Lecture 6 - statistical decision theory

Thursday, September 21, 2023 2:26 PM

today: statistical decision theory
evaluation of estimators

statistical decision theory - formal setup

1) • a random observation $D \sim P$ ← unknown distribution which models the world/phenomenon

(often P_θ)

2) • action space $\mathcal{A}$

3) • (statistical) loss $L(P, a)$ = statistical loss of doing action $a \in \mathcal{A}$ when the world is P } describe the goal/task

↳ often write $L(\theta, a)$ if we have a parametric model of world i.e. P is a pdf/pmf p_θ for some $\theta \in \Theta$

• $\delta : \mathcal{D} \to \mathcal{A}$ "decision rule"

$\mathcal{D}$: Ω_D

examples: a) parameter estimation:

$\mathcal{A} = \Theta$ for parametric family P_Θ

δ is a parameter estimator from data $D = (X_1, \dots, X_n)$ (typically)

[usually $X_i \overset{iid}{\sim} P_\theta$]

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

"squared loss"

but another loss is $KL(P_\theta \| P_a)$

b) $\mathcal{A} = \{0, 1\}$; this is hypothesis testing

δ describes a statistical test

$\mathbb{1}_A(u) = \begin{cases} 1 \text{ if } u \in A \\ 0 \text{ o.w.} \end{cases}$

loss → usually 0-1 loss $L(\theta, a) = \mathbb{1}_{\{\theta\}^c}(a)$ $\left(\triangleq \mathbb{1}\{\theta \neq a\}\right)$

$\{\theta\}^c = \Theta \setminus \{\theta\}$

$\{\theta\}^c = \Theta \setminus \{\theta\}$

c) prediction in ML = learn a prediction fct. in supervised learning (function estimation)

here $D = (X_i, Y_i)_{i=1}^n$ $X_i \in \mathcal{X}$ (input space) $Y_i \in \mathcal{Y}$ (output space)

$\mathcal{Y} = \{0,1\}$ → classification

$\mathcal{Y} = \mathbb{R}$ → regression

P_θ gives joint on (X,Y)

$\mathcal{A} = \mathcal{Y}^{\mathcal{X}}$ (set of fct's from $\mathcal{X}$ to $\mathcal{Y}$)

$|\mathcal{A}| = |\mathcal{Y}|^{|\mathcal{X}|}$

$D \sim P$ where $P = \underbrace{P_\theta \otimes P_\theta \otimes \cdots P_\theta}_{n \text{ times}}$

$p(x_1, \ldots, x_n) = \prod_{i=1}^n p_\theta(x_i)$

in ML $\boxed{L(P_\theta, f) \triangleq \mathbb{E}_{P_\theta}[\ell(Y, f(X))]}$ "generalization error"

$(X,Y) \sim P_\theta$ "prediction loss" "classification error"

in ML, is often call the "risk"

↳ e.g. classification $\ell(Y, f(X)) = \mathbb{1}\{Y \neq f(X)\}$ 0-1 error

Svrion calls it "Vapnik risk" to distinguish it from frequentist risk $\mathbb{E}_D[L(P_\theta, \delta(D))]$

↳ think cross-validation e.g.

* decision rule $\hat{f} = \delta(D)$

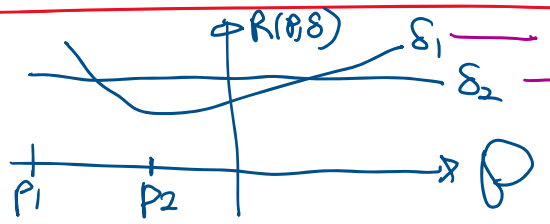
prediction fct. / classifier / etc. "learning alg."

Comparing procedures?

δ_1 vs. δ_2

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

"risk profiles"



* transform to scalar

• "minimax" analysis : $\max_{P \in \mathcal{P}} R(P, \delta)$ "worst case"

- "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

[5h31]

PAC theory vs frequentist risk :

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

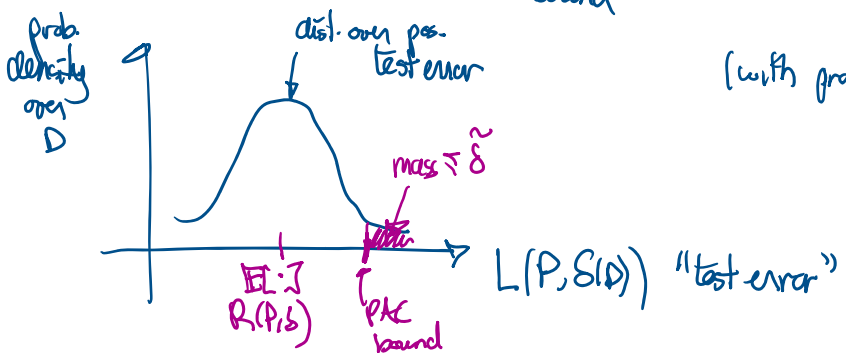
PAC theory
↳ "probably approx. correct"

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

example of generalization error bound

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

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



Bayesian decision theory

→ condition on data D

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

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

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

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

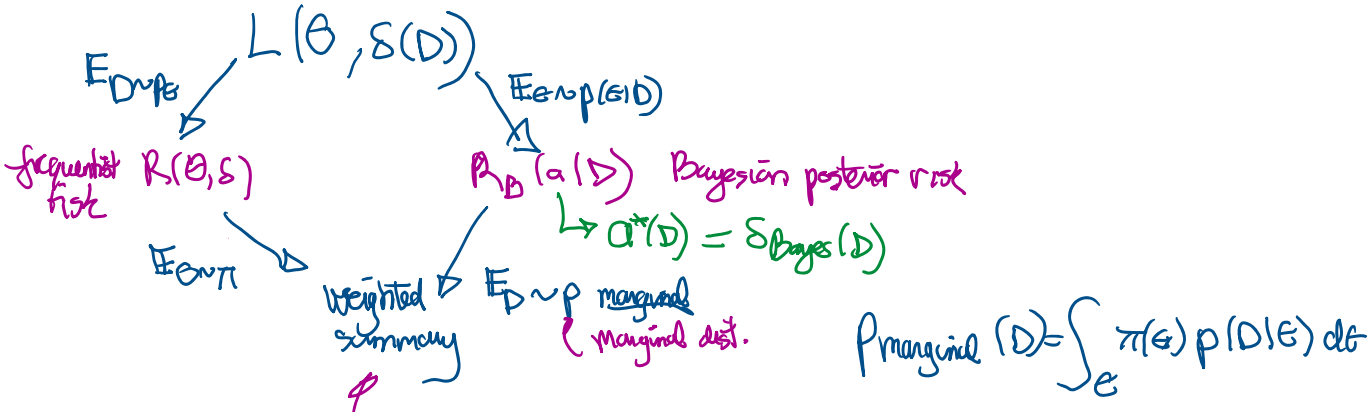
example : if $\mathcal{A} = \mathcal{H}$ ("estimation")

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

then (exercise) $\delta_{Bayes}(D) = \mathbb{E}[\theta \mid D]$ (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 among all δ's

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

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.

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

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

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

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

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 → see later]

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

(mixture of Gaussian)

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

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

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

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

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

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

↳ hypothesis class

James-Stein estimator:

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

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

for $d \geq 3$ (d: dimension of $\vec{\mu}$)

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

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

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

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