Lecture 4 - prob. / frequentist vs. Bayesian

September 14, 2021 4:31 PM

today: - finish prob.
- frequentist vs. Bayesian

binomial distribution:

model n indep. coin flips

sum of n indep. Bern(θ) R.V.

let $X_i \overset{\text{iid}}{\sim} \text{Bern}(\theta)$ (mutually) independent and identically distributed

implicitly talking $\rightarrow$ $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim}$

let $X = \sum_{i=1}^{n} X_i$ then we have $X \sim \text{Bin}(n, \theta)$

"Binomial with parameter n & θ"

$\Omega_X = \{0, 1, \dots, n\}$

$\binom{n}{x} = \frac{n!}{x!(n-x)!}$

pmf: $p(x; \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$ for $x \in \Omega_X$

# of ways to choose x elements out of n "n choose x"

$p(x_1, \dots, x_n) = \prod_{i=1}^{n} \text{Bern}(x_i; \theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}$

$= \theta^{\sum x_i} (1-\theta)^{n - \sum x_i}$, where $\sum x_i = x$

mean: $X = \sum_i X_i$

$\mathbb{E}[X] = \sum_i \mathbb{E}[X_i] = n\theta$

Similarly, $\text{Var}\left(\sum_i X_i\right) \overset{\text{by indep.}}{=} \sum_i \text{Var}(X_i) = n\theta(1-\theta)$

other distributions:

Poisson(λ) $\Omega_X = \{0, 1, \dots\} = \mathbb{N}$ [count data]

λ: mean

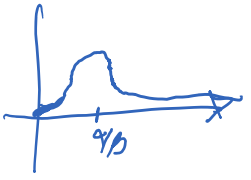
Gaussian in 1D $\mathcal{N}(\mu, \sigma^2)$ $\Omega_X = \mathbb{R}$

(μ: mean, σ^2: variance)

gamma $\Gamma(\alpha, \beta)$ $\Omega_X = \mathbb{R}^+$

α: shape, β: inverse scale aka. rate

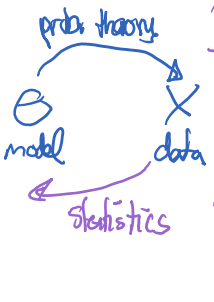
mean $\frac{\alpha}{\beta}$ variance $\frac{\alpha}{\beta^2}$



other: Laplace, Cauchy, exponential $\hookrightarrow \Gamma(1, \beta)$, beta $\rightarrow \Omega_X = [0,1]$, Dirichlet $\hookrightarrow$ Dirichlet on 2 elements

Statistical Concepts

Cartoon



} well defined

} ill defined "inverse problem"

epistemology of science

example: model n indep. coin flips

prob. theory → compute prob. k heads in a row

statistics: I have observed k heads and 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 iid 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.]

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 answers

Justification for frequentist semantic ⇝ see notes from last year (lecture4 scribbles 2020)

Coin flips → Bayesian approach

biased coin flip — unknown ⇒ model it as a R.V.

biased coin flip — unknown ⇒ model it as a R.V.

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

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

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

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

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

- $p(\Theta = \theta \mid X = x)$: posterior belief
- $p(x \mid \theta)$: observation model / likelihood
- $p(\theta)$: "prior belief"
- $p(x)$: normalization, "marginal likelihood"

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

example:

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

$$p(\theta \mid x) \propto p(x \mid \theta) p(\theta) \propto \underbrace{\theta^x (1-\theta)^{n-x}}_{p(x \mid \theta) \text{ up to a constant}} \underbrace{\mathbb{1}_{[0,1]}(\theta)}_{p(\theta)}$$

("proportional to")

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

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

$$B(a,b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

beta fct. — gamma fct.

$$\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du$$

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

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

(a, β: parameters)

- uniform distribution $B(\theta \mid 1, 1)$
- posterior $B(\theta \mid x+1, n-x+1)$

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

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

17h37

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

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

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

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

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

(by our model)

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

"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 → Θ

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

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

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

↳ 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"

Obtain multiple estimators
- MLE
- moment matching
- Bayesian posterior mean
- MAP
- regularized MLE

and then analyze the statistical properties:
- biased?
- variance?
- consistent?

Maximum likelihood principle

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

we want to estimate/learn θ from x

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

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

$\mathcal{L} \triangleq L(\theta)$ "likelihood function" of θ

$p(x|\theta)$

MLE example I: binomial:

n coin flips $\quad \Omega_x = 0:n$

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

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

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

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

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

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

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

$f'(\theta)=0$

$f(\theta)\nearrow \quad f(\theta)\searrow$

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

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

$x(1-\theta) - \theta(n-x) = 0$

$\theta^* = x$ use often as estimator

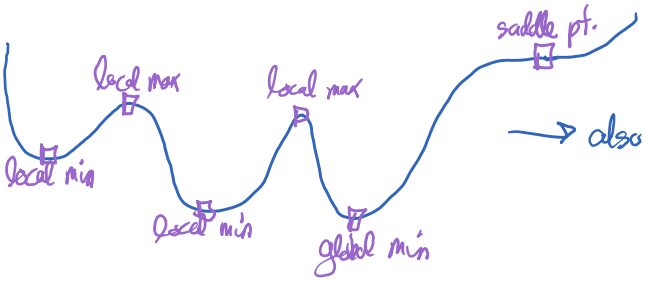
$x(1-\theta) - \theta(n-x) = 0$

$\theta = \frac{x}{n}$ — use often as soln in optimization

hence $\hat{\theta}_{ML}(x) = \frac{x}{n}$ i.e. relative frequency

same optimization comments:

$\min_{\theta \in \Theta} f(\theta)$ (*) $\nabla f(\tilde{\theta}) = 0$ ("stationary pts.") (if f is diff) on Θ is a necessary condition for $\tilde{\theta}$ being a local min when $\tilde{\theta}$ is in the interior of Θ



→ also need to check Hessian$(f)(\tilde{\theta}) \geq 0$ for a local min

scalar: $(f'')(\tilde{\theta}) > 0$

$H \geq 0 \iff u^T H u \geq 0 \quad \forall u \neq 0 \in \mathbb{R}^d$

(*) only local result in general

- but if Hessian$(f(\theta)) \geq 0 \ \forall \theta \in \Theta$, fct. is said "convex"

  and in this case, $\nabla f(\tilde{\theta}) = 0 \Rightarrow$ sufficient for $\tilde{\theta}$ to be a global min

- otherwise, for smooth fct, look at zero gradient pts and boundary pt. give you enough information to find global optima