Lecture 6 - scribbles

Friday, September 21, 2018 13:39

today: finish decision theory
gen/disc. learning

stat decision theory cted. :

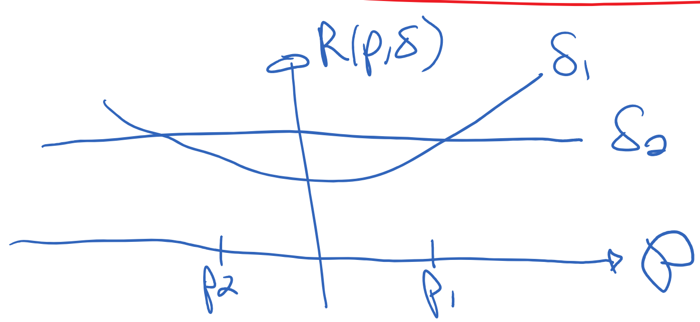
given previous framework, how can we compare stat procedures? e.g. δ_1 vs. δ_2

first property

$$\boxed{(\text{frequentist})\ \text{risk}\quad R(P,\delta) \triangleq \mathbb{E}_{D\sim P}\left[L(P,\delta(D))\right]}$$

$\delta_i : D \to \mathcal{A}$

"risk profiles"



* transform to scalar :

- "minimax" analysis : $\max_{p \in \Theta} R(p,\delta)$ "worst case"

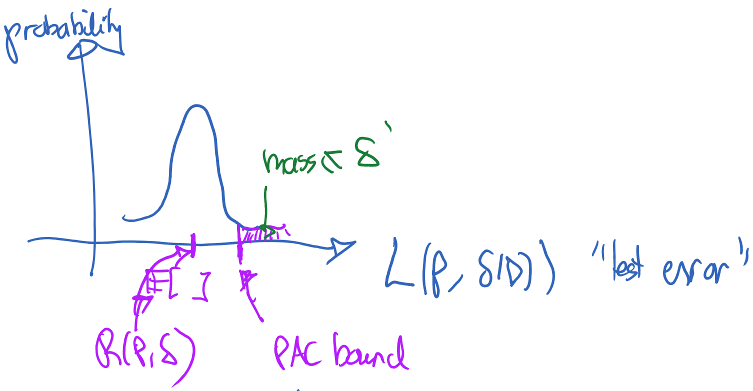
- weighted average : $\int_{\Theta} \underbrace{R(\theta,\delta)}_{\text{frequentist}} \pi(\theta)\, d\theta$

- weighted average : $\int_{\Theta} R(\theta, \delta)\, \pi(\theta)\, d\theta$

* alternatively: in ML theory, is PAC theory

→ look at tail bounds for dist. of $L(P, \delta(D))$ [D is random]

center



↳ $P\{ L(P, \delta(D)) \geq \underset{\substack{\uparrow \\ \text{bound}}}{\text{stuff}} \} \leq \underset{\substack{\uparrow \\ \text{small number}}}{\delta'}$

"with high prob" statement

Bayesian decision theory:

→ condition on data D

Bayesian posterior risk $\quad R_B(a \mid D) = \int_{\Theta} L(\theta, a)\, \underbrace{p(\theta \mid D)}_{\text{posterior} \,\propto\, p(\theta)\, p(D \mid \theta)}\, d\theta$

Bayesian optimal action : $\quad \delta_{Bayes}(D) \triangleq \underset{a \in \mathcal{A}}{\operatorname{argmin}}\; R_B(a \mid D)$

example: if $\mathcal{A} = \Theta$ ("estimation")

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

then (exercise) $\delta_{Bayes}(D) = \mathbb{E}[\theta \mid D]$ (posterior mean)

$$\begin{tikzcd}
 & L(\theta, \delta(D)) \arrow[dl, "\mathbb{E}_{D \sim p_\theta}"'] \arrow[dr, "\mathbb{E}_{\theta \sim p(\theta \mid D)}"] & \\
R(\theta, \delta) \arrow[dr, "\mathbb{E}_{\theta \sim \pi}"'] & & \text{Bayesian posterior risk} \arrow[dl, "\mathbb{E}_{D \sim p(D)}"] \\
 & \text{weighted summary} &
\end{tikzcd}$$

frequentist risk $R(\theta, \delta)$

$\mathbb{E}_{D \sim p(D)}$ ← marginal dist $\quad p(D) \triangleq \int_\theta p(\theta)\, p(D \mid \theta)\, d\theta$

Bayesian procedure δ_{Bayes} minimizes this when using a prior $p(\theta) = \pi(\theta)$ ← weighting fct.

Examples of estimators: $\delta : \mathcal{D} \to \Theta$

1) • MLE

2) • MAP

3) • method of moments

idea: find an injective mapping from Θ to "moments" $\mathbb{E}X$, $\mathbb{E}X^2$

....

and surjective on "possible moments"

and then invert it from empirical moments $\hat{\mathbb{E}}[X] \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i$

$\hat{\mathbb{E}}[X^2]$

...

example: for Gaussian $X \sim N(\mu, \sigma^2)$

$$\mathbb{E}[X] = \mu$$

$$\mathbb{E}[X^2] = \sigma^2 + \mu^2$$

$$f(\mu, \sigma^2) \triangleq \begin{pmatrix} \mathbb{E}[X] \\ \mathbb{E}[X^2] \end{pmatrix}$$

$$\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} \triangleq f^{-1}\begin{pmatrix} \hat{\mathbb{E}}[X] \\ \hat{\mathbb{E}}[X^2] \end{pmatrix}$$

(here, this estimator is same as MLE)
[property of exponential family]

(*) this is useful for latent variable models (e.g. mixture of Gaussians)
("spectral methods" e.g.)

4) in the context of prediction $\mathcal{A} = \{(f: \mathcal{X} \to \mathcal{Y})\}$ $\mathcal{X}$ input space
$\mathcal{Y}$ output "

example of $\delta: \mathcal{D} \to \mathcal{A}$

is using empirical "risk" minimization (ERM)

↳ "Vapnik risk" i.e. generalization error

recall: Vapnik risk:

$$L(p,f) = \mathbb{E}_{(X,Y)\sim p}[\ell(Y, f(X))]$$

replace with

$$\hat{\mathbb{E}}[\ell(Y,f(X))] = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$$

$$\hat{f}_{ERM} = \underset{f\in\mathcal{F}}{\arg\min}\ \hat{\mathbb{E}}[\ell(Y,f(X))]$$

← hypothesis class

14h34

bias-variance decomposition:

iid setting $\quad \mathcal{D} = p_\theta^{\otimes n} \quad\quad D_n = (X_i)_{i=1}^{n}$

$X_i \overset{iid}{\sim} p_\theta$

notation: $\hat{\theta}_n = \delta_n(D_n)$

to highlight dependence on n

study $R(\theta, \delta_n)$ as a function of n

in particular, would like $R(\theta,\delta_n) \xrightarrow{n\to\infty} R(\theta, \delta_\theta^*)$

in particular, would like $R(\theta, \delta_n) \to R(\theta, \delta_\theta)$ "consistency"

$\delta_\theta = \underset{\delta}{\operatorname{argmin}}\, R(\theta, \delta)$

* for estimation, typical loss: squared loss $L(\theta, \delta_n(D)) = \|\theta - \delta_n(D)\|_2^2$

<u>standard statistical consistency</u> $\hat{\theta}_n \xrightarrow{P} \theta$ "in probability"

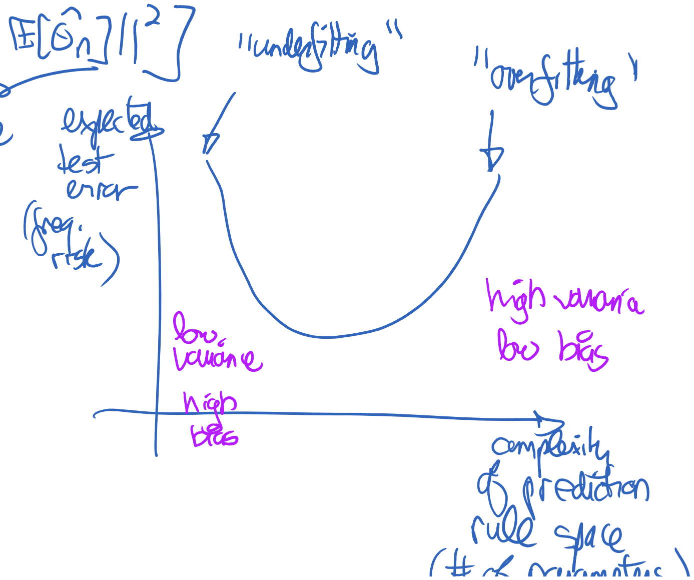
i.e. $\forall \varepsilon > 0,\ P\{\|\hat{\theta}_n - \theta\| > \varepsilon\} \xrightarrow{n\to\infty} 0$

randomness is from $\underline{D_n}$

$\hat{\theta}_n = \delta_n(D_n)$

* last time, $R(\theta, \delta_n) = \mathbb{E}_{D \sim P}\left[\|\theta - \hat{\theta}_n\|_2^2\right]$

$$= \underbrace{\|\theta - \mathbb{E}[\hat{\theta}_n]\|^2}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\|^2\right]}_{\text{variance}}$$



<u>James-Stein estimator</u>:

for estimating the mean of $N(\vec{\mu}, \sigma^2 I)$

δ_{JS} is <u>biased</u>, but lower variance than MLE

δ_{JS} actually strictly dominates δ_{MLE} for $d \geq 3$

δ_{JS} is biased, but lower variance than MLE

δ_{JS} actually strictly dominates δ_{MLE} for $d \geq 3$ (p; dimension($\vec{\mu}$))

(# prediction rule space (# of parameters))

i.e. $R(\theta, \delta_{JS}) \leq R(\theta, \delta_{MLE}) \quad \forall \theta$

and $\exists \theta$ s.t. $R(\theta, \delta_{JS}) < R(\theta, \delta_{MLE})$

$\Rightarrow$ MLE is sometimes inadmissible !?

---

in assignment, "consistency" mean $R(\theta, \delta_n) \xrightarrow{n\to\infty} 0$ $\quad$ ($\mathbb{E}\|\theta - \hat{\theta}_n\|^2 \xrightarrow{n\to\infty} 0$)

"convergence in $\mathcal{L}_2$"

[it turns out that $\mathcal{L}_2$-converges $\Rightarrow$ convergence in probability i.e. $\hat{\theta}_n \xrightarrow{P} \theta$]

by bias-variance decomposition:

$$\left.\begin{array}{l} \text{bias}(\delta_n) \xrightarrow{n\to\infty} 0 \\ \text{variance}(\delta_n) \xrightarrow{n\to\infty} 0 \end{array}\right\} \Rightarrow R(\theta, \delta_n) \to 0 \Rightarrow \text{consistency } \hat{\theta}_n \xrightarrow{P} \theta$$

properties (asymptotic) of MLE:

under regularity conditions on Θ & $p(x;\theta)$

a) $\hat{\theta}_n \xrightarrow{P} \theta$ "consistent"

b) CLT: (central limit theorem) $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \underbrace{I(\theta)}_{\text{information matrix}}^{-1})$ "in distribution" (iid data) theorem (iid data) information matrix

c) asymptotically optimal (Cramer-Rao lower bound)

i.e. it has minimal asymptotic variance among all "reasonable" estimators
↳ • consistent
• ...

d) invariance: MLE is preserved under reparameterization

suppose have bijection $f: \Theta \to \Theta'$

then $\widehat{f(\theta)} = f(\hat{\theta})$

* if not a bijection, can generalize MLE with "profile likelihood"

suppose $g: \Theta \to \Lambda$ profile likelihood $L(\eta) \triangleq \max_{\theta : \eta = g(\theta)} p(\text{data}; \theta)$, $\eta \in \Lambda$

define $\hat{\eta}_{MLE} \triangleq \arg\max_{\eta \in g(\Theta)} L(\eta)$

then we have $\boxed{\hat{\eta}_{MLE} = g(\hat{\theta}_{MLE})}$ "plug in" estimator

example: $\widehat{(\sigma^2)} = (\hat{\sigma})^2$

$\widehat{\sin \sigma^2} = \sin \hat{\sigma}^2$