Welcome! Ask Qs or say hi in chat before/during/after class

CSE 312
Foundations of Computing II

Lecture 24: Maximum Likelihood Estimation (MLE)

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Rachel Lin, Hunter Schafer

Slide Credit: Based on Stefano Tessaro’s slides for 312 19au incorporating ideas from Alex Tsun’s and Anna Karlin’s slides for 312 20su and 20au

Music: Carly Rae Jepsen
Agenda

- Idea: Estimation
- Maximum Likelihood Estimation (example: mystery coin)
- Continuous random variables
- General Steps
Probability: View Point up to Now

\( X \sim \text{Ber}(\theta) \)
\[ \mathbb{P}(X=1; \theta) = \theta \]

\( \theta \) = known parameter

\( \theta \) tells us how samples are distributed.
\( \mathbb{P}(x; \theta) \) viewed as a function of \( x \) (fixed \( \theta \))
Statistics: Parameter Estimation – Workflow

$X_i \sim \text{Ber}(\theta)$

$\theta = \text{unknown}$ parameter

Don’t know how samples are distributed.

$\mathcal{L}(x|\theta)$ viewed as a function of $\theta$ (fixed $x$)

**Example:** $\mathcal{L}(x|\theta) = \text{coin flip distribution with unknown } \theta = \text{probability of heads}$

Observation: HTTHHHHTHTTTTHTHTTTTTTH

**Goal:** Estimate $\theta$ from data
Example

Suppose we have a mystery coin with some probability $p$ of coming up heads. We flip the coin 8 times, independent of other flips and see the following sequence of flips:

$\text{TTHTHTTH}$

Given this data, what would you estimate $p$ is?

Poll: pollev.com/hunter312

- a. $\frac{1}{2}$
- b. $\frac{5}{8}$
- c. $\frac{3}{8}$
- d. $\frac{1}{4}$
Agenda

• Idea: Estimation
• Maximum Likelihood Estimation (example: mystery coin)
• Continuous random variables
• General Steps
Likelihood

\[ X_i \sim \text{Ber}(\theta) \]

\[ P(X_i = H \mid \theta) = \theta \]

Say we see outcome HHTHH.

\[ \mathcal{L}(\text{HHTHH} \mid \theta) = \theta^4(1 - \theta) \]

Probability of observing the outcome HHTHH if \( \theta = \text{prob. of heads} \). This is a function of \( \theta \).

\[ \frac{d}{d\theta} \mathcal{L}(\text{HHTHH} \mid \theta) = \frac{d}{d\theta} \theta^4(1 - \theta) \]

\[ = \frac{d}{d\theta} \theta^4 - \theta^5 \]

\[ = 4\theta^3 - 5\theta^4 \]

\[ 4\hat{\theta}^3 - 5\hat{\theta}^4 = 0 \]

\[ \hat{\theta}^3 (4 - 5\hat{\theta}) = 0 \]

\[ \hat{\theta} = 0 \text{ or } \left( \frac{4}{5} \right) \]

Max Prob of seeing HHTHH

Technically need 2nd derivative test.

But we skip this step in 312 this quarter.
Definition. The **likelihood** of independent observations \( x_1, \ldots, x_n \) is
\[
L(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} P(x_i; \theta)
\]

**Maximum Likelihood Estimation (MLE).** Given data \( x_1, \ldots, x_n \), find \( \hat{\theta} \) ("the MLE") of model such that \( L(x_1, \ldots, x_n | \hat{\theta}) \) is maximized!
\[
\hat{\theta} = \arg\max_\theta L(x_1, \ldots, x_n | \theta)
\]

Usually: Solve \( \frac{\partial L(x_1, \ldots, x_n | \theta)}{\partial \theta} = 0 \) or \( \frac{\partial \ln L(x_1, \ldots, x_n | \theta)}{\partial \theta} = 0 \) [+check it’s a max!]
Likelihood vs. Probability

A probability function \( \Pr(x ; \theta) \) is a function with input being an event \( x \) for some fixed probability model (w/ param \( \theta \)).

\[
\sum \Pr(x ; \theta) = 1
\]

A likelihood function \( \mathcal{L}(x | \theta) \) is a function with input being \( \theta \) (the param of the prob. Model) for some fixed dataset \( x \).

These notions are very closely connected, but answer different questions. We are trying to find the \( \theta \) that maximizes likelihood, thus we are looking for the maximum likelihood estimator.
Example – Coin Flips

Observe: Coin-flip outcomes $x_1, \ldots, x_n$, with $n_H$ heads, $n_T$ tails

- i.e., $n_H + n_T = n$

Goal: estimate $\theta = \text{prob. heads.}$

\[
L(x_1, \ldots, x_n | \theta) = \theta^{n_H} (1 - \theta)^{n_T}
\]

\[
\frac{\partial}{\partial \theta} L(x_1, \ldots, x_n | \theta) = ???
\]

While it is possible to compute this derivative, it’s not always nice since we are working with products.

\[X_i \sim \text{Ber}(\theta)\]

\[
\Pr(X_i = 1; \theta) = \theta \quad \Pr(X_i = 0; \theta) = 1 - \theta
\]
Log-Likelihood

We can save some work if we work with the log-likelihood instead of the likelihood directly.

Definition. The log-likelihood of independent observations $x_1, \ldots, x_n$ is

$$\mathcal{L}(x_1, \ldots, x_n | \theta) = \ln \mathcal{L}(x_1, \ldots, x_n | \theta)$$

$$= \ln \prod_{i=1}^{n} \mathbb{P}(x_i; \theta) = \sum_{i=1}^{n} \ln \mathbb{P}(x_i; \theta)$$

Useful log properties

$$\log(ab) = \log(a) + \log(b)$$

$$\log(a/b) = \log(a) - \log(b)$$

$$\log(a^b) = b\log(a)$$
Example – Coin Flips

Observe: Coin-flip outcomes $x_1, \ldots, x_n$, with $n_H$ heads, $n_T$ tails

- I.e., $n_H + n_T = n$

Goal: estimate $\theta = \text{prob. heads.}$

$$\mathcal{L}(x_1, \ldots, x_n | \theta) = \theta^{n_H} (1 - \theta)^{n_T}$$

$$\ln \mathcal{L}(x_1, \ldots, x_n | \theta) = n_H \ln \theta + n_T \ln(1 - \theta)$$

$$\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \ldots, x_n | \theta) = n_H \cdot \frac{1}{\theta} - n_T \cdot \frac{1}{1 - \theta}$$

Solve $n_H \cdot \frac{1}{\theta} - n_T \cdot \frac{1}{1 - \theta} = 0$

$$\hat{\theta} = \frac{n_H}{n}$$
Brain Break
Agenda

- Idea: Estimation
- Maximum Likelihood Estimation (example: mystery coin)
- Continuous random variables
- General Steps
The Continuous Case

Given \( n \) samples \( x_1, \ldots, x_n \) from a Gaussian \( \mathcal{N}(\mu, \sigma^2) \), estimate \( \theta = (\mu, \sigma^2) \)

Definition. The likelihood of independent observations \( x_1, \ldots, x_n \) is

\[
\mathcal{L}(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} f(x_i; \theta)
\]

Density function! (Why?)
Why density?

• Density ≠ probability, but:
  – For maximizing likelihood, we really only care about relative likelihoods, and density captures that
  – has desired property that likelihood increases with better fit to the model
$n$ samples $x_1, \ldots, x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, 1)$. Most likely $\mu$?

[i.e., we are given the promise that the variance is one]
$n$ samples $x_1, \ldots, x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, 1)$. Most likely $\mu$?

\[ \mu = 0? \]

Unlikely …
$n$ samples $x_1, \ldots, x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, 1)$. Most likely $\mu$?

$\mu = 3$?

Better, but optimal?
Example – Gaussian Parameters

\[ x_i \sim \mathcal{N}(\theta, 1) \]

Normal outcomes \( x_1, \ldots, x_n \), known variance \( \sigma^2 = 1 \)

**Goal:** estimate \( \theta \) expectation

\[
\mathcal{L}(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - \theta)^2}{2}} = \left( \frac{1}{\sqrt{2\pi}} \right)^n \prod_{i=1}^{n} e^{-\frac{(x_i - \theta)^2}{2}}
\]

\[
\ln \mathcal{L}(x_1, \ldots, x_n | \theta) = -n \frac{\ln 2\pi}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2}
\]
Example – Gaussian Parameters

Normal outcomes $x_1, \ldots, x_n$, known variance $\sigma^2 = 1$

\[
\ln \mathcal{L}(x_1, \ldots, x_n | \theta) = -n \frac{\ln 2\pi}{2} - \frac{1}{2} \sum_{i=1}^{n} (x_i - \theta)^2
\]

Note:
\[
\frac{\partial}{\partial \theta} \frac{(x_i - \theta)^2}{2} = \frac{1}{2} \cdot 2 \cdot (x_i - \theta) \cdot (-1) = \theta - x_i
\]

\[
\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \ldots, x_n | \theta) = \sum_{i=1}^{n} (x_i - \theta) = \sum_{i=1}^{n} x_i - n\theta = 0
\]

\[
\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}
\]

In other words, MLE is the sample mean of the data.
Next: $n$ samples $x_1, ..., x_n \in \mathbb{R}$ from Gaussian $\mathcal{N} (\mu, \sigma^2)$. Most likely $\mu$ and $\sigma^2$?
Agenda

• Idea: Estimation
• Maximum Likelihood Estimation (example: mystery coin)
• Continuous random variables
• General Steps
General Recipe

1. **Input** Given \( n \) iid samples \( x_1, \ldots, x_n \) from parametric model with parameters \( \theta \).

2. **Likelihood** Define your likelihood \( L(x_1, \ldots, x_n | \theta) \).
   - For discrete \( L(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} \Pr(x_i ; \theta) \)
   - For continuous \( L(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} f(x_i ; \theta) \)

3. **Log** Compute \( \ln L(x_1, \ldots, x_n | \theta) \)

4. **Differentiate** Compute \( \frac{\partial}{\partial \theta} \ln L(x_1, \ldots, x_n | \theta) \)

5. **Solve for** \( \hat{\theta} \) by setting derivative to 0 and solving for max.

Generally, you need to do a second derivative test to verify it is a maximum, but we won’t ask you to do that in CSE 312.