Lecture 6 - properties of estimators

September 21, 2021 4:27 PM

today: statistical decision theory ~~of~~ properties of estimators

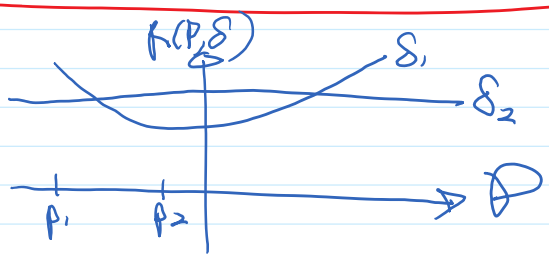
comparing procedures?

δ_1 vs. δ_2

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

(think cross-validation)

"risk profiles"



* transform to scalar
  - "minimax" analysis: $\max_{P \in \mathcal{P}} R(P, \delta)$ "worst case"
  - weighted average $\int_{\mathcal{H}} R(P_\theta, \delta)\, \pi(\theta)\, d\theta$ (kind of a Bayesian tool)
    - $\pi(\theta)$: weighting

PAC theory vs. frequentist risk:

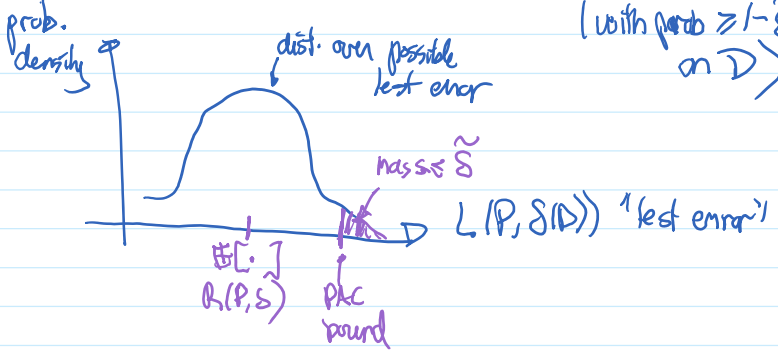
in ML, usually they look at tail bounds for dist. of $L(P, \delta(D))$ where D is random

PAC theory
↳ "probably approx. correct" $P\{L(P, \delta(D)) \geq \text{stuff}\} \leq \tilde{\delta}$

example of generalization error bound

$\text{test error}(\hat{f}) \leq \text{train error}(\hat{f}) + \frac{1}{\sqrt{n}} \sqrt{\text{complexity}(\hat{f}) + \log\left(\frac{1}{\tilde{\delta}}\right)}$

(with prob. $\geq 1 - \tilde{\delta}$ on D)



Bayesian decision theory

→ condition on data D

Bayesian posterior risk $R_B(a|D) = \int_{\Theta} L(\theta, a)\, p(\theta|D)\, d\theta$

$p(\theta|D)$: posterior over "possible worlds"

Θ (4)

posterior over "possible worlds"
$\propto p(\theta)\, p(D|\theta)$

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

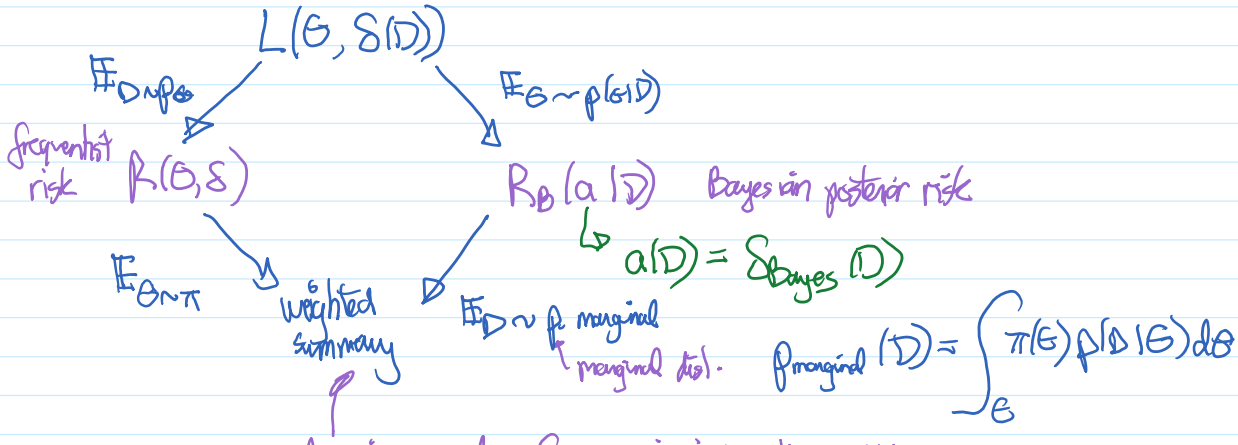
example: if $\Theta = \mathbb{R}$ ("estimation")

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

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

but if use $L(\theta, a) = |\theta - a|$ (1D)

then $\delta_{Bayes}(D)$ = posterior median



Bayesian procedure δ_{Bayes} minimizes the weighted summary

when using prior fct. as the weighting fct.
i.e. $\pi(\theta) = p(\theta)$

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

1) MLE
2) MAP
3) method of moments (MoM)

idea: find an injective mapping from Θ to "moments" of R.V.

$\mathbb{E}X, \mathbb{E}X^2, \ldots$

and then invert it from empirical moments to get $\hat{\theta}$

$\hat{\mathbb{E}}[X] \triangleq \frac{1}{n}\sum_{i=1}^n x_i$

$\hat{\mathbb{E}}[X^2] \triangleq \frac{1}{n}\sum_{i=1}^n x_i^2 \ldots$

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

$\mathbb{E}X = \mu$

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

$$f(\mu, \sigma^2) \triangleq \begin{pmatrix} \mu \\ \sigma^2 + \mu^2 \end{pmatrix}$$

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

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

⊛ MoM is quite used for latent variable models
↳ ("spectral methods" e.g.)

(mixture of Gaussian)

17h30

4) prediction example $\mathcal{A} = \{(f : \mathcal{X} \to \mathcal{Y})\}$ $\mathcal{X}$ ← input space, $\mathcal{Y}$ ← output "

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

is using empirical "risk") minimization (ERM)
↳ Vapnik risk ie. generalization error / test error

ie. $L(P, f) = \mathbb{E}_{(x,y)\sim P}[\ell(Y, f(X))]$

replace this with $\hat{\mathbb{E}}[\ell(Y, f(X))] = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i))$

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

$\mathcal{F}$ ← hypothesis class

James-Stein estimator:

estimator to estimate the mean of $N(\vec{\mu}, \sigma^2 I)$ ← d independent Gaussian variables $X_i \overset{\text{iid}}{\sim} N(\mu_i, \sigma^2)$

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

recall bias-variance decomposition $R(\theta, \hat{\theta}) = \mathbb{E}\|\theta - \hat{\theta}\|_2^2$

$$= \underbrace{\|\mathbb{E}\hat{\theta} - \theta\|_2^2}_{\text{bias}^2} + \underbrace{\mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2]}_{\text{variance}}$$

δ_{JS} actually strictly dominates δ_{MLE} for $d \geq 3$ ← dimension of $\vec{\mu}$

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

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

$\Rightarrow$ MLE is inadmissible in this case [note $n=1$ here $d \geq 3$]

(can interpret the δ_{JS} as an "empirical" Bayesian method)

(asymptotic) properties of MLE:

under suitable regularity conditions on $\mathcal{H}$ & $p(x;\theta)$ $\quad \hat{\theta}_n = \underset{\theta \in \mathcal{H}}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \log p(x_i;\theta)$

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

b) CLT (central limit theorem) $\quad \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{dist.} N(0, I(\theta)^{-1/2})$ [$P_n \sim P_\theta^{\otimes n}$]

information matrix

c) asymptotically optimal (Cramer-Rao lower bound)

i.e. it has minimal asymptotic scaled variance among all "reasonable" estimators

d) invariance: MLE is preserved under reparameterization

suppose have a bijection $f: \mathcal{H} \to \mathcal{H}'$

then $g(\theta) = g(\hat{\theta})$

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

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

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

suppose $g: \mathcal{H} \to \Lambda$

profile likelihood $\triangleq L(\eta) = \max_{\theta;\, \eta = g(\theta)} p(\text{data};\theta)$

define $\hat{\eta}_{MLE} \triangleq \underset{\eta \in \Lambda}{\operatorname{argmax}}\ L(\eta)$

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

"plug in estimator"

$N(\mu, \sigma^2)$

e.g. $g(\mu) = \mu^2$

---

prediction want to learn prediction fct. $h: \mathcal{X} \to \mathcal{Y}$ $\quad \mathcal{Y} = \{0,1\} \to$ binary classification

prediction want to learn prediction fct. $h: \mathcal{X} \to \mathcal{Y}$

$\mathcal{Y} = \{0,1\} \to$ binary classification
$\mathcal{Y} = \{0,\dots,k-1\} \to$ multiclass classif.
$\mathcal{Y} = \mathbb{R} \to$ regression

$x \in \mathbb{R}^d$

$$p(x,y) = \underbrace{p(y|x)}_{\text{"prediction model"}} \underbrace{p(x)}_{\text{model over } \mathcal{X}}$$

$$= \underbrace{p(x|y)}_{\text{"class conditional"}} \underbrace{p(y)}_{\text{prior over class}}$$

"generative perspective" (in context of classification) $\rightsquigarrow$ model $p(x)$ as well

"conditional perspective" $\rightsquigarrow$ only model $p(y|x)$
(traditionally called "discriminative")

"more discriminative"

generative	conditional	"fully discriminative"
model $p_\theta(x,y)$	model $p_\theta(y\|x)$	model $h_\theta: \mathcal{X} \to \mathcal{Y}$ (not nec. derived from $p(y\|x)$)
MLE	max conditional likelihood	reg. ERM; etc. $\frac{1}{n}\sum_{i=1}^{n} \ell(y^{(i)}, h_\theta(x^{(i)}))$
more assumptions $\Rightarrow$ less robust for prediction		less assumptions more robust

$$\hat{h}(x) \triangleq \operatorname{argmin}_{\tilde{y} \in \mathcal{Y}} \sum_{y} p_\theta(y|x)\, \ell(y,\tilde{y})$$

if $\ell(y,\tilde{y}) = \mathbb{1}\{y \neq \tilde{y}\}$ (0-1 loss) then $\hat{h}(x) = \operatorname{argmax}_{\tilde{y} \in \mathcal{Y}} p_\theta(\tilde{y}|x)$