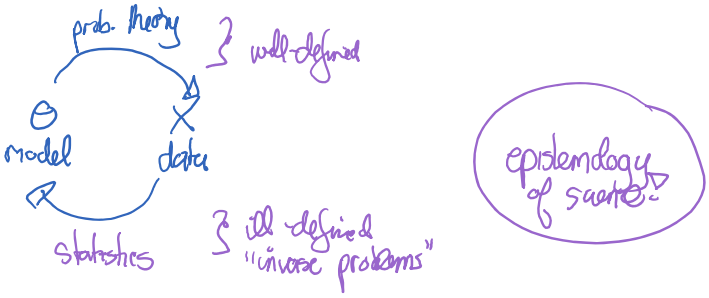
Lecture 4 - prob. / frequentist vs. Bayesian

Thursday, September 14, 2023 2:30 PM

today: statistics frequentist vs. Bayesian

Statistical concepts



example: model ∧ 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 and then use laws of prob. (and Bayes rule) to get updated beliefs and answers

justification for frequentist semantic

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

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

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

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

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

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

↑ limiting frequency

by C.L.T. $\quad \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} B_i - \theta\right) \xrightarrow{d} N(0, \theta(1-\theta))$

$\sum_{i=1}^{n} B_i \sim \text{Bin}(n, \theta)$

Coin flips — Bayesian approach

biased coin flips

unknown $\Rightarrow$ model it as R.V.

we believe $X \sim \text{Bin}(n, \Theta)$ $\quad \Rightarrow$ 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

$$\underbrace{p(\Theta=\theta \mid X=x)}_{\text{posterior belief}} = \frac{\overbrace{p(X=x \mid \Theta=\theta)}^{\text{observation model / likelihood}} \overbrace{p(\Theta=\theta)}^{\text{prior belief}}}{\underbrace{p(X=x)}_{\text{normalization "marginal likelihood"}}}$$

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

$p(\theta) \to$ pdf

example:

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

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

↑ "proportional to"

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

up to a constant in θ

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

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

normalization constant $\int_\theta p(\theta|x)\, d\theta = 1$ $\quad B(a,b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ (Beta fct., gamma fct.) $\quad \Gamma(a) = \int_0^\infty u^{a-1} e^{u}\, du$

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

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

(α, β: parameters)

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

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

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

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

15h32

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

e.g. question: what is prob. of head ($F=1$, flip outcome) on the next flip?

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

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

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

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

("posterior mean" of θ)

* a meaningful "Bayesian" estimator of θ

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

notation: $\hat{\theta}$ & observation $\to \Omega_\Theta$

Our coin example: $p(\theta|x) = \text{Beta}(\theta|\alpha = x+1, \beta = n-x+1)$

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

thus $\hat{\theta}_{\text{Bayes}}(x) = \mathbb{E}[\theta|x] = \frac{x+1}{n+2}$

thus $\hat{\theta}_{Bayes}(x) = \mathbb{E}[\theta|x] = \frac{x+1}{n+2}$

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

$= \mathbb{E}\left[\frac{X+1}{n+2}\right] = \frac{\mathbb{E}X+1}{n+2} = \frac{n\theta+1}{n+2} \neq \theta$

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

compare & contrast with $\hat{\theta}_{MLE}(x) = \frac{x}{n}$ [unbiased $\mathbb{E}\left[\frac{X}{n}\right] = \frac{n\theta}{n} = \theta$]

to summarize:

- 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

and then analyze the statistical properties of estimator:

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

Maximum likelihood principle

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

we want to estimate/learn θ from x

$\hat{\theta}_{MLE}(x) \triangleq \arg\max_{\theta\in\Theta} p(x;\theta)$

$L \triangleq L(\theta)$ "likelihood function" 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 $L(\theta)$

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

$\triangleq \ell(\theta)$ **log likelihood**

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

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

$$\Rightarrow \operatorname*{argmax}_{\theta \in \Theta} \log p(x;\theta) = \operatorname*{argmax}_{\theta \in \Theta} p(x;\theta)$$

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

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

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

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

$x - x\theta = n\theta - \theta x$

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

Used often as solution in optimization

$\theta^* = \frac{x}{n}$