Lecture 5 - scribbles

Tuesday, September 19, 2017
14:37

aside on proofs:

assumptions ⇝ use logical laws like $A \Rightarrow B$ ⇝ conclusion

example of thm: $X \perp\!\!\!\perp Y \Rightarrow p(x|y) = p(x) \quad \forall x, y$

$X \perp\!\!\!\perp Y$ (assumption)

by definition $\Rightarrow$ $(*)$ $p(x,y) = p(x)p(y) \quad \forall x, y$

suppose y s.t. $p(y) \neq 0$

$$p(x|y) \overset{\text{by definition}}{=} \frac{p(x,y)}{p(y)} \overset{\text{by } (*)}{=} \frac{p(x)p(y)}{p(y)} = p(x)$$

(proof by contradiction: use fact that $(A \Rightarrow B) \iff (\neg B \Rightarrow \neg A)$) ("equivalent")

---

today: statistical decision theory

(frequentist) statistical decision theory

formal setup:

- random observation $D \sim P$ (perhaps P_θ) — unknown distribution which models the "world"
- action space $\mathcal{A}$
- loss: $L(D, a)$ = loss of doing action a ? describes the

when "true world" is p $\mathcal{S}$ goal/task

(if you have parametric model in mind; often write $L(\theta, a)$ where θ is s.t. $p = p_\theta$)

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

$\mathcal{D}$ — set of parameters for $\mathcal{P}_\Theta$

examples: a) if $\mathcal{A} = \Theta$ for a parametric family $\mathcal{P}_\Theta$

δ is then parameter estimator from data

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

[more specifically: $D = (X_1, \dots, X_n)$ where $X_i \overset{iid}{\sim} p_\theta$ (θ is unknown)

$\delta(D) = \hat{\theta}$ $\quad L(\theta, \delta(D)) = \|\theta - \delta(D)\|_2^2$]

suppose $\mathcal{P}_\Theta$ is Gaussian family $\{N(\mu, 1) : \mu \in \mathbb{R}\}$

$\Theta = \mathbb{R}$ (set of possible means)

$\delta(D) = \frac{1}{n} \sum_i X_i$

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

here δ describes a statistical test

c) machine learning: learning a prediction rule

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

e.g. $\mathcal{Y} = \{0, 1\}$ for binary classification

if p_θ gives joint on (X_i, Y_i); then $D \sim P$ where $P = P^{\otimes n} = P_1(x) P_2(x) \cdots \otimes P_n$

... then $D \sim P$ where $P = P_0^{\otimes n} = P_0 \otimes P_0 \otimes \cdots \otimes P_0$ (iid model)

(n times)

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

in machine learning, $L(P, f) \triangleq \mathbb{E}_{P_0}[\ell(Y, f(X))]$ "generalization error"

(ℓ: prediction loss; $(X,Y) \sim P_0$)

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

($\hat{f}$: prediction function; δ: "learning algorithm"; D: training dataset)

in M.L., often called "risk"

Suman calls it "Vapnik risk" to distinguish from statistical freq. risk

procedure analysis:

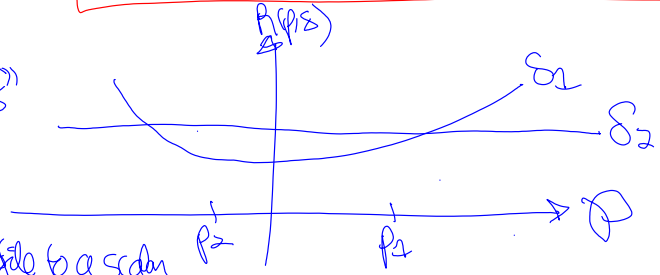
given this framework, how do we compare procedures? e.g. δ_1 vs δ_2

first property:

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

(P: $D \sim P$)

"how well δ does in average"



"risk profiles"

how to go from risk profile to a scalar number

* "minimax" → look at $\max R(P, \delta)$

alternative in ML theory is PAC theory → looking at tail bounds

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

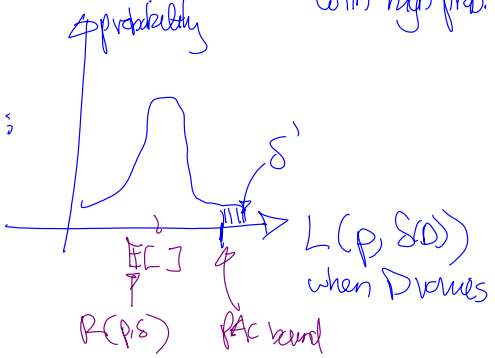
"with high prob." statement

* weighting = "prior" over $\mathcal{H}$

$$\int_{\mathcal{H}} R(\theta, \delta)\, \pi(\theta)\, d\theta$$

θ ∈ $\mathcal{H}$

cartoon:



(*) Bayesian decision theory

→ condition on data D

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

posterior ∝ $p(\theta)\, p(D|\theta)$

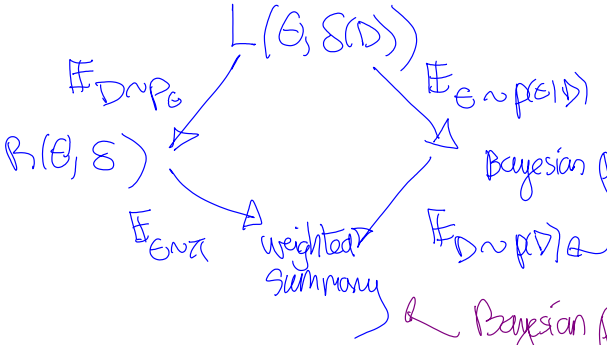
Bayesian optimal action $\delta_{\text{Bayesian}}(D) = \underset{a \in A}{\operatorname{argmin}}\, R_B(a|D)$

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

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

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

$L(\theta, \delta(D))$



↳ Bayesian procedure minimize this when using prior $= \pi(\theta)$ ↳ weighting function

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

- MLE
- another: MAP, given a prior $p(\theta)$ ("maximum a posteriori")

then pick $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ p(\theta|D)$

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

- method of moments

idea: find an injective mapping from Θ to "moments" $\mathbb{E}[X]$, $\mathbb{E}[X^2]$, etc...

and surjective on "possible moments"

and then invert it from empirical moments $\hat{\mathbb{E}}[X] \triangleq \frac{1}{n}\sum_i X_i$, $\hat{\mathbb{E}}[X^2]$

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

$\mathbb{E}[X] = \mu$

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

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

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

(here, this same as MLE (property of exponential family))

⊛ this is useful for latent variable models (e.g. mixture of Gaussians) ("spectral methods") e.g.

- in the context of prediction $\mathcal{A} = \{f : \mathcal{X} \to \mathcal{Y}\}$ $\mathcal{X}$ 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 error

recall: $L(p, f) = \mathbb{E}_{(x,y)\sim p}[\ell(y, f(x))]$

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

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

$\mathcal{F}$ hypothesis class