Lecture 4 - scribbles

Friday, September 15, 2017
13:23

today: - continue Bayesian approach
- MLE

(continuation from last class:)

as a Bayesian, posterior $p(\theta \mid \underbrace{X=x}_{\text{observation}})$ contains all info we need

e.g. here is a question: what is the probability that the next coin flip is $=1$

$$P(\underset{F}{\text{next-flip}}=1 \mid \underset{\Theta=\theta}{\theta}) = \theta$$

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

(always true) $= \int_\theta P(F=1 \mid \theta, X=x)\, P(\theta \mid X=x)\, d\theta$

|| (by our model)

$\underbrace{P(F=1 \mid \theta)}_{\theta}$

conditional expectation

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

"posterior mean" of θ

a meaningful Bayesian estimator of θ

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

statistical

notation: $\hat{\theta}: obs \to \Theta$: "statistical estimator"

"frequentist statistics" (traditional statistics)

consider multiple possible estimators
- MLE
- moment matching
- bayesian posterior mean
- MAP

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

---

coming back to coin example: $p(\theta | X=x) = Beta(\theta; \underbrace{\alpha = x+1, \beta = n-x+1}_{\text{parameters}})$

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

thus $\boxed{\mathbb{E}[\theta | X=x] = \frac{x+1}{n+2} = \hat{\theta}_{Bayes}(x)}$ here (X here is a binomial R.V.) (biased, but asymptotically unbiased)

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

ie. $\mathbb{E}_{X|\theta}[\hat{\theta}_{MLE}(X)] = \theta$

Maximum likelihood principle

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

we want to estimate θ

$\hat{\theta}_{ML}(x) \triangleq \underset{\theta\in\Theta}{\operatorname{argmax}}\ p(x;\theta)$ i.e. $\hat{\theta}_{ML}(x)$ maximizes $p(x;\cdot)$

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

example: 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 $L(\theta)$

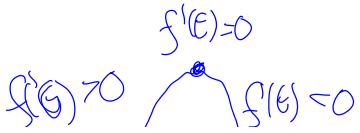
justification: $\log(\cdot)$ is strictly increasing

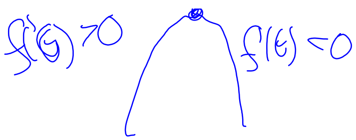
i.e. $a < b \iff \log a < \log b$

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

log-likelihood $\log p(x;\theta)$

$= \underbrace{\log\binom{n}{x}}_{\text{constant}} + x\log\theta + (n-x)\log(1-\theta) = \ell(\theta)$





look for $\frac{\partial l(\theta)}{\partial \theta} = 0$

i.e. $\frac{x}{\theta} - \frac{n-x}{1-\theta} = 0$

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

$\theta^* = \frac{x}{n}$ ← often used when optimizing

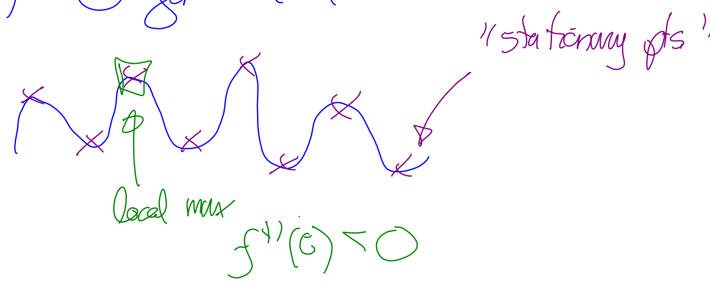
here $\boxed{\hat{\theta}_{ML}(x) = \frac{x}{n}}$

i.e. relative frequency

some optimization comments:

($\nabla f(\theta) = 0$) • $f'(\theta) = 0$ is necessary condition for local max when θ in interior(Θ)

→ also need to check $f''(\theta) < 0$ for a local max



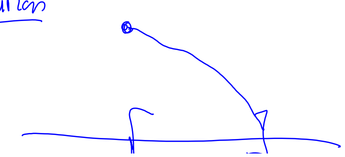
dxd matrix → ($\text{Hessian}(\theta) \preceq 0$) negative semi-definite

→ only local result in general

but if $f''(\theta) \leq 0 \ \forall \theta \in \Theta$, function is "concave"

in this case, $f'(\theta) = 0$ is sufficient condition for global max

• be careful with boundary cases i.e. $\theta^* \in$ boundary(Θ)



Some notes about MLE

- does not always exist [e.g. $\hat{\theta} \in \text{bd}(\Theta)$ but Θ is open]
  e.g. $\Theta =]0,1[$
- is not necessarily unique (i.e. multiple maxima)
- is not "admissible" in general [see next class]

Example #2: Multinomial dist.

Suppose X_i is discrete R.V. on K choices → Multinoulli

(we could choose $\Omega_{X_i} = \{1, \dots, K\}$)

but instead, convenient to encode with the unit basis in $\mathbb{R}^K$

i.e. $\Omega_{X_i} = \{e_1, \dots, e_K\}$ where $e_i \in \mathbb{R}^K$ "one hot encoding"

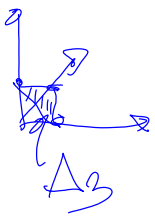
$e_i = (0, \dots, 0, 1, 0, \dots, 0)$ ← i^{th} coordinate

parameter for discrete R.V. $\pi \in \Delta_K$

probability simplex on K choices

$$\Delta_K \triangleq \{\pi \in \mathbb{R}^K \text{ s.t. } \pi_j \geq 0 \ \forall j, \ \sum_{j=1}^{K} \pi_j = 1\}$$

$(\Theta = \Delta_K)$



we will write $X_i \sim \text{Mult}(\pi)$ "multinoulli"

parameter

* Consider $X_i \overset{iid}{\sim} \text{Mult}(\pi)$

then $X = \sum_{i=1}^{n} x_i \sim \text{Mult}(n, \pi)$ (analog of binomial for k choices)

$x_i \in \mathbb{N}^k$; "multinomial"

$$\Omega_X = \left\{ (n_1, \ldots, n_k) : n_k \in \mathbb{N}, \sum_{j=1}^{k} n_j = n \right\}$$

consider MLE for R.V. $X = \sum_{i=1}^{n} x_i$ (vector)

$p(x|\pi)$ $\qquad p(x_i|\pi) = \text{Mult}(x_i|\pi)$

$= \prod_{j=1}^{k} \pi_j^{x_{ij}}$ ← j^{th} component of vector x_i

$$p(x|\pi) = \prod_{i=1}^{n} p(x_i|\pi) = \prod_{i=1}^{n} \left(\prod_{j=1}^{k} \pi_j^{x_{ij}} \right)$$

by indep.

$$= \prod_{j=1}^{k} \left(\prod_{i=1}^{n} \pi_j^{x_{ij}} \right)$$

$$= \prod_{j=1}^{k} \pi_j^{\sum_{i=1}^{n} x_{ij}}$$

$\sum_{i=1}^{n} x_{ij} = n_j$

log-likelihood: $\ell(\pi) = \log p(x|\pi) = \sum_{j=1}^{k} n_j \log \pi_j$

↳ note that $n_j(x)$

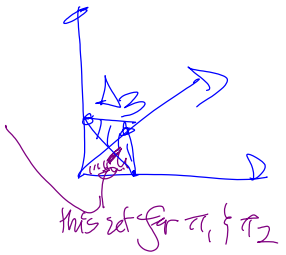
we want $\max_{\pi} \ell(\pi)$

s.t. $\pi \in \Delta_k$ } constraints

two options a) could reparameterize with $\pi_1, \dots, \pi_{K-1} \in [0,1]$

with constraint $\sum_{j=1}^{K-1} \pi_j \leq 1$

and use $\pi_K = 1 - \sum_{j=1}^{K-1} \pi_j$



and then do unconstrained optimization over $\pi_1, \dots, \pi_{K-1}$

b) use Lagrange multipliers approach for equality constraint $\sum_{j=1}^{K} \pi_j = 1$

$\Rightarrow 1 - \sum_{j=1}^{K} \pi_j = 0$

$g(\pi) = 0$

look for stationary points (0 gradient) of

$J(\pi, \lambda) \triangleq \ell(\pi) + \lambda \underbrace{\left(1 - \sum_{j=1}^{K} \pi_j\right)}_{g(\pi)}$

λ: Lagrange multiplier

$\max f(\pi)$ s.t. $g(\pi) = 0$

i.e. want $\nabla_\pi J(\pi, \lambda) = 0$ and $\nabla_\lambda J(\pi, \lambda) = 0 \rightarrow$ equivalent $g(\pi) = 0$

$\ell(\pi) = \sum_{j=1}^{K} N_j \log \pi_j$ $\quad \frac{\partial J}{\partial \pi_j} = 0 \Rightarrow \frac{N_j}{\pi_j} - \lambda = 0$

$\Rightarrow \pi_j^* = \frac{N_j}{\lambda}$ λ: scaling constant

want $\sum_j \pi_j^* = 1 \Rightarrow \lambda = \sum_j N_j = N$

$\Rightarrow \boxed{\pi_j^* = \frac{N_j}{N}}$ MLE for multinomial

Similarly: $\pi_j = N_j \geq 0$

Sidenote: $\pi_j = \frac{n_j}{n} \geq 0$

picture behind Lagrange multiplier technique

$$J(x, \lambda) = f(x) + \lambda g(x)$$

$$D_x J(x, \lambda) = 0 \Rightarrow Df(x) = -\lambda Dg(x)$$