Lecture 4 - scribbles
Friday, September 13, 2019 14:03

Today: Bayesian approach
MLE

Coin Flips - Bayesian Approach:

Biased Coin Flips

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

unknown \( \Rightarrow \) model it as a R.V.

\[ \theta = [0, 1] \]

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

then we can "update" our belief about \( \theta \) using Bayes rule

\[
p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}
\]

\( p(x) \) = normalization "marginal likelihood"

\[ p(\theta|x) \] is posterior belief

\[ p(x|\theta) \] = observation model

\[ p(\theta) \] = prior belief

\[ p(\theta) \] is a "mixture" dist.

Note: \( p(\theta|x) \) is a pmf

\( p(x|\theta) \) is a pdf
\[ \text{Note: } p(x|\theta) \xrightarrow{\theta \to 0} p(x) \quad \text{is a pdf} \quad p(x) \quad \text{is a "mixed" dist.} \]

**Example:**

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

\[ p(G|\theta) \propto \theta^{(1-\theta)^{n-2}} \chi_{[0,1]}(\theta) \]

"proportional to" \( p(x|\theta) \) up to \( p(\theta) \) scaling

\[ \left( \frac{n}{x} \right) \]

Scaling : \[ \int_{0}^{1} \theta^{(1-\theta)^{n-2}} d\theta = B(x+1, n-x+1) \]

Normalization constant \( \infty \) so that \[ \int_{0}^{1} p(G|\theta) \, d\theta = 1 \]

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

\( \Gamma(a) = \int_{0}^{\infty} u^{a-1} e^{-u} \, du \)

Here, \( p(G|\theta) \) is called a "beta distribution"

\[ B(\theta|\alpha, \beta) \triangleq \frac{\Gamma(n-1)(1-\theta)^{n-2}}{\Gamma(n)B(\alpha, \beta)} \chi_{[0,1]}(\theta) \]

Parameters
- uniform distribution: \( B(\theta | 1, 1) \)
- posterior here was \( B(\theta | x+1, n-x+1) \)

exercise to the reader: if we \( B(\theta_0 | \theta_0) \) as prior

posterior will be \( B(\theta | x+b_0, n-x+b_0) \)

posterior belief \( \mathbb{P}(\theta | x=x) \) contains all the info from data \( x \) that we need to answer new queries

eg. question: what is probability of head \( (F=1) \) on the next flip

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

as a Bayesian \( \mathbb{P}(F=1 | x=x) = \int \mathbb{P}(F=1 | \theta, x=x) d\theta \)

\( = \mathbb{E}[\theta | x=x] \) [by our model]
a meaningful "Bayesian" estimator of $\theta$

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

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

Our case example: $p(\Theta | x) = \text{Beta}(\Theta | x=x+1, \beta = n-x+1)$

mean of a beta r.v. $\frac{\alpha}{\alpha + \beta}$

Thus:

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

Here, biased estimator $\mathbb{E}_x[\hat{\theta}(x)] \neq \Theta$

but asymptotically unbiased

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

Compare/contrast with $\hat{\theta}_{\text{MLE}}(x) = \frac{x}{n}$

Unbiased: $\mathbb{E}\hat{\theta}_{\text{MLE}}(x) = \frac{\mathbb{E}X}{n} = \frac{n\bar{X}}{n} = \Theta$

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

consider multiple possible estimators

and then analyze their 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 \( \theta \)

\[ \hat{\theta}_{ML}(x) = \arg \max_{\theta \in \Theta} p(x; \theta) \quad p \in L(\Theta) \]

"Likelihood function" of \( \theta \)

**Example:** \( \theta \) can flips
\( X \sim \text{Bin}(n, \theta) \)
\( l_x = 0 : n \)
\[ X \sim \text{Bin}(n, \theta) \]

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

**Trick:** to maximize \( \log L(\theta) \) instead of \( L(\theta) \)

\[ \Rightarrow \log L(\theta) \quad \text{[log-likelihood]} \]

**Justification:** \( \log(.) \) is strictly increasing

i.e. \( a < b \iff \log a < \log b \)

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

\[ \log p(x; \theta) = \log \binom{n}{x} + x \log \theta + (n-x) \log (1-\theta) = \ell(\theta) \]

- constant.

\[ f'(\theta) = 0 \quad f''(\theta) < 0 \]

look for \( \theta \) s.t. \( \ell'(\theta) = 0 \)

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

\[ x(1-\theta) - (n-x)\theta = 0 \]

\[ \hat{\theta}_{\text{ML}}(x) = \frac{x}{n} \]

hence \( \hat{\theta}_{\text{ML}}(x) = x \) i.e. relative
Some optimization comments:

\[ \min_{\Theta} f(\Theta) \quad \implies \quad \nabla f(\Theta) = 0 \]

("stationary pt")

is a necessary condition for a local min when \( \Theta \) is in the interior of \( \Theta \) (when \( f \) is differentiable).

\[ \nabla^2 f(\Theta) > 0 \] also need to check Hessian for a local min

\[ f''(\Theta) > 0 \]

\( \mathbf{H} > 0 \)

\[ \forall \mathbf{u} \in \mathbb{R}^d \]

\( \mathbf{H} \mathbf{u} > 0 \)

\( \Rightarrow \)

only local result in general

but if Hessian \( f''(\Theta) > 0 \) for all \( \Theta \), function is said "convex"

and in this case \( \nabla f(\Theta) = 0 \) \( \Rightarrow \) sufficient for global min.
a otherwise, for a smooth fct., checking at zero gradient & boundary gives you enough information

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

  e.g. \( f(0) \neq 0 \)

  \[
  \begin{array}{c}
  \text{Another example} \\
  \end{array}
  \]

- example where the MLE does not exist

Some notes about MLE

- does not always exist \([ \Theta^* \in \text{bd}(\Theta) \text{ but } \Theta^* \text{ is open} ] \text{ or when } \Theta^* = \{ \theta_0 \} \]

  \( \Theta_0 = [0, 1] \)

- is not necessary unique \( \text{ i.e. multiple maxima } \)

  e.g. mixture models

- is not "admissible" in general \( \text{ [see later] } \)

  stricty "better" estimates
Example 2: Multinomial distribution

Suppose $X_i$ is discrete R.V. on $k$ choices \( \text{"multinoulli"} \)

(we could choose $\Omega = \{1, \ldots, k\}$

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

i.e. $\sum x_i = e_1, \ldots, e_k$ where $e_j \in \mathbb{R}^k$ \( \text{"one hot encoding"} \)

$e_j = (0, \ldots, 1, \ldots, 0)$ $j$th coordinate

Parameter for discrete R.V.: $\pi \in \Delta_k$ \( \text{($\Omega = \Delta_k$)} \)

\[ \Delta_k = \{ \pi \in \mathbb{R}^k : \pi_j > 0 \text{ all } j, \sum_{j=1}^k \pi_j = 1 \} \]

\( \text{probability simplex on } k \text{ choices} \)

We will write $X \sim \text{Mult}(\pi)$ \( \text{"multinoulli"} \)

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

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

\( \text{\"multinomial distribution\"} \)
$X \in \mathbb{N}^K$

$\mathcal{L}_X = \{ (n_1, \ldots, n_K) : n_j \in \mathbb{N}, \sum_{j=1}^K n_j = N \}$

pmf for $X$:

$p(x | \pi) = \binom{n}{n_1 \ldots n_K} \prod_{j=1}^K \pi_j^{n_j}$

$x = \{ n_1, \ldots, n_K \}$

2. multinomial coeff. \[
\binom{n}{n_1 \ldots n_K} \stackrel{\mathbb{N}}{=} \frac{n!}{n_1! \cdots n_K!}
\]