36-705 06/06/2020

YouTube lecture 30/09/16

- Have seen following methods for constructing estimators
  1. MoM
  2. MLE
  3. Bayes

Example 8

A1: Derive this ✓

- $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ σ^2 is known
- Let $\mu \sim N(m, \tau^2)$ - m, τ^2 fixed unknown nos.
- following procedure from Bayes estimators:-

$$p(\mu|X_1,\dots,X_n) \propto p(X_1,\dots,X_n|\mu)p(\mu) \propto e^{-\frac{(\mu-m)^2}{2\tau^2}}$$

$$\propto e^{-\frac{(\mu-a)^2}{2b^2}}$$

$$\Rightarrow \hat{\mu} = \mathbb{E}[\mu|X] = \frac{\overset{(i)}{\tau^2}}{\underset{(ii)}{\tau^2 + \frac{\sigma^2}{n}}}\bar{X} + \frac{\sigma^2/n}{\tau^2 + \sigma^2/n} m$$

(i). prior variance
(ii). variance of $\bar{X}$

- convex combination of sample mean under MLE and prior mean

ⓧ: Mode of posterior, as part of Bayes estimator procedure, yields MAP (rather than mean in our case)

LW: we need a way of evaluating estimators. Focus on mean-squared-error as 1st step. Minimax theory supplements this as a formal way of evaluating estimator quality. And then large sample theory

MSE ⟶ minimax ⟶ large sample theory / asymptotics.

5. MSE

- Recall $\hat{\theta}$ is an r.v.
- Heuristically, MSE is mean of how far an estimator $\hat{\theta}$ deviates from its true value θ.
- $\mathbb{E}_\theta(\hat{\theta}-\theta)^2 = \int \cdots \int (\hat{\theta}(x_1,\dots,x_n)-\theta)^2 p(x_1;\theta)\dots p(x_n;\theta)\,dx_1\dots dx_n$

(*) Expectation is wrt joint distribution (assuming IID) (?) AI clarify
(*) Computationally → we don't evaluate this integral (or rarely)

- MSE = Bias^2 + Variance (I)
- Bias: $B = \mathbb{E}_\theta(\hat{\theta}) - \theta$ ("mean of an estimator minus true value")
- Variance: $V = Var_\theta(\hat{\theta})$

(*) Many problems in ML involve a trade-off between B and V

Theorem 9

$$MSE = B^2 + V$$

- add/subtract m

Proof: $MSE = \mathbb{E}_\theta(\hat{\theta}-\theta)^2 = \mathbb{E}_\theta(\hat{\theta}-m+m-\theta)^2$ where $m = \mathbb{E}_\theta(\hat{\theta})$

$$= \mathbb{E}_\theta(\hat{\theta}-m)^2 + (m-\theta)^2 + 2\mathbb{E}_\theta(\hat{\theta}-m)(m-\theta)$$

$$= \mathbb{E}_\theta(\hat{\theta}-m)^2 + (m-\theta)^2 + \underbrace{2(m-\theta)\mathbb{E}_\theta(\hat{\theta}-m)}_{=0 \text{ as } \mathbb{E}_\theta(\hat{\theta})=m}$$

$$= \mathbb{E}_\theta(\hat{\theta}-m)^2 + (m-\theta)^2$$

$$= V + B^2$$

LW: Some parametric estimators have 0 bias; then MSE = variance.

- lots of focus on unbiasedness in 1950s/60s, now combo of bias-variance is emphasised.

Example 10:

- Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$
- consider MLE estimates of μ and σ^2:-

$$\hat{\mu}_{MLE,MOM} = \bar{X}_n \qquad \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

- we make adjustment $S_n^{(\Lambda)} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ to ensure unbiasedness

?

$\Rightarrow E[S_n] = \sigma^2$

- LW: not entirely significant; made more for historical reasons.
- $MSE(\hat{\mu}) = \frac{\sigma^2}{n}$ (equal to variance as $\hat{\mu}$ is unbiased.)
- $MSE(\hat{\sigma}^2) = \mathbb{E}[(S^2 - \sigma^2)^2] = \frac{2\sigma^4}{n-1}$ (A3) - derivation

· note $MSE \to 0$ as $n \to \infty$ $MSE = O(\frac{1}{n})$ · note $MSE = f(\sigma^2)$

- characteristic of many significant parametric estimators

(*) for many non-parametric estimators cannot achieve this kind of convergence for MSE

LW: But, still doesn't quite give clear prescriptions on what estimators to select out of a subset of good estimators.

- As it contains an unknown parameter σ^2 (is a function of it)
- computing MSE is a first step

6. Best unbiased estimators

- Idea is that we restrict ourselves to unbiased estimators.
- Then we can answer question of which estimator has lowest variance

LW: Many theorems, textbooks on this

LW: His view is that these results are not particularly 'useful' (without qualification); but important historically; hence he does not emphasise

- Rao-Blackwell: For an unbiased estimator W, you can take $E[W|T]$ where T is a sufficient statistic; and this is still an estimator
- in the sense that an estimator can only depend on the data
- $E[W|T]$ is guaranteed to depend on the data and not on the parameter because conditioning on the sufficient statistic $\Rightarrow$ distn no longer depends on parameter θ.
- so $E[W|T]$ defines a new estimator which automatically gives us another estimator with a unilateral decrease in variance.

Lecture Notes 8 - Minimax Theory

- A theoretical construct to evaluate quality of estimators
- see a lot of NeurIPS conferences/papers with minimax
- covered in more detail in 36-702.
- This covers basic idea.

1. Minimax Theory

- More concretely
- Suppose we want to estimate a parameter $\underline{\theta}$ using data $X^n = (X_1, \ldots, X_n)$
- What is best possible estimator $\underline{\hat{\theta}} = \hat{\theta}(X_1, \ldots, X_n)$ of θ?
- Minimax $\rightarrow$ provides framework for answering.

1.1. Introduction

- What do we mean by a parameter estimator being close to the truth?
- Problem dependent $\rightarrow$ if you tell me what you mean by 'closeness', theory will assist.

- Let $\underline{\hat{\theta}} = \underline{\hat{\theta}}(X^n)$ be an estimator for the parameter $\underline{\theta} \in \Theta$ *

* - dropping vector notation.

- Define a loss function $L(\theta, \hat{\theta})$ that measures how good an estimator is. ×

× - (AI) - Notes have sep. not. for vectors.

(*) Can take many functional forms; MSE $\rightarrow$ squared error loss (squared distance as measure of loss)

Examples $\rightarrow$ e.g. $L(\theta,\hat{\theta}) = (\theta-\hat{\theta})^2$ - squared error loss

(scalar) $L(\theta,\hat{\theta}) = |\theta-\hat{\theta}|$ - abs.

$L(\theta,\hat{\theta}) = \mathbb{I}\{\theta \neq \hat{\theta}\}$ - zero-one loss

(*) Minimax theory is general; gives you an optimal estimator with respect to a loss function.

(*) Classification:- (zero-one loss)

- Predict $Y \in \{0,1\}$ $\quad L(Y,h(X)) = \mathbb{I}\{Y \neq h(X)\}$
- classifier $h(x)$

(*) Real-valued predic:-

$L(Y,\hat{Y}) = (Y-\hat{Y})^2$

- value of the loss function $L(\theta,\hat{\theta})$ is a random quantity, due to presence of $\hat{\theta}$.

(A2) - commit → rem. - (all alph.) ✓

The risk of an estimator $\hat{\theta}$:- (expected value of loss)

$$R(\theta,\hat{\theta}) = \mathbb{E}_\theta(L(\theta,\hat{\theta})) = \int L(\theta,\hat{\theta}(x_1,\dots,x_n))\, p(x_1,\dots,x_n;\theta)\, dx$$

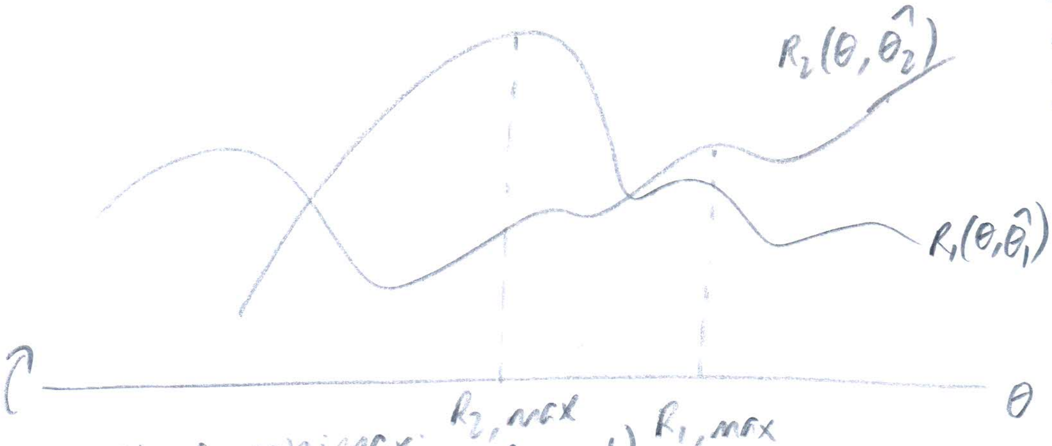
$$= \int\dots\int L(\theta,\hat{\theta}(x_1,\dots,x_n)\, p(x_1;\theta)\dots p(x_n;\theta)\, dx_1\dots dx_n$$

(*) Loss depends on data when we evaluate it, the risk is not, as we are integrating the data out; risk still depends on the parameter θ and is a function of θ

(*) For L_2 loss: the risk

$$L(\theta,\hat{\theta}) = (\theta-\theta)^2 \Rightarrow R(\theta,\hat{\theta}) = \mathbb{E}_\theta[(\theta-\hat{\theta})^2] = MSE$$

(E1): MSE is a ~~more~~ specific case of a more general concept within minimax



- If you give me an estimator, I can compute the risk $R(\theta, \hat{\theta})$, (function) but it is a function of param θ.
- still have an issue of selecting estimator from a set.

- motivation for minimax: (informal)
- 'Protect ourselves' from worst case, given we do not true value of θ (unknown param)
- look at $\max(R_i(\theta, \hat{\theta}_i))$ and say an estimator is 'better' if the maximum of corresponding risk is smaller than another estimator.
- i.e. for 2 estimators $\hat{\theta}_i$ and $\hat{\theta}_R$
- If $\max(R_i(\theta, \hat{\theta}_i)) < \max(R_R(\theta, \hat{\theta}_R)) \Rightarrow \hat{\theta}_i$ is a 'better estimator'

(A3) - Review concept. ✓

- formally:

(*) The minimax risk :-

$$R_n = \underbrace{\inf_{\hat{\theta}} \underbrace{\sup_{\theta} R(\theta, \hat{\theta})}_{(I)}}_{(II)}$$

or

(I) Take estimator $\hat{\theta}$, compute function $R(\theta, \hat{\theta})$ and find its maximum over its argument (the unknown parameter θ). ('max ≈ sup')

-(II) Find the smallest you can make quantity (I) over all possible estimators $\hat{\theta}$. ('min ≈ inf')

(*) In a certain sense, minimax risk is a quantification of how 'difficult' a problem is. (as the best you can do).

(*) An estimator $\hat{\theta}$ is minimax estimator if:-

$$\sup_{\theta} R(\theta, \hat{\theta}) = \inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta}) = R_n$$

(Q): Prior on θ? i.e. introduce averaging.

(A): Looking ahead. Integrating $R(\theta, \hat{\theta})$ gives you Bayes estimators

θ

(*) Aside:-

Parametric problems:- $R_n = O(\frac{1}{n})$ - risk goes to 0, but at what rate?

nonparametric -"- :- $R_n = O(\frac{1}{n^a})$ $a<1$

- in CS, we have sample complexity; how big a sample size do I need to ensure that risk $R_n < \varepsilon$
(theoretical)