Lecture 4 - prob. / frequentist vs. Bayesian

Monday, September 16, 2024 12:31 PM

today: statistical frequentist vs. Bayesian

statistical concepts

Cartoon

prob. theory } well-defined

$$\begin{tikzcd} \underset{\text{model}}{\theta} \arrow[r, bend left, "\text{prob. theory}"] & \underset{\text{data}}{X} \arrow[l, bend left, "\text{statistics}"] \end{tikzcd}$$

statistics } ill-defined "inverse problems"

epistemology of science?

example: model n indep coin flips

prob. theory → prob. k heads in a row

statistics: I have observed k heads, $n-k$ tails, what is θ?

frequentist vs Bayesian

semantic of prob.?: meaning of a prob?

a) (traditional) frequentist semantic

$P\{X=x\}$ represents the limiting frequency of observing $X=x$

if I could repeat ∞ # of i.i.d experiments

b) Bayesian (subjective) semantic

$P\{X=x\}$ encodes an agent "belief" that $X=x$

laws of prob. characterizes a "rational" way to combine "beliefs" and "evidence" [observations]

[has motivation in terms of gambling, utility/decision theory, etc.]

operationally: Bayesian approach: (*) very simple philosophically

treat all uncertain quantities as R.V.

i.e. encode all knowledge about the system ("beliefs") as a "prior" on probabilistic models as a "prior" on probabilistic models

and then use law of prob. (and Bayes rule) to get updated beliefs and answer?

justification for frequentist semantic:

for discrete R.V., suppose $P\{X=x\}=\theta$

$$\Rightarrow P\{X\neq x\}=1-\theta$$

$B \triangleq \mathbb{1}_{\{X=x\}}$ $\Rightarrow B \sim \text{Bern}(\theta)$ R.V.

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

repeat i.i.d. ie. $B_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$

by L.L.N. (law of large numbers) $\frac{1}{n}\sum_{i=1}^{n} B_i \xrightarrow{a.s.} \mathbb{E}[B_1]=\theta$

limiting frequency

by C.L.T. $\sqrt{n}\Big(\frac{1}{n}\underbrace{\sum_{i=1}^{n} B_i}_{\sim \text{Bin}(n,\theta)} - \theta\Big) \xrightarrow{d} N(0, \theta(1-\theta))$

coin flips – Bayesian approach

biased coin flips

unknown ⇒ model it as R.V.

we believe $X \sim \text{Bin}(n,\theta)$ ⇒ need a $p(\theta)$ "prior distribution"

$$\Omega_{\Theta}=[0,1]$$

suppose we observe $X=x$ (result of n coin flips)

then we can "update" our belief about Θ by using Bayes rule

$$\boxed{p(\Theta=\theta \mid X=x)} = \frac{\boxed{p(X=x\mid\Theta=\theta)}\,\boxed{p(\Theta=\theta)}}{p(X=x)}$$

posterior belief; observation model / likelihood; prior belief; "marginal likelihood" (normalization)

[note: $p(x|\theta)\rightarrow$ pmf $\quad p(x,\theta)$ is "mixed distribution"]

$p(\theta) \rightarrow$ pdf

Example: suppose $p(\theta)$ is uniform on $[0,1]$ "no specific preference"

$$p(\theta|x) \underset{\text{"proportional to"}}{\propto} p(x|\theta)\, p(\theta)$$

$$p(\theta|x) \propto \theta^{x}(1-\theta)^{n-x} \overbrace{\mathbb{1}_{[0,1]}(\theta)}^{\text{prior}}$$

up to a constant in θ

scaling $\int_0^1 \theta^x (1-\theta)^{n-x}\, d\theta = B(x+1, n-x+1)$

normalization constant $\int_\theta p(\theta|x)\, d\theta = 1$

$\underbrace{B(a,b)}_{\text{beta fct.}} \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ $\qquad \underset{\text{gamma function}}{\Gamma(a)} = \int_0^\infty u^{a-1} e^{-u}\, du$

here $p(\theta|x)$ is called "beta distribution"

$$B(\theta|\underset{\text{parameters}}{\alpha,\beta}) \triangleq \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \mathbb{1}_{[0,1]}(\theta)$$

- uniform dist density: $B(\theta|1,1)$

posterior density: $B(\theta|x+1, n-x+1)$ "as prior counts"

exercise to the reader: if use $B(\alpha_0, \beta_0)$ as prior

posterior will be $B(x+\alpha_0, n-x+\beta_0)$

13h34

(*) posterior $p(\theta|X=x)$ contains all the info from data x that we need to answer questions about θ

e.g. question: what is prob. of head $(\underset{\text{flip outcome}}{F=1})$ on the next flip?

as a frequentist $P(F=1|\text{data } X=x) = \hat{\theta}$ ← notation to mean "estimates"

as a Bayesian $P(F=1|X=x) = \int_\theta P(F=1, \Theta=\theta|X=x)\, d\theta$

product rule

$$= \int_\theta \underbrace{p(F=1|\Theta=\theta, X=x)}_{\substack{=\theta \\ \text{(by our model)}}} \underbrace{p(\theta|X=x)}_{\text{posterior}}\, d\theta$$

$$= \int \theta\, p(\theta|X=x)\, d\theta = \mathbb{E}[\theta|X=x]$$

$$= \int_{\Theta} \theta \, p(\theta | X=x) \, d\theta = \mathbb{E}[\theta | X=x]$$

"posterior mean" of θ

* as a meaningful "Bayesian" estimator of θ

$$\hat{\theta}_{Bayes}(x) \triangleq \mathbb{E}[\theta | X=x] \quad \text{(posterior mean)}$$

notation: $\hat{\theta}$: observation $\rightarrow \Omega_\theta$

<u>On coin example</u>: $p(\theta|x) = \text{Beta}(\theta | \alpha = x + \alpha_0, \beta = n - x + \beta_0)$

mean of a beta R.V. $\frac{\alpha}{\alpha+\beta}$

thus $\boxed{\hat{\theta}_{Bayes}(x) = \mathbb{E}[\theta|x] = \frac{x+\alpha_0}{n+\alpha_0+\beta_0}}$

here, biased estimator $\mathbb{E}_X[\hat{\theta}(X)] \neq \theta$

$$\mathbb{E}\left[\frac{X+\alpha_0}{n+\alpha_0+\beta_0}\right] = \frac{\overbrace{\mathbb{E}X}^{n\theta}+\alpha_0}{n+\alpha_0+\beta_0} \neq \theta \text{ unless } \alpha_0 = \beta_0 = 0 \text{ (not allowed)}$$

but asymptotically unbiased $\xrightarrow[n\to\infty]{} \theta$

compare & contrast $\hat{\theta}_{MLE}(x) = \frac{x}{n}$ (unbiased)

<u>to summarize</u>:

- as a Bayesian: get a posterior + use law of probabilities
- in "frequentist statistics"

consider multiple estimators
- MLE
- moment matching
- Bayesian posterior mean
- MAP
- regularized MLE
- etc.

and then analyze the statistical properties of estimator?

- biased?
- variance?
- consistent?
- frequentist risk?

<u>Maximum likelihood principle</u>

Maximum likelihood principle

setup: given a parametric family $p(x;\theta)$ for $\theta \in \Theta$

we want to estimate / learn θ from x

$$\boxed{\hat{\theta}_{MLE}(x) \triangleq \underset{\theta \in \Theta}{\arg\max}\ \underbrace{p(x;\theta)}_{\triangleq L(\theta)}}$$

"likelihood fct." of θ

$\hat{\theta}_{MLE}(x)$ maximizes $p(x;\cdot)$

MLE example I: binomial

n coin flips $\Omega_X = 0{:}n$

$X \sim Bin(n,\theta)$ $\quad p(x;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$

trick: to maximize $\log L(\theta)$ instead of $L(\theta)$

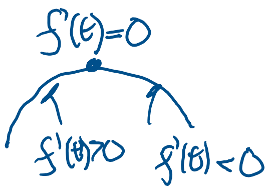
$\triangleq \ell(\theta)$ $\boxed{\text{log-likelihood}}$

justification: $\log(\cdot)$ is strictly increasing

i.e. $a<b \iff \log a < \log b \quad (\forall a,b>0)$

$\Rightarrow \underset{\theta\in\Theta}{\arg\max}\ \log p(x;\theta) = \underset{\theta\in\Theta}{\arg\max}\ p(x;\theta)$

$$\log p(x;\theta) = \underbrace{\log\binom{n}{x}}_{\text{constant w.r. to }\theta} + x\log\theta + (n-x)\log(1-\theta) = \ell(\theta)$$



look for θ s.t. $\frac{\partial \ell}{\partial \theta} = 0$

want: $\frac{x}{\theta} - \frac{n-x}{1-\theta} = 0$

$x(1-\theta) = \theta(n-x)$

$x - \cancel{\theta x} = n\theta - \cancel{\theta x}$ $\Rightarrow \boxed{\theta^* = \frac{x}{n}}$

used as solution in optimization

hence $\boxed{\hat{\theta}_{MLE}(x) = \frac{x}{n}}$