Lecture 4 - scribbles

Friday, September 14, 2018 13:34

today: - continue Bayesian approach
- MLE

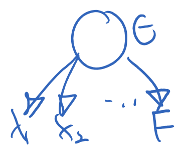
(continuation from last class):

posterior belief $p(\theta \mid \underbrace{X=x}_{\text{observations}})$ contains all the info from data that we need to answer new questions

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

as frequentist: $p(F=1 \mid \theta) = \hat{\theta}$

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



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

|| (by our model)

$P(F=1 \mid \theta) = \theta$

$p(x, y \mid z) = p(x \mid y, z)\, p(y \mid z)$

$p(x, y \mid z, w) = p(x \mid y, z, w)\, p(y \mid z, w)$

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

a meaningful Bayesian estimator of θ

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 $\to \Theta$)

$p(\theta) = \mathrm{unif}(\theta)$
$= \mathrm{Beta}(\theta \mid \alpha=1, \beta=1)$

our coin example: $p(\theta \mid X=x) = \mathrm{Beta}(\theta \mid \alpha = x + \underbrace{\boxed{1}}_{\alpha_{prior}}, \beta = n-x+\underbrace{\boxed{1}}_{\beta_{prior}})$

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

thus $\boxed{\hat{\theta}_{Bayes}(x) = \mathbb{E}[\theta \mid X=x] = \frac{x+1}{n+2}}$ ← here biased → but asymptotically unbiased

i.e. $\mathbb{E}_X \hat{\theta}(X) \neq \theta$

asympt. unbiased

$$\mathbb{E}[\hat{\theta}_{Bayes}(X)] = \frac{\mathbb{E}X+1}{n+2} = \frac{n\theta+1}{n+2} \xrightarrow[n\to\infty]{} \theta$$

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

i.e. $\mathbb{E}_{X\mid\theta}[\hat{\theta}_{MLE}(X)] = \frac{\theta n}{n}$
$= \theta$

to summarize:

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

consider multiple possible estimators ← MLE
moment matching
Bayesian posterior mean

moment matching
Bayesian posterior mean
MAP

and then analyze their statistical properties

- biased?
- variance?
- consistent?

14h20

Maximum likelihood principle

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

• we want to estimate θ

$$\hat{\theta}_{ML}(x) \triangleq \underset{\theta \in \Theta}{\operatorname{argmax}} \; p(x;\theta)$$

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

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

example: n coin flips $\quad \Omega_X = 0:n$

$X \sim \mathrm{Bin}(n,\theta)$

$$p(x;\theta) = \binom{n}{x} \theta^{x} (1-\theta)^{n-x}$$

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

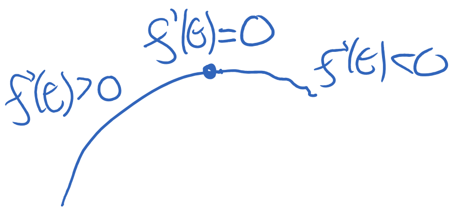
$\triangleq \ell(\theta)$ $\boxed{\text{log likelihood}}$

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

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

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

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



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

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

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

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

often used in optimization

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

i.e. relative frequency

Some optimization comments

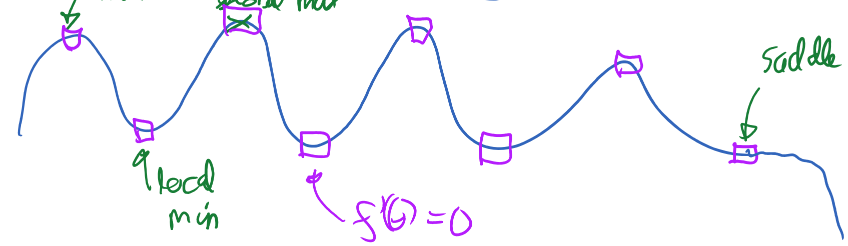
in multiple dim: $(\nabla f(\theta) = 0)$

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

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

  local max ✓ global max

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



d×d matrix

"curved"

Hessian$(\theta) \preceq 0$ ⟸ $u^T A u \leq 0$ $\forall u \in \mathbb{R}^d$

i.e. negative semi-definite

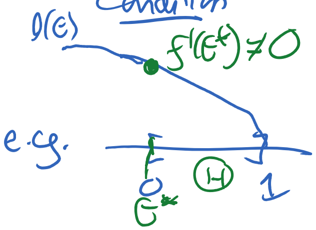
→ only local result in general

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

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

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

e.g.



another example

$\Theta = [0, +\infty[$

Some notes about MLE

• does not always exist [$\theta^* \in bd(\Theta)$ but Θ is open] or when "$\theta^* = +\infty$"

e.g. $\Theta =]0,1[$

• is not necessarily unique [i.e. multiple maxima]

• is not "admissible" in general [see next class]
(basically, ∃ strictly "better" estimators)

example 2 : multinomial distribution

suppose X_i is discrete R.V. on K choices "Multinoulli"

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

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

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

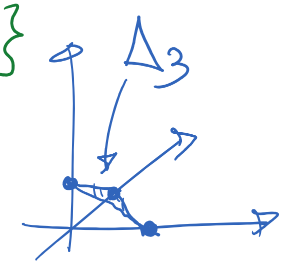
$$e_j = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \leftarrow j^{th} \text{ coordinate}$$

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

$(\Theta = \Delta_K)$

$$\Delta_K \triangleq \{\pi \in \mathbb{R}^K : \pi_j \geq 0 \; \forall j ; \sum_{j=1}^{K} \pi_j = 1\}$$

probability simplex on K choices



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

parameter

(*) consider $X_i \overset{iid}{\sim} \mathrm{Mult}(\pi)$

then $X \triangleq \sum_{i=1}^{n} X_i \sim \mathrm{Mult}(n, \pi)$

"multinomial" distribution

$X \in \mathbb{N}^K$

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