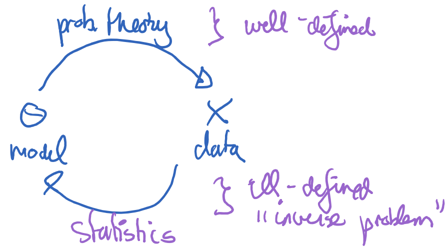
Lecture 4 - prob. / frequentist vs. Bayesian

Friday, September 16, 2022 2:57 PM

today: statistics frequentist vs. Bayesian

Statistical concepts

cartoon



epistemology of science?

example: model n indep. coin-flips

prob. theory → prob. k heads in 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 ~~a lot 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.

ie. 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 answer?

justification for frequentist semantic:

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

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

$B \triangleq \mathbb{1}_{\{X=x\}}$ $\Rightarrow B \sim \text{Bern}(\theta)$ 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 experiments i.e. $B_i \overset{iid}{\sim} \mathrm{Bern}(\theta)$

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

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

coin flips — Bayesian approach

biased coin flips

we believe $X \sim \mathrm{Bin}(n, \theta)$ — unknown $\Rightarrow$ model it as R.V. $\Rightarrow$ 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

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

$p(x)$ } normalization "marginal likelihood"

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

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

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

"proportional to" — up to a constant

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

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

$\underbrace{B(a,b)}_{\text{Beta fct.}} \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ — Γ: 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 $B(\theta | 1, 1)$
- posterior $B(\theta | x+1, n-x+1)$ — as "prior counts"

exercise to the reader: if use $B(\alpha_0, \beta_0)$ as prior posterior will be $B(x+\alpha_0, n-x+\beta_0)$

16h07

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

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

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

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

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

$$= \int_\theta \theta\, p(\theta \mid X=x)\, d\theta = \underbrace{\mathbb{E}[\theta \mid X=x]}_{\text{"posterior mean of } \theta\text{"}}$$

* a meaningful "Bayesian" estimator of θ

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

notation: $\hat{\theta}$: observation → Θ

<u>our coin example</u>: $p(\theta \mid x) = \text{Beta}(\theta \mid \alpha = x+1, \beta = n-x+1)$

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

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

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

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

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

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

<u>to summarize</u>:

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

consider multiple estimators — MLE, moment matching, ...

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

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

"likelihood function" of θ

MLE example I: binomial:

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

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

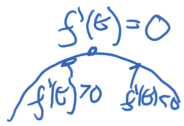
trick: to maximize $\underbrace{\log L(\theta)}_{\triangleq \ell(\theta)}$ instead of $L(\theta)$ $\quad$ $\boxed{\text{log likelihood}}$

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

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

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

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



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

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

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

$x - \cancel{x\theta} = n\theta - \cancel{x\theta}$

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

$\boxed{\theta^{(*)} = \frac{x}{n}}$ used often as solution in optimization hence $\boxed{\hat{\theta}_{MLE}(x) = \frac{x}{n}}$