$$

$$\delta_{Bayes,\pi}(x) = \underset{a \in A}{\operatorname{argmin}}\ \mathbb{E}_{\Pi}\left[L(\theta, a)\right].$$

e.g. $L(\theta, a) = (a - g(\theta))^2$.

$$\delta_{Bayes,\pi} = \int_{\Theta} g(\theta)\, \Pi(d\theta \mid x).$$

Conjugate prior.

e.g. $X_1, X_2, \cdots X_n \overset{iid}{\sim} N(\theta, 1)$.

$$\pi = N(0, \tau^2)$$

$$\pi(\theta)\cdot p_\theta(X_1)\cdot p_\theta(X_2)\cdots p_\theta(X_n)$$

$$\propto \exp\left(-\frac{\theta^2}{2\tau^2}-\frac{(X_1-\theta)^2}{2}-\frac{(X_2-\theta)^2}{2}-\cdots-\frac{(X_n-\theta)^2}{2}\right)$$

$$\propto \exp\left(-\left(\frac{1}{2\tau^2}+\frac{n}{2}\right)\theta^2+\left(\sum_{i=1}^n X_i\right)\theta\right)$$

$$\pi(\theta|X_1^n)=N\left(\frac{\tau^2 n\bar{X}_n}{\tau^2 n+1},\ \frac{\tau^2}{\tau^2 n+1}\right)$$

Another e.g.

Beta-Bernoulli.

$\theta\sim \mathrm{Beta}(\alpha,\beta)$

$\pi(\theta)\propto\theta^{\alpha-1}\cdot(1-\theta)^{\beta-1}$ for $\theta\in[0,1]$.

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathrm{Ber}(\theta)$

$$\pi(\theta|X_1^n)=\mathrm{Beta}\left(\alpha+\sum_{i=1}^n X_i,\ \beta+\left(n-\sum_{i=1}^n X_i\right)\right).$$

— Minimax rule

minimize $\max_{\theta \in \Theta} R(\theta, \delta)$.

i.e. $\inf_{\delta} \sup_{\theta \in \Theta} R(\theta, \delta)$.

Relation between criteria.

— Unique Bayes rules are admissible.

Proof: If δ' satisfying $R(\theta, \delta') \le R(\theta, \delta)$ $(\forall \theta \in \Theta)$

then $r_\pi(\delta') \le r_\pi(\delta)$ δ' is also Bayes.

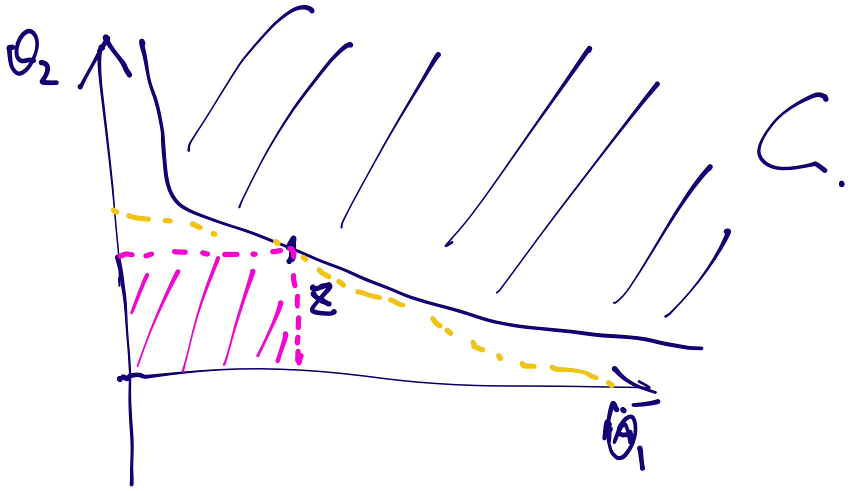
— Suppose Θ is finite.

All admissible rules are Bayes.

Proof: $\Theta = \{\theta_1, \theta_2, \cdots, \theta_K\}$.

$C := \{(R(\theta_j, \delta))_{j=1}^{K} : \delta$ is a rule$\}$

δ adm. risk vector $z \in \mathbb{R}^K$.



C is convex

separating hyperplane thm.

$\exists\ \lambda \in \mathbb{R}^k$ s.t. $x^T\lambda \begin{cases} \geq a & x \in C \\ \leq a & 0 \leq x \leq z \end{cases}$