Lecture 5 - scribbles

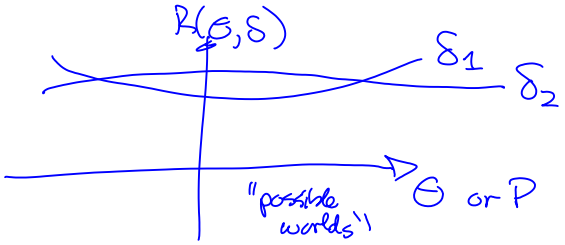
Friday, September 16, 2016
13:27

today:
1) finish decision theory
2) prediction models → linear regression
logistic regression

(frequentist) statistical decision theory

$L(P, a)$ ← loss describing task
$\mathcal{A}$ action space
$\delta: \mathcal{D} \to \mathcal{A}$
$a = \delta(D)$
↑ random observation $D \sim P$

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



- minimax

"prior" over Θ to get weighted summary of performance

$$\int_\Theta R(\theta, \delta)\, p(\theta)\, d(\theta)$$

(*) Bayesian decision theory: → we condition on data D
instead of averaging

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

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

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

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

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

→ if consider $\int_\Theta p(\theta)\, R(\theta, \delta_{Bayesian})\, d\theta$
↳ I use same $p(\theta)$ to define it

$L(\theta, \delta(D))$
frequentist $\mathbb{E}_{D|P}$ $R(\theta, \delta)$
$\mathbb{E}_{\theta \sim p(\theta)}$
Bayesian notions then this has minimal value

$R(\theta, \delta)$ ... Bayesian risk ... then this has minimal value amongst "all" estimators δ (under regularity conditions)

$\mathbb{E}_{\theta \sim p(\theta)}$ $\mathbb{E}_{D \sim P}$ same answer

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

- MLE
- another: MAP, given a prior $p(\theta)$,
  then pick $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ p(\theta \mid D)$
  $\propto p(D \mid \theta) p(\theta)$
- method of moments:

  idea: find injective and surjective mapping from $\mathcal{H}$ to "moments" on "possible moments"
  - $\mathbb{E}[X]$
  - $\mathbb{E}[X^2]$
  - etc.

  and then invert it from empirical distribution $\hat{\mathbb{E}}[X]$, $\hat{\mathbb{E}}[X^2]$ etc...

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

  $$\mathbb{E}[X] = \mu$$
  $$\mathbb{E}[X^2] = \sigma^2 + \mu^2$$
  $$\begin{pmatrix} \mathbb{E}[X] \\ \mathbb{E}[X^2] \end{pmatrix} = f(\mu, \sigma^2)$$
  $$\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} = f^{-1}\left(\begin{pmatrix} \hat{\mathbb{E}}(X) \\ \hat{\mathbb{E}}(X^2) \end{pmatrix}\right)$$

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

  $\mathcal{X}$ ← input
  $\mathcal{Y}$ ← output space

example of $\delta : \mathcal{D} \to \mathcal{F}$
is using empirical "risk" minimization (ERM)
↓
"Vapnik risk"

recall $L(P, f) = \mathbb{E}_{(x,y) \sim P}[\ell(Y, f(X))]$ ← "generalization error"
↑ prediction loss

ERM $\quad \hat{\mathbb{E}}[\ell(Y, f(X))]$

$$= \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) \quad \text{for } \{(x_i, y_i)\}_{i=1}^{n}$$

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

iid setting

suppose that $P = p^{\otimes n}$ i.e. $D = \{X_i\}_{i=1}^n$, $X_i \overset{iid}{\sim} P$

write $\begin{cases} \delta_n \\ \hat{\theta}_n \end{cases}$ to emphasize dependence on n

study $R(\theta, \delta_n)$ as function of n

in particular $R(\theta, \delta_n) \xrightarrow{n\to\infty} 0$ ("consistency")

for estimation, had $L(\theta, \delta_n(D)) = \|\theta - \delta_n(D)\|_2^2$

risk: $\mathbb{E}[\|\theta - \hat{\theta}_n\|^2]$

standard statistical consistency: $\hat{\theta}_n \xrightarrow{P} \theta$ "in probability"

$\forall \varepsilon > 0$, $P\{\|\theta - \hat{\theta}_n\| > \varepsilon\} \xrightarrow{n\to\infty} 0$

$$\mathbb{E}_{D\sim P}[\|\theta - \hat{\theta}_n(D)\|^2] = \mathbb{E}[\|\theta - \mathbb{E}[\hat{\theta}_n] + \mathbb{E}[\hat{\theta}_n] - \hat{\theta}_n\|^2]$$

$$= \mathbb{E}[\|\theta - \mathbb{E}[\hat{\theta}_n]\|^2] + \mathbb{E}[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\|^2] + 2\mathbb{E}[\langle \theta - \mathbb{E}[\hat{\theta}_n], \mathbb{E}[\hat{\theta}_n] - \hat{\theta}_n \rangle]$$

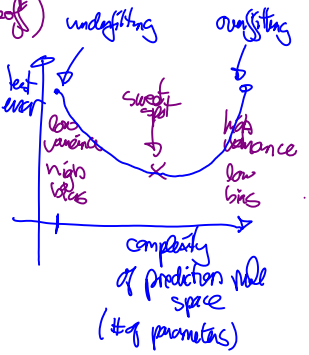
($\theta - \mathbb{E}[\hat{\theta}_n]$: constants; $\mathbb{E}[\hat{\theta}_n] - \mathbb{E}[\hat{\theta}_n] = 0$)

$$= \|\mathbb{E}[\hat{\theta}_n] - \theta\|^2 + \mathbb{E}[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\|^2]$$

$\|\mathbb{E}[\hat{\theta}_n] - \theta\|^2 \triangleq$ bias; $\mathbb{E}[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\|^2]$ = variance $(\hat{\theta}_n)$

risk for squared loss $= \|\text{bias}(\hat{\theta}_n)\|^2 + \text{variance}(\hat{\theta}_n)$

bias-variance decomposition (tradeoff)



James-Stein estimator: estimating mean of $N(\mu, \sigma^2 I)$

↳ is biased, but lower variance than MLE

and so JS dominates MLE for $d \geq 3$

note: $R(\theta, \delta) \xrightarrow{n\to\infty} 0$ for squared loss

$\mathbb{E}[\|\theta - \hat{\theta}_n\|^2] \to 0$ "convergence in L_2"

convergence in L_2 $\Rightarrow$ convergence in prob

so $R(\theta,\delta) \to 0 \Rightarrow \hat{\theta}_n \xrightarrow{p} \theta$ i.e. is consistent
& squared loss

{ bias $\to 0$, variance $\to 0$ } $\Rightarrow$ consistency

asymptotic properties of MLE: under regularity conditions on (X) & $p(x;\theta)$

a) $\hat{\theta}_n \xrightarrow{p} \theta$

b) CLT: $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \underbrace{I(\theta)^{-1/2}}_{\text{information matrix}})$

c) asymptotically optimal
ie. has minimal asymptotic variance (Cramer-Rao lower bound)
among all 'reasonable' estimators
↳ • consistent
& ...

d) invariance:
if I reparameterize (H)
ie. $f: (H) \to (H')$
bijection
then $\widehat{f(\theta)} = f(\hat{\theta})$

if you don't use bijection, can generalize the MLE
with "profile likelihood" $\log: (H) \to \Lambda$

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

define $\hat{\eta}_{MLE} = \underset{\eta \in g(\Theta)}{\operatorname{argmax}} L(\eta)$

then we have $\hat{\eta}_{MLE} = g(\hat{\theta}_{MLE})$ { "plug-in"

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

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