Lecture 4 - scribbles

Friday, September 13, 2019 14:03

today: • Bayesian approach
• MLE

coin flips – Bayesian approach:

biased coin flips

we believe $X \sim Bin(n, \Theta)$

Θ: unknown $\Rightarrow$ model it as a R.V.

$\Rightarrow$ need a $p(\theta)$ "prior distribution"

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

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

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

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

$p(X = x \mid \theta)$: observation model / likelihood

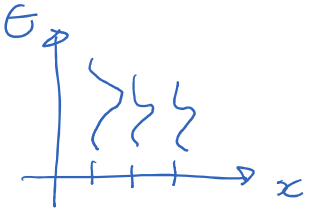
$p(x)$: normalization

$P(\{\Theta = \theta\})$

$p(\Theta = \theta)$ I mean pdf of R.V. Θ $p(\theta)$

[note: $p(x|\theta)$ → is a pmf $p(x, \theta)$ is a "mixed" dist.

[note: $p(x|\theta)$ → is a pmf, $p(\theta)$ → pdf $p(x,\theta)$ is a "mixed" dist.



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

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

"proportional to"

$\binom{n}{x}$

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

normalization constant so that $\int_\theta p(\theta|x)\, d\theta = 1$

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

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

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 here was $B(\theta|x+1, n-x+1)$

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

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

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

e.g. Question: what is probability of head $(\overbrace{F=1}^{\text{flip outcome}})$ on the next flip

as a frequentist: $P(F=1|\text{data}) = \hat{\theta}$

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

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

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

"posterior mean" of θ

* a meaningful "Bayesian" estimator of θ

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

notation: $\hat{\theta}$: observation $\rightarrow \Theta$

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

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

but asymptotically 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: $\mathbb{E}\hat{\theta}_{MLE}(X) = \frac{\mathbb{E}X}{n} = \frac{n\theta}{n} = \theta$

to summarize

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

Consider multiple possible estimators
- MLE
- moment matching
- Bayesian posterior mean
- MAP
- regularized MLE

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

15h35

Maximum likelihood principle

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

we want to estimate/learn θ

$\hat{\theta}_{ML}(x) \triangleq \underset{\theta \in \Theta}{\arg\max}\ p(x;\theta)$ $\quad \hat{\theta}_{ML}(x)$ maximizes $p(x;\cdot)$

$\ell \triangleq L(\theta)$

"likelihood function" of θ

Example: n coin flips

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

$x = 0:n$

$X \sim \text{Bin}(n,\theta)$ $\quad \downarrow x = 0:n$

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

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

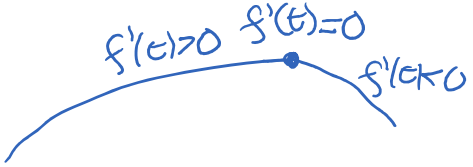
$\triangleq \ell(\theta)$ log-likelihood

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

ie. $a < b \iff \log a < \log b$

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

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



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

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

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

used often as solution in optimization

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

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

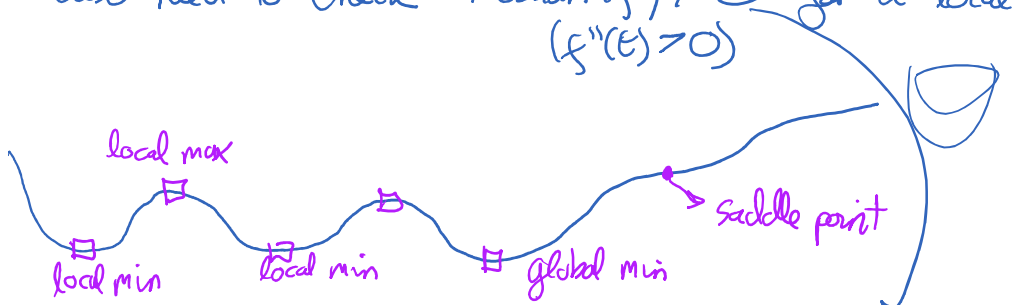
hence $\boxed{\hat{\theta}_{ML}(x) = \frac{x}{n}}$ i.e. relative frequency

Some optimization comments

$\min_{\theta \in \Theta} f(\theta)$ (*) $\nabla f(\theta) = 0$ is a necessary condition for a local min when θ is in the interior of Θ (when f is differentiable on Θ)

"stationary pts."

→ also need to check Hessian$(f) > 0$ ($f''(\theta) > 0$) for a local min



$H > 0$
$\Leftrightarrow$
$u^T H u > 0$
$\forall u \neq 0 \in \mathbb{R}^d$

(*) only local result in general

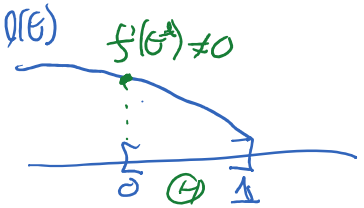
- but if Hessian$(f(\theta)) \geq 0$ $\forall \theta \in \Theta$, fct. is said "convex"

and in this case $\nabla f(\theta) = 0 \Rightarrow$ sufficient for global min

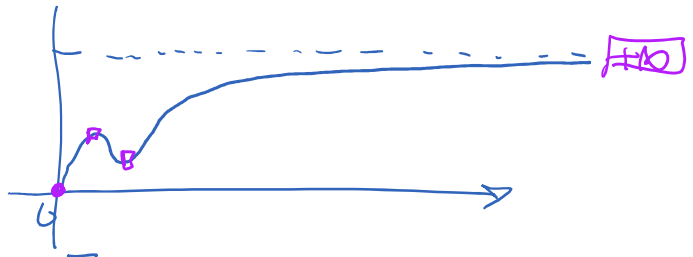
- otherwise, for a smooth fct., looking at zero gradient + boundary gives you enough information

⊛ be careful with boundary cases i.e. $\theta^* \in \text{boundary}(\Theta)$

e.g.



another example



example where the MLE does not exist

⊛ Some notes about MLE

- does not always exist [$\theta^* \in \text{bd}(\Theta)$ and Θ is open] or when "$\theta^* = +\infty$"

  $\Theta =]0,1[$

- is not necessarily unique [i.e. multiple maxima
  e.g. mixture models]
- is not "admissible" in general [see later]
  ∃ strictly "better" estimators

15h36

example 2 : multinomial distribution

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 unit basis in $\mathbb{R}^K$

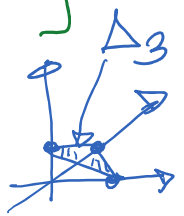
i.e. $\Omega_{X_i} = \{e_1, \dots, e_K\}$ where $e_j \in \mathbb{R}^K$ "one hot encoding"

$e_j = (0, \dots, 0, \underset{j^{th} \text{ coordinate}}{1}, 0, \dots, 0)$

parameter for discrete R.V. : $\pi \in \Delta_K$ $\quad (\Theta = \Delta_K)$

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

probability simplex on K choices



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

π : parameter

⊗ consider iid. $X_i \sim \text{Mult}(\pi)$

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

"multinomial distribution"

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

$$\Omega_X = \{ (n_1, \dots, n_K) : n_j \in \mathbb{N}, \sum_{j=1}^{K} n_j = n \}$$

pmf for X:

$$p(x \mid \pi) = \binom{n}{n_1, \dots, n_K} \prod_{j=1}^{K} \pi_j^{(x)_j} \qquad x = (n_1, \dots, n_K)$$

multinomial coef. $\binom{n}{n_1 \dots n_K} \triangleq \frac{n!}{n_1! \cdots n_K!}$