Lecture 6 - statistical decision theory

Monday, September 23, 2024 12:32 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 world is P } describe the goal/task

↳ often write $L(\theta, a)$ if we have a parametric model of world
i.e. P has 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 some parametric family $\mathcal{P}_\Theta$

δ is a parameter estimator **from data** $D \overset{\text{typically}}{=} (X_1, X_2, \ldots, X_n)$

[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 $L(\theta, a) = \mathbb{1}_{\{\theta\}^c}(a)$ $(\triangleq \mathbb{1}\{\theta \neq a\})$

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

→ related to type I & type II errors $\{\theta\}^c$ "whole...

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

→ related to type I & type II errors

$\{\theta\}^C$ "whole world"

$A^C = B \setminus A$

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

let p_θ be joint on (X, Y)

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

$\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$ (set of functions from $\mathcal{X}$ to $\mathcal{Y}$)

$p(\tilde{x}_1, \ldots, \tilde{x}_n) = \prod_{i=1}^{n} p_\theta(\tilde{x}_i)$

in ML

$$L(p_\theta, f) \triangleq \mathbb{E}_{(x,y)\sim p_\theta}[\ell(Y, f(X))]$$

"generalization error"

"classification error"

ℓ: "prediction loss"

→ e.g. classification $\ell(Y, f(X)) = \mathbb{1}\{Y \neq f(X)\}$

0-1 error

→ in ML, is often called the "risk"

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

→ think cross-validation e.g.

* decision rule

$\hat{f} = \delta(D)$ ← "learning alg"

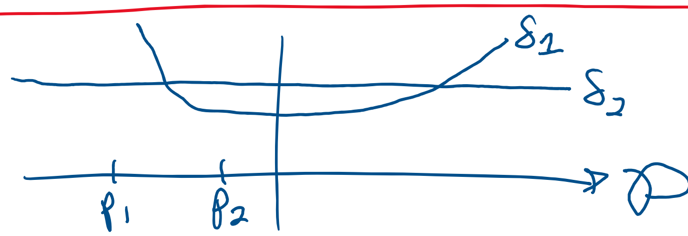
↑ prediction fct. classifier / regressor / etc. ..

comparing procedures?

δ_1 vs. δ_2

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

"risk profiles"



* transform to scalar

- "minimax" analysis : $\max_{P \in \mathcal{D}} R(P, \delta)$ "worst case"
- weighted average : $\int_{\Theta} R(\theta, \delta)\, \pi(\theta)\, d\theta$ (kind of a Bayesian feel)
  - $\pi(\theta)$: weighting

13h34

PAC theory vs. frequentist risk :

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

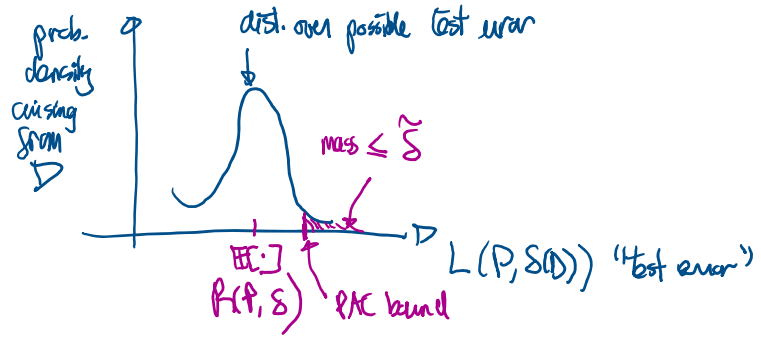
PAC theory
↳ "probably approx. correct"

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

example of generalization error bound

$$\text{test\_error}(\vec{f}) \leq \text{train\_error}(\vec{f}) + \frac{1}{\sqrt{n}} \sqrt{\text{complexity}(\vec{f}) + \log \frac{1}{\delta}}$$

(with prob $\geq 1 - \tilde{\delta}$
↳ randomness of D



Bayesian decision theory

→ condition on data D

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

↳ posterior over "possible worlds"

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

Bayesian optimal action $\delta_{\text{Bayes}}(D)$

$\triangleq \operatorname{argmin}_{a \in A} R_B(a|D)$

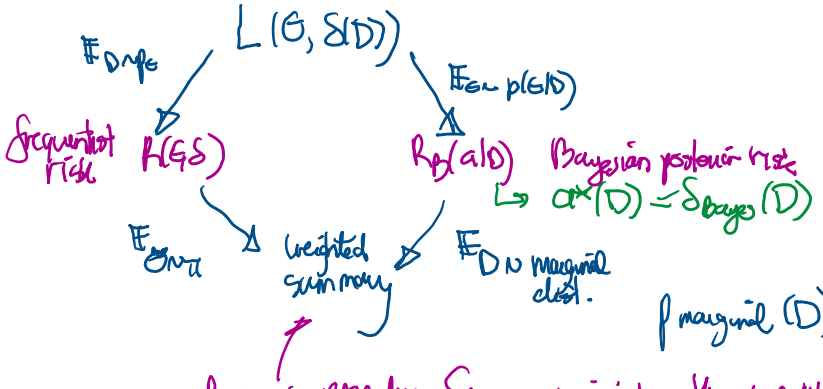
example : if $A = \Theta$ "estimation"

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

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

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

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



$$p_{marginal}(D) = \int_\Theta \pi(\theta) p(D|\theta) d\theta$$

Bayesian procedure δ_{Bayes} minimizes the weighted frequentist risk among all δ's when using prior fct. as the weighting fct. i.e. $\pi(\theta) = p(\theta)$

Examples of estimators: $\delta: D \to \Theta$

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 $\hat{\theta}$

$$\hat{\mathbb{E}}[X] \triangleq \frac{1}{n}\sum_i x_i$$

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

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 gives same answer as MLE)

[general property in exponential family → see later]

(*) MoM is quite useful for latent variable models (mixture of Gaussian)

↳ ("spectral methods" e.g.)

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

$\mathcal{Y}$ ← output space

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

is using empirical "risk" minimization (ERM)

↳ Vapnik risk i.e. test error

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

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

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

↑ hypothesis class

James-Stein estimator:

estimator to estimate the mean of $N(\vec{\mu}, \sigma^2 I)$ ← d independent Gaussian 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. $R(\theta, \delta_{JS}) \leq R(\theta, \delta_{MLE})\ \forall \theta$

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

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

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

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

here $d \geq 3$