36-705

11/06/2020

Youtube lecture 05/10/16

(*) note that you will have to 'splice' this lecture with Youtube lecture 03/10/16, as that lecture had 30 mins missing ~~at the~~ at the end. So there will be deductions/material in this that might feel unfamiliar. (*) ✓complete

Lecture notes 8 - Minimax Theory (cont.)

- define $X_1, \ldots, X_n \sim N(\theta, 1)$
- use L_2 loss: $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$

Theorem: $\hat{\theta} = \bar{X}$ is minimax

- cannot deal with this as have done previously; finding Bayes estimator ∴ does not work.
- LW: strategy:- find an upper and lower bound (separately) (which are very close together)
- we have:- $R_n = \inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta})$ (minimax risk)

- can we upper bound R_n?
- Recall $R_n = \inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta})$ - inf/'min' over all estimators of the maximum risk.
- select an arbitrary estimator $\hat{\theta}_*$; and ensure its maximum risk $\sup_{\theta} R(\theta, \hat{\theta}_*)$ is greater than $\inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta})$
- so we upper bound R_n by taking the maximum risk of some arbitrary estimator $\hat{\theta}_*$
- so $\forall$ any estimator $\hat{\theta}_*$ -

$$R_n \leq \sup_{\theta} R(\theta, \hat{\theta}_*)$$

- can we lower bound R_n?
- Pick any prior π and get the Bayes estimator $\hat{\theta}_B$

(*) compute its Bayes risk $B_\pi(\hat{\theta}_\pi)$ → this is a lower bound on R_n

- let's see why:

- minimax risk: $R_n = \inf_{\hat{\theta}} \underbrace{\sup_{\theta} R(\theta, \hat{\theta})}_{(i)}$ "min/inf over all estimators of maximum risk"

- (i) note maximum is always greater than the average,

- Hence $$R_n = \inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta}) \geq \inf_{\hat{\theta}} \int R(\theta, \hat{\theta}) \pi(\theta) d\theta = \inf_{\hat{\theta}} B_\pi(\hat{\theta}) = B_\pi(\hat{\theta}_\pi)$$

- note that $\inf_{\hat{\theta}} B_\pi(\hat{\theta})$ i.e. 'minimum'/infimum of the Bayes risk $B_\pi(\hat{\theta})$ for an est $\hat{\theta}$ over all estimators $\hat{\theta}$, is precisely the definition of the Bayes estimator

so, by definition, we have upper and lower bounds on the minimax risk:-

$$B_\pi(\hat{\theta}_\pi) \leq R_n \leq \sup_{\theta} R(\theta, \hat{\theta}_*)$$

LW: select an arbitrary Bayes estimator (which is a minimise of the Bayes risk), giving you a lower bound on minimax risk R_n.
- select an arbitrary estimator and take its max risk, to give you an upper bound.
- poor choices of $B_\pi(\hat{\theta}_\pi)$ and $\sup_{\theta} R(\theta, \hat{\theta}_*)$ will yield very little info on R_n (i.e. difference is very large).
- clever choices of $B_\pi(\hat{\theta}_\pi)$ and $\sup R(\theta, \hat{\theta}_*)$ i.e. close together can yield info on R_n.

(*) Apply this to the Normal case:-

- As an upper bound, select $\bar{X}$ as our arbitrary estimator.
- Risk of $\bar{X}$, under L_2 loss is MSE.
- MSE = Bias^2 + variance
- Bias = 0 ⇒ MSE = variance = $\frac{1}{n}$

- So we have upper bound $R_n \leq \frac{1}{n}$
- To get lower bound on R_n, select a Normal prior $\pi = N(0, c^2)$

LW: As an exercise, make sure you are acquainted with conjugacy (of likelihood, prior, post.)

- under this prior, we have a posterior Normal $p(\theta | x_1, \ldots, x_n) \sim N(a, b)$ for some a, b.
- We want Bayes estimator $\longrightarrow$ which under L_2 loss is posterior mean (convex combo of MLE, prior mean)

$$\hat{\theta}_\pi = \frac{nc^2 \bar{X}}{1 + nc^2}$$

- Now have to get Bayes risk of $\hat{\theta}_\pi$ (by getting risk and then Bayes risk).
- Risk of $\hat{\theta}_\pi$ under L_2 loss is MSE
- MSE = bias^2 + variance
- use these to assist:-

$$\text{Bias} = \mathbb{E}(\hat{\theta}_\pi) - \theta = \frac{nc^2 \mathbb{E}[\bar{X}]}{1 + nc^2} - \theta$$

$$= \frac{-\theta}{1 + nc^2}$$

- noting: only $\bar{X}$ is an r.v.
$\mathbb{E}[\bar{X}] = \theta$.
- for fut ref: bias in terms of sample size is $O(\frac{1}{n})$

(A1) exercise review (prior, post, like)
(A2) - check discrep against notes

$$\text{var}(\hat{\theta}_\pi) = \frac{n^2 c^4}{(1+nc^2)^2} \text{Var}[\bar{X}] = \frac{n^2 c^4}{(1+nc^2)^2} \cdot \frac{1}{n} = \frac{nc^4}{(1+nc^2)^2}$$

- The risk in this case is MSE = bias^2 + variance (L_2 loss)
- If not L_2 loss, will have to take expected value of loss to get risk.

$$R(\theta, \hat{\theta}_\pi) = \text{bias}^2 + \text{variance} = \frac{\theta^2 + nc^4}{(1+nc^2)^2}$$

- So we now have the risk; but we require Bayes risk

- Bayes risk of the Bayes estimator:-

$$B_\pi(\hat{\theta}_\pi) = \int \frac{\theta^2 + nc^4}{(1+nc^2)^2}\,\pi(\theta)\,d\theta \quad \sim ② \text{ can't see.}$$

- $\pi(\theta)$ is just a normal prior with mean 0, variance c^2
- we are just taking expected value under prior $\pi(\theta)$ $\frac{\theta^2+nc^4}{(1+nc^2)^2}$ and treating θ as a r.v. (that is Normal).

$$= \frac{\mathbb{E}_\pi[\theta^2] + nc^4}{(1+nc^2)^2}$$

- $\mathbb{E}_\pi[\theta^2] = c^2$
- 2nd moment of θ, with mean 0 is just variance $\pi \sim N(0, c^2)$

$$= \frac{c^2 + nc^4}{(1+nc^2)^2}$$

$$= \frac{c^2(1+nc^2)}{(1+nc^2)^2} = \frac{c^2}{(1+nc^2)}$$

- Hence, we have an upper and lower bound on R_n:-

$$\frac{c^2}{(1+nc^2)} \leq R_n \leq \frac{1}{n} \qquad \forall c$$

- As we selected arbitrary value of c (s.d. of prior on θ), above holds for $\forall c$
- so we can take limit of LHS:- $\lim_{c\to\infty}\left(\frac{c^2}{1+nc^2}\right) = \frac{1}{n}$ - can't see this limit artist ⑬⑦

- Hence $\frac{1}{n} \leq R_n \leq \frac{1}{n} \Rightarrow R_n = \frac{1}{n}$

(*) Finding Bayes estimator doesn't make, as we would need to find a prior that is uniform over the whole real line

(*) Hence we know <u>minimax risk</u> (in context) is $\frac{1}{n}$.

- To find minimax estimator that achieves the minimax risk, we have to find an estimator $\hat{\theta}$ whose maximum risk ($\sup_\theta R(\theta, \hat{\theta})$) is $\frac{1}{n}$. ~~That is~~

(*) If we take $\hat{\theta} = \bar{X}$; $\sup_\theta R(\theta, \hat{\theta}) = \frac{1}{n} = R_n$ (*)

(*) max risk does not depend on θ here, but we write sup anyway.

- So we have:-

1) Found minimax risk ($R_n = \frac{1}{n}$)

2) Found an estimator that achieves minimax risk ($\hat{\theta} = \bar{X}$)

- Hence $\bar{X}$ in this context is the <u>minimax estimator</u>

test prep for gen. case

LW: Details here make proof strategy more explicit than notes.

LW: Above strategy is contingent on <u>artful choices</u> of $B_\pi(\hat{\theta}_\pi)$ and $\sup R(\theta, \hat{\theta}_*)$.

1.5 maximum likelihood

(*) Parametric models that satisfy certain conditions (weak regularity);

$\hat{\theta}_{MLE}$ is <u>approximately minimax</u>.

- Making this precise → tricky (pioneered by Lucien LeCam)
  - beyond scope of this course.
- You will get a feeling for assumptions in parametric models for above to be true.
- This assumes that the dimension of θ is <u>fixed</u>, n is <u>increasing</u>
- fixed, finite dimensional parametric models with smoothness condition;
- $\hat{\theta}_{MLE}$ is approximately minimax.

(*) See "Asymptotic Statistics", van der Waart for <u>further exposition</u>

1.6 Hodges example

- Skip; give it a read.
- Related to Lecture Notes 9

36-705 11/06/2020

Youtube Lecture 05/10/16

Lecture Notes 9 - Asymptotic Theory

- Notation:- Recall:-

$$\mathbb{E}_\theta[g(X)] = \int g(x)p(x;\theta)\,dx$$

- DOES NOT MEAN θ is random (in this course)
- A reminder that the true distn is parametrised by θ.
- θ is fixed, not random (unknown)

- 1. Review of o and O

- These are included in notes for reference; should all be familiar and intuitive.

- 2. Distances between probability distributions

- Included for reference; not expected to memorise.
- Common distances:-

KL distance:- $K(P,Q) = \int p \log(p/q)$ mean

Hellinger: $h(P,Q) = \sqrt{\int (\sqrt{p} - \sqrt{q})^2}$

TW: we are typically interested in behaviour/properties of an estimator as $n \to \infty$ (i.e. we have more and more samples).

- We are concerned with:- For an est. $\hat{\theta}_n = g(X_1, \ldots, X_n)$

1) Consistency of an estim. : $\hat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$ i.e. $\hat{\theta}_n - \theta = o_p(1)$

i.e. estimator converges in probability to the true value of the parameter. (in particular ML estimator)

2) Asymptotic normality: under some conditions, a CLT for in particular ML estimators.

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \tau^2)$$

3. Consistency

- 2 modes of showing consistency; directly show via definition:-

Method 1 :- show that for all $\varepsilon > 0$ $\quad \hat{\theta}_n \xrightarrow{p} \theta \quad (\text{as } n \to \infty)$

$$P(|\hat{\theta}_n - \theta| \geq \varepsilon) \longrightarrow 0$$

Method 2 : Prove convergence in quadratic mean :- $\hat{\theta}_n \xrightarrow{q.m} \theta$

- $\mathbb{E}[(\hat{\theta}_n - \theta)^2] \longrightarrow 0$
- But in more upgraded vocab:-

$$MSE(\hat{\theta}_n) = Bias^2(\hat{\theta}_n) + Var(\hat{\theta}_n) \longrightarrow 0$$

- If Bias $\to 0$ and var $\to 0$ then $\hat{\theta}_n \xrightarrow{qm} \theta$ (estimator converges in q.m to param.)
- $\hat{\theta}_n \xrightarrow{q.m} \theta \Rightarrow \hat{\theta}_n \xrightarrow{p} \theta \Rightarrow$ estimator $\hat{\theta}_n$ is consistent

LW: Run through examples of consistent estimators and estimators which are not.

Example 1

- $X_1, \ldots, X_n \sim Ber(p)$
- $\hat{p}_{MLE} = \bar{X} \qquad \hat{p}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- $Bias = \mathbb{E}_p[\hat{p}_{MLE}] - p = 0$
- $variance = \frac{p(1-p)}{n} \to 0 \text{ as } n \to 0$
- Hence $MSE(\hat{p}_{MLE}) \longrightarrow 0$; and $\hat{p}_{MLE} \xrightarrow{q.m} p \Rightarrow \hat{p}_{MLE} \xrightarrow{p} p$

Suppose we want to compute log-odds $\psi = \log\left(\frac{p}{(1-p)}\right)$

- via equivariance: $\hat{\psi} = \log\left(\frac{\hat{p}_{MLE}}{(1-\hat{p}_{MLE})}\right)$ A

Q Is $\hat{\psi}$ a consistent estimator of ψ? YES; via continuous mapping theorem.

Example 2

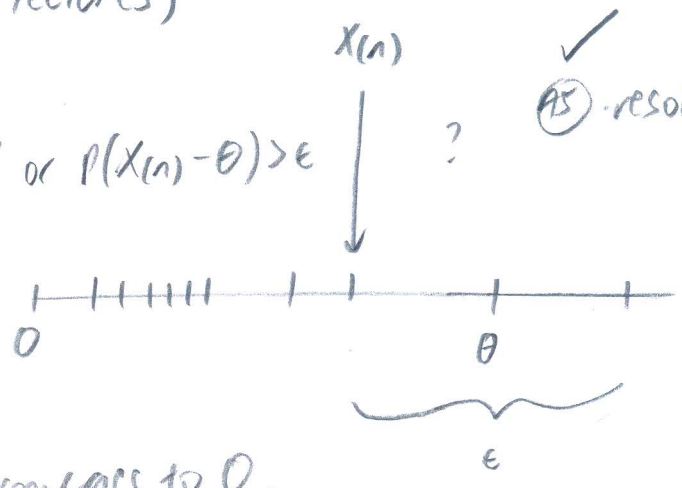
(A4) - review proof

$X_1, \ldots, X_n \sim \text{unif}(0, \theta)$

- $\hat{\theta}_{n,MLE} = X_{(n)} = \max\{X_1, \ldots, X_n\}$
- $\hat{\theta}_n \xrightarrow{p} \theta$ (proof directly on previous lectures)

(*) To refresh :-

- what is $P(\max\{X_1, \ldots, X_n\} - \theta > \epsilon)$? or $P(X_{(n)} - \theta) > \epsilon$



(A5) - resolve ✓

- probability of being outside interval size ϵ $\Rightarrow X_{(1)}, \ldots, X_{(n)}$ all of observations outside the interval are less than $(\theta - \epsilon)$.
- Can compute that probability; that it converges to 0.
- not hugely important - move on.

Hence: $P(|\hat{\theta}_n - \theta| > \epsilon) \longrightarrow 0 \quad \forall\, \epsilon > 0$

and $\hat{\theta}_{n,MLE} \xrightarrow{p} \theta$ i.e. $\hat{\theta}_{n,MLE}$ is consistent

Example 5

- Suppose I decide to measure everyone's blood pressure
- $\Rightarrow$ blood pressure is an r.v.
- we each have a ~~mean~~ blood pressure μ_i
- There is measurement noise from the machine σ^2
- Take 2 measurements per person

$Y_{11}, Y_{12} \sim N(\mu_1, \sigma^2)$

$Y_{21}, Y_{22} \sim N(\mu_2, \sigma^2)$

$\vdots$

$Y_{n1}, Y_{n2} \sim N(\mu_n, \sigma^2)$

(*) Find likelihood $L(\theta)$ which is just a large product

μ - assume blood pressure is fixed, unknown no. (as opposed to varying)

(*) $$\hat{\sigma}^2 = \sum_{i=1}^{n} \sum_{j=1}^{2} \frac{(Y_{ij} - \bar{Y}_i)^2}{2n}$$

$\bar{Y}_i$ - average of 2 measurements for each individual i.

(PS) - check calc of $\hat{\sigma}^2$ and $\xrightarrow{P}$ of $\hat{\sigma}^2$ (test prep).

- we can show:-

$$\hat{\sigma}^2 \xrightarrow{P} \frac{\sigma^2}{2}$$

(*) so $\hat{\sigma}^2$ is not consistent; but $2\hat{\sigma}^2$ is consistent

- some of issue:- no. of params increases with no. of observations (not par?)
  - an issue for maximum likelihood
  - dimension of parameter space increasing with n

LW: when is the maximum likelihood estimator consistent?

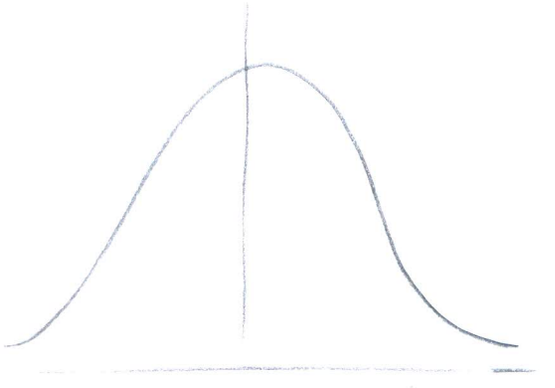
- Full answer → see can do weird
- Adhin in notes to give intuition on when ML estimator is consistent and what makes it consistent

LW: we need two regularity conditions on model

1) $\dim(\theta) = d$ is fixed

2) $p(x;\theta)$ is a smooth function of θ (*)

↳



- intuitively:- perturbing θ only informally shifts $p(x;\theta)$ a small amount

Theorem 6

- under regularity conditions on the model $\{p(x;\theta): \theta \in \Theta\}$, the maximum likelihood estimator is consistent

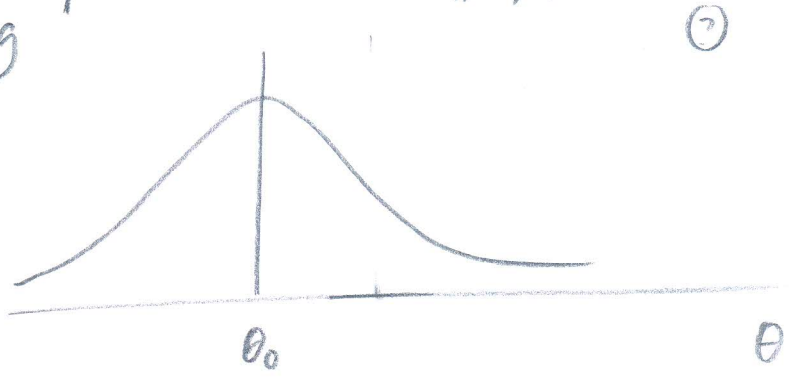
- Proof: needs probability tools not covered. But shows some intuition with other theorems.

LW: Theorem 3 is optional. Focus on Theorem 4: also reveals an affinity between maximum likelihood and KL-divergence

---

Theorem 4

- Assume we have a true value of the parameter θ_0 $X_1, \ldots, X_n \sim$
- generating data from distn using true param θ_0
- Intuitively, would like to show that likelihood at θ_0 is high; likelihood at all other points is small.



- select another arbitrary value of θ; compare the likelihoods:-

$$\frac{L(\theta_0)}{L(\theta)} > 1 \quad \leftarrow \text{this encodes above intuition.}$$

- switching to log-like form:-

$$l(\theta_0) - l(\theta) > 0$$

$$\Rightarrow \frac{1}{n}(l(\theta_0) - l(\theta)) > 0$$

- we now assess if this is positive:

$$\frac{1}{n}(l(\theta_0) - l(\theta)) = \frac{1}{n}\sum_{i=1}^{n} \log p(x_i;\theta_0) - \frac{1}{n}\sum_{i=1}^{n} \log p(x_i;\theta)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i;\theta_0)}{p(x_i;\theta)}$$

(*) Doing some variable renaming:-

$$\log \frac{p(X_i;\theta_0)}{p(X_i;\theta)} = Y_i$$

- We can view $\frac{1}{n}\sum_{i=1}^{n} \log \frac{p(X_i;\theta_0)}{p(X_i;\theta)} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

i.e.
- An average of IID r.v.s.
- We know via WLLN that $\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{P} \mu_Y$
- And that in this case:-

$$\mu_Y = \int \log \frac{p(x;\theta_0)}{p(x;\theta)}\, p(x;\theta_0)\, dx$$

(16) - Why using $p(x;\theta_0)$ as density?
↙ i.e. expectation not true data generating distri

(*) So:-

$$\frac{1}{n}(\ell(\theta_0) - \ell(\theta)) \xrightarrow{P} KL\left(p(x;\theta_0) \,\|\, p(x;\theta)\right) > 0$$

- This shows we need to assume identifiability (more regularity conditions)
- Two different values of param θ_1, θ_2 ; $\theta_1 \neq \theta_2$
  then $KL(\theta_1, \theta_2) > 0$
- i.e. we require that these be different densities/distinct.

(*) So $\frac{1}{n}(\ell(\theta_0) - \ell(\theta))$ is converging in probability to a strictly positive number

LN: This informs us that $L(\theta_0)/L(\theta) \xrightarrow{P} R$ where $R > 1$

- (*) So likelihood of true value of param θ_0 as $n \to \infty$, is always larger than the likelihood of another arbitrary value θ
- (*) Suggests that maximum likelihood (? / $\theta_{n,MLE}$?) est. should somehow be close to θ_0. (17)
- Making the deduction that $\hat{\theta}_{n,MLE}$ is consistent : $\hat{\theta}_{n,MLE} \to \theta_0$ requires more work on top of what we have shown.