Lecture 6 - scribbles

Friday, September 18, 2020 13:29

today: statistical decision theory
properties of estimator

Statistical decision theory

Formal setup:
- a random observation $D \sim P$ (often P_θ) — unknown distribution which models the world/phenomenon
- action space $\mathcal{A}$
- loss $L(P, a)$ = statistician's 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 had pdf/pmf p_θ for some $\theta \in \Theta$
- $\delta : \mathcal{D} \to \mathcal{A}$ "decision rule"
  - $\mathcal{D} = \Omega_D$

Examples: a) parameter estimation:

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

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

[usually, $X_i \overset{\text{iid}}{\sim} p_\theta$, θ unknown]

typical loss $L(\theta, a) = \|\theta - a\|_2^2$, $a \in \mathcal{A} = \Theta$ "squared loss"

$\hat{\theta} = \delta(D)$

but other losses are used, eg. $KL(p_\theta \| p_a)$

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

δ describes a statistical test

loss → usually 0-1 loss $L(\theta, a) = \mathbb{1}\{\theta \neq a\}$

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

p_θ gives joint (X, Y)

$D \sim P$ where $P = p_{\theta_0} \otimes p_{\theta_1} \cdots \otimes p_{\theta_n}$

P_θ gives joint (X,Y)

$D \sim P$ where $P = P_\theta \otimes P_\theta \cdots \otimes P_\theta$ (n times)

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

in machine learning $L(P_\theta, f) \triangleq \mathbb{E}_{(X,Y)\sim P_\theta}[\ell(Y, f(X))]$

"generalization error"

"classification error"

in ML, is often called the "risk"

$(X,Y) \sim P_\theta$

prediction loss

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

Simon calls it "Vapnik risk" to distinguish it from frequentist risk

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

prediction fct. / classifier / etc..

"learning algorithm"

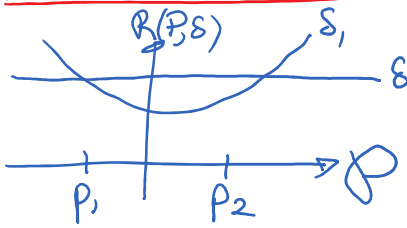
frequentist risk $\mathbb{E}_D[L(P_\theta, \delta(D))]$

comparing procedure?

δ_1 vs. δ_2

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

"risk profiles"



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

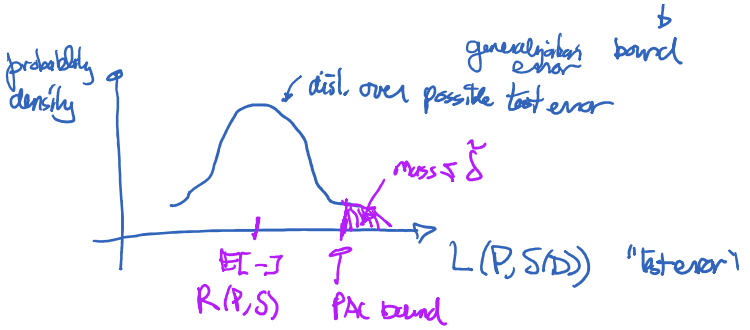
weighting

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

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

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

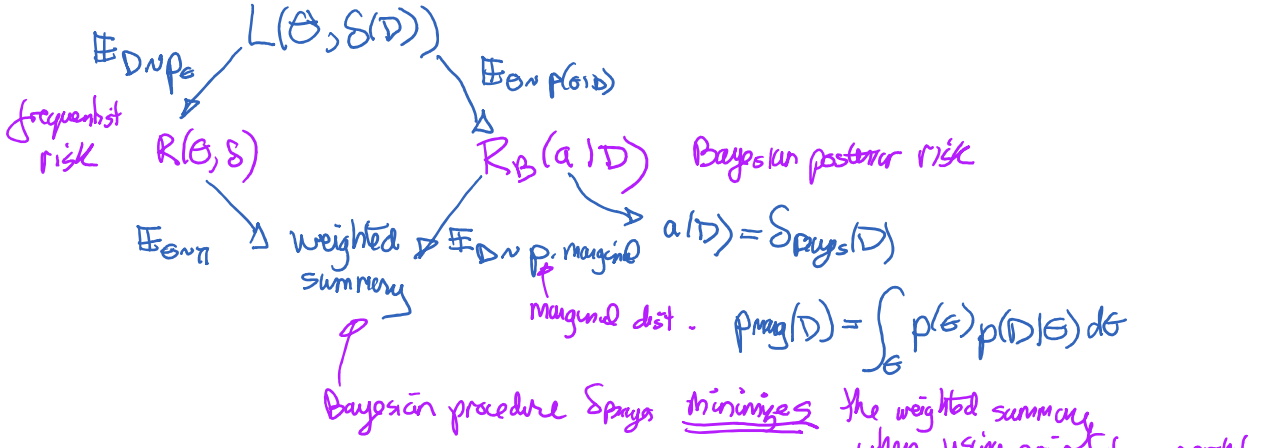
example: if $\mathcal{A} = \Theta$ ("estimation")

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

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



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

$\pi(\theta) = p(\theta)$

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 RVs

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

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

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

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

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 of exponential family])

(*) MoM is quite used for latent variable models (e.g. mixture of Gaussian)

↳ ("spectral methods" e.g.)