STA 3000F Lecture 1

Instructor : Wenlong Mou.

Course website : mouwenlong.github.io/teaching/sta3000f25

Queraus: homework submission / Lecture recording.

Grading: 3 homeworks 20% each.

Take-home final 40%.

Poll: 1. 75 min Lecture 10 min break
75 min Lecture.
(end 10 min earlier).

2. 80 min 10 min 80 min (9, 50.

- Basic prob tools and review.
- Decision theory. (optimality, classical theory).
- Modern tools: empirical process theory (basics).
(application to asymptotics).

- Basics of nonparametric estimation.
  (including minimax lower bounds)

Probability review.

Radon-Nikodym derivative.

Idea: generalization of pdf.

μ : measure of interest

λ : base measure. (e.g. Lebesgue msr)

If $\mu \ll \lambda$ (i.e., $\forall A$ measurable, $\lambda(A) = 0 \Rightarrow \mu(A) = 0$)

then $\exists$ a density function.

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

For any measurable set A.

$$\mu(A) = \int_A \frac{d\mu}{d\lambda}(x) \cdot d\lambda(x).$$

Basic inequalities.

- Markov inequality.

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

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

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

Decaying only at $1/a$ rate.

Higher moment leads to better tail

Chebyshev inequality.

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

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

Proof: Let $Y = (X - \mathbb{E}[X])^2$.

Apply Markov ineq to Y.

Extension to higher moments.

If $\mathbb{E}[|X|^p] < +\infty$.

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

Larger p leads to faster-decaying tail behavior.

What if all moments exist for X?

Moment generating functions.

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

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

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

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

Fact (direct consequence of Markov ineq).

$\forall \lambda > 0$.

$$\mathbb{P}(X \geq \mathbb{E}[X] + t) = \mathbb{P}\left(e^{\lambda X} \geq e^{\lambda \mathbb{E}[X] + \lambda t}\right)$$

$$\text{(By Markov)} \leq \frac{m_X(\lambda)}{\exp(\lambda \mathbb{E}[X] + \lambda t)}.$$

(λ is up to our choice. optimize λ).

eg. for $X \sim N(\mu, \sigma^2)$.

$$\mathbb{P}(X \geq \mu + t) \leq \frac{\exp(\mu\lambda + \frac{\lambda^2\sigma^2}{2})}{\exp(\mu\lambda + \lambda t)}$$

$$= \exp(-\lambda t + \frac{\lambda^2\sigma^2}{2}).$$

(Take $\lambda = \frac{t}{\sigma^2}$)

$$\mathbb{P}(X \geq \mu + t) \leq \exp(\frac{-t^2}{2\sigma^2}).$$

(Exponentially decaying tail).

Def. X "sub Gaussian" random variable if.

$\forall \lambda$, $m_X(\lambda)$ exists; and.

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp(\frac{\lambda^2\sigma^2}{2}).$$

We call σ "the sub Gaussian parameter".

eg. Rademacher r.v. is sub-Gaussian.

$$X = \begin{cases} +1 & \text{w.p. } \frac{1}{2} \\ -1 & \text{w.p. } \frac{1}{2}. \end{cases}$$

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

$$= 1 + \frac{\lambda^2}{2!} + \frac{\lambda^4}{4!} + \frac{\lambda^6}{6!} + \cdots$$

On the other hand,

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

Each term:

$$\frac{1}{(2k)!} \quad \text{vs.} \quad \frac{1}{k!}\cdot\frac{1}{2^k}.$$

$$\leq.$$

So $m_X(\lambda) \leq \exp\left(\frac{\lambda^2}{2}\right) \quad (\forall \lambda)$.

X is. 1- sub Gaussian.

Moving to i.i.d. sums.

Suppose $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$.

$\mathbb{P}$ is σ- sub Gaussian.

How about $\frac{1}{n}\sum_{i=1}^{n} X_i$. (compared to $\mathbb{E}(X)$)?

For $\lambda \in \mathbb{R}$

$$\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}(X_i))\right)\right]$$

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

$$\overset{\text{(sub-Gaussian)}}{\leq} \left(\exp\left(\frac{\lambda^2\sigma^2}{2n^2}\right)\right)^n = \exp\left(\frac{\lambda^2\sigma^2}{2n}\right).$$

So $\frac{1}{n}\sum_{1}^{n} X_i$ is $\frac{\sigma}{\sqrt{n}}$-sub Gaussian.

Corollary. $\mathbb{P}\left(\left|\frac{1}{n}\sum_{1}^{n} X_i - \mathbb{E}[X]\right| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2\sigma^2}\right)$.

$(t>0)$

eg $X_i \overset{i.i.d}{\sim}$ Rademacher.

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{1}^{n} X_i\right|\right) \leq 2\exp\left(-\frac{nt^2}{2}\right).$$

Matching the tail for limiting Gaussian r.v.

This can be extended to general bounded r.v.

X supported on $[a, b]$. $(a<b)$

then X is sub-Gaussian w. parameter $\frac{b-a}{2}$.

Rademacher is the "worst-case" bounded r.v.

Corollary (Hoeffding bound).

$X_1, X_2 \cdots X_n$ i.i.d. $\quad X_i \in [a,b]$ a.s.

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right| \geq t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

(Chernoff bound).

e.g. $X \sim \text{Ber}(p)$.

$$\mathbb{P}\left(|\bar{X}_n - p| \geq t\right) \leq \exp\left(-2nt^2\right).$$

Remarks: independence is needed,

"identically distributed" can be relaxed.

"union bound".

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

Usually, we use A_i to model "bad thing at certain direction".

Tail behavior matters — # of terms.

e.g. Probabilistic method.

Consider Hamming cube $\{+1,-1\}^n$

$$d_{Ham}(x,y) = \sum_{i=1}^{n} 1_{x_i \neq y_i}.$$

Question: can we pack a lot of points in the Hamming cube, s.t. each one is far from each other.

i.e. Does there exist $\{x_1, x_2, \cdots x_M\} \subseteq \{+1,-1\}^n$.

s.t. $d_{Ham}(x_i, x_j) \geq \frac{1}{4} \cdot n$.

Fact: we can make M exponentially large

Proof. Let $X_1, X_2, \cdots X_M \overset{iid}{\sim} Unif(\{\pm 1\}^n)$.

$$\mathbb{P}\left(\underbrace{d_{Ham}(X_i, X_j)}_{\text{sum of } n \text{ iid Bernoulli r.v.'s.}} \leq \frac{n}{4} \right) \leq \exp\left(-2n \cdot \left(\frac{1}{4}\right)^2\right)$$

$$\leq \exp\left(-\frac{n}{8}\right).$$

$$\mathbb{P}\left(\exists i,j \in [M],\ d_{\text{Ham}}(X_i, X_j) \le \frac{n}{4}\right)$$

$$\le M^2 \cdot \mathbb{P}\left(d_{\text{Ham}}(X_1, X_j) \le \frac{n}{4}\right)$$

$$\le M^2 \cdot \exp\left(-\frac{n}{8}\right).$$

Let $M = \left\lfloor \frac{1}{2}\exp\left(\frac{n}{16}\right) \right\rfloor$.

$$\mathbb{P}(\text{bad event}) \le \frac{1}{4}.$$

Classical decision theory.

Key question: compare stats methods and prove optimality.

A statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$

Observe $X \sim P_{\theta^*}$ for some $\theta^* \in \Theta$.

Goal: to make decisions (about θ^*) based on data.

— A decision rule: mapping from data δ to action $a \in \mathcal{A}$.

— Loss function:

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

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

($\mathbb{E}_\theta$, $\mathbb{P}_\theta$ denote expectation, prob under θ).

e.g. estimation. $g : \Theta \to \mathbb{R}^k$ $\quad A = \mathbb{R}^k$.

$\hat{g} = \delta$: data $\to \mathbb{R}^k$.

$$L(\theta, a) = \|g(\theta) - a\|_2^2.$$

$$R(\theta, \hat{g}) = \mathbb{E}_\theta\left[\|g(\theta) - \hat{g}(X)\|_2^2\right].$$

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

$\Theta_0 \subseteq \Theta$.

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

$$R(\theta; \delta) = \begin{cases} \mathbb{P}(\text{type-I error}) & \theta \in \Theta_0 \\ \mathbb{P}(\text{type-II error}) & \theta \notin \Theta_0. \end{cases}$$

eg. Learning a binary classifier.

$\mathcal{A}$: the set of functions from input domain $\to \{0,1\}$.

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

$$L(\theta, f) = \mathbb{P}_\theta(f(z) \neq Y).$$

$(f \in \mathcal{A})$

$$R(\theta; \delta) = \mathbb{E}_\theta\left[L(\theta; \delta(X))\right].$$

(usually evaluated on a test set)

Comparing decision rules

Idea: $\min_{\delta}$ $R(\theta; \delta)$.

Multi-objective: there could be δ_1, δ_2 each performing better on a subset of models.

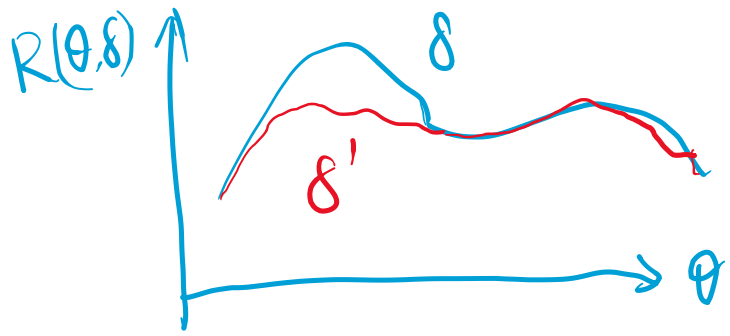
—— Admissibility.

If $\exists \delta'$ s.t. $R(\theta; \delta') \leq R(\theta; \delta)$ $(\forall \theta \in \Theta)$

with strict inequality for some θ.

then we call δ inadmissible.

A decision rule δ is called admissible if not inadmissible.



Problem: stupid decision rules can be admissible

e.g. $X_1, \dots X_n \overset{iid}{\sim} N(\theta, 1)$ $\theta \in \mathbb{R}$.

Goal: estimate θ.

$$\delta(X) = \theta.$$

$R(\theta, \delta) = 0$ if $\theta = 0$.

(But the risk is very large when $\theta \neq 0$).

Some "good" estimators are inadmissible.

"Stein's phenomenon":

$X_1, X_2, \dots, X_n \overset{iid}{\sim} N(\theta, I_d)$ $\theta \in \mathbb{R}^d$.

Goal: estimate θ under MSE Loss.

$$\delta(X_1, \cdots X_n) = \frac{1}{n}\sum_1^n X_i$$

Fact: δ is not admissible when $(d \geq 3)$.