Lecture 6 - scribbles

Friday, September 20, 2019 14:01

today: finish decision theory
more properties of estimators

PAC theory vs. frequentist risk

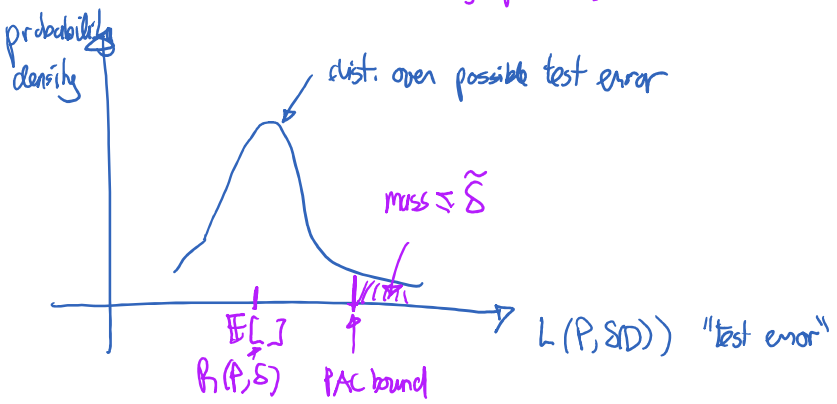
in ML, usually they look tail bounds for dist. $L(P, S(D))$ where D is random

PAC theory
↳ "probably approx. correct"

$$P\{ L(P, S(D)) \geq \text{stuff} \} \leq \tilde{\delta}$$

stuff ↑ bound "with high prob." statement

$\tilde{\delta}$ ↑ small #



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"
$\propto p(\theta)\, p(D|\theta)$

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

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

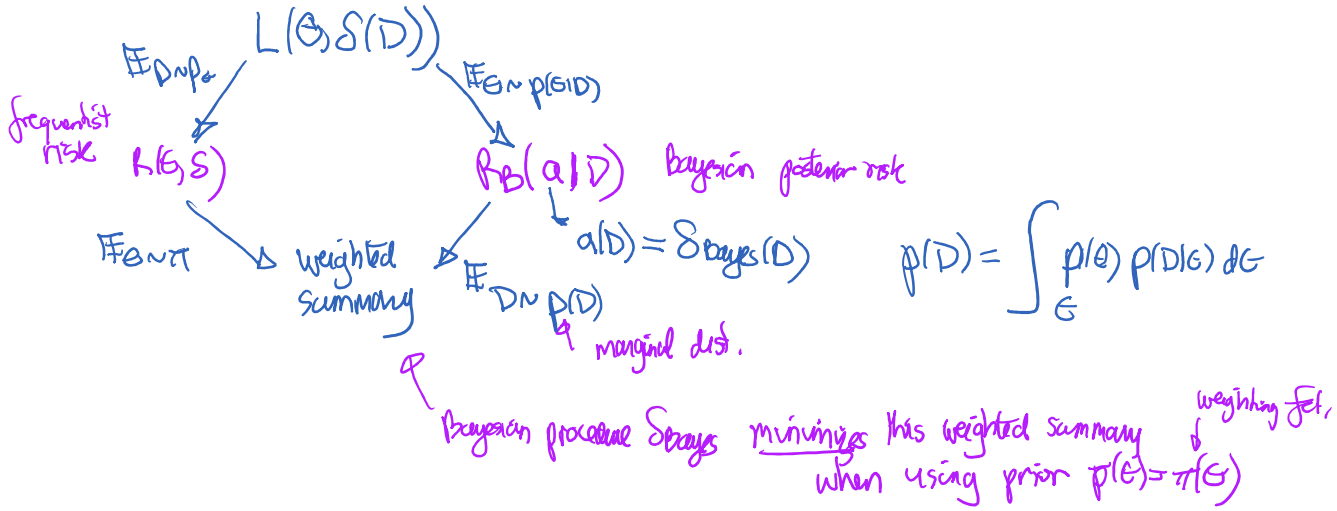
$L(\theta, a) = \|\theta - a\|_2^2$

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

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

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

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



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

1) MLE

2) MAP

3) method of moments (MoM)

idea: find an injective mapping from $\mathcal{H}$ to "moments" of R.V.

$\mathbb{E}X$, $\mathbb{E}X^2$, etc.

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

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

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

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

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

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

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

$\mathbb{E}[X]$

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

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

e.g. mixture of Gaussians

15h10

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

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

is using empirical "risk" minimization (ERM)
↳ "Vapnik risk" i.e. generalization error

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

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

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

↑ hypothesis class

James-Stein estimator:

estimator to estimate the mean of $N(\vec{\mu}, \sigma^2 I)$ ← d independent Gaussian variables with means μ_i & variance σ^2

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

recall the bias-variance decomposition for squared loss

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

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

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

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

↪ MLE is inadmissible in this case

(can interpret the $\hat{\theta}_{JS}$ as an "empirical" Bayesian method)

properties (asymptotic) of MLE:

under suitable regularity conditions on $(\mathcal{H}, p(x;\theta))$ $\quad \hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}} \log p(x_i;\theta)$

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

b) CLT (central limit theorem) $\quad \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I(\theta)^{1/2})$ — $I(\theta)$: information matrix

c) asymptotically optimal (Cramer-Rao lower bound)

ie. it has minimal asymptotic variance among all "reasonable" estimators
- consistent
- ...

d) invariance: MLE is preserved under reparameterization

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

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

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

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

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

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

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

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

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

'plug-in' extender