Learning From Data: MLE

Maximum Likelihood Estimators
Parameter Estimation

• Assuming sample \( x_1, x_2, \ldots, x_n \) is from a parametric distribution \( f(x|\theta) \), estimate \( \theta \).

• E.g.: Given sample HHTTTTTHTHTTTTHH of (possibly biased) coin flips, estimate

\[ \theta = \text{probability of Heads} \]

\( f(x|\theta) \) is the Bernoulli probability mass function with parameter \( \theta \)
Likelihood

• \( P(x \mid \theta) \): Probability of event \( x \) given model \( \theta \)
• Viewed as a function of \( x \) (fixed \( \theta \)), it’s a *probability*
  • E.g., \( \sum_x P(x \mid \theta) = 1 \)
• Viewed as a function of \( \theta \) (fixed \( x \)), it’s a *likelihood*
  • E.g., \( \sum_{\theta} P(x \mid \theta) \) can be anything; *relative* values of interest.
    E.g., if \( \theta = \) prob of heads in a sequence of coin flips then
    \[ P(\text{HHTHH} \mid .6) > P(\text{HHTHH} \mid .5), \]
    i.e., event HHTHH is *more likely* when \( \theta = .6 \) than \( \theta = .5 \)
  • And what \( \theta \) make HHTHH *most likely*?
Likelihood Function

\[ P(\text{HHTHH} | \theta) \]: Probability of HHTHH, given \( P(H) = \theta \):

<table>
<thead>
<tr>
<th>( \theta )</th>
<th>( \theta^4(1-\theta) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.0013</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0313</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0819</td>
</tr>
<tr>
<td>0.95</td>
<td>0.0407</td>
</tr>
</tbody>
</table>
Maximum Likelihood Parameter Estimation

• One (of many) approaches to param. est.

• Likelihood of (indp) observations \( x_1, x_2, \ldots, x_n \)

\[
L(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)
\]

• As a function of \( \theta \), what \( \theta \) maximizes the likelihood of the data actually observed

• Typical approach:

\[
\frac{\partial}{\partial \theta} L(\bar{x} \mid \theta) = 0 \quad \text{or} \quad \frac{\partial}{\partial \theta} \log L(\bar{x} \mid \theta) = 0
\]
Example 1

• $n$ coin flips, $x_1, x_2, \ldots, x_n$; $n_0$ tails, $n_1$ heads, $n_0 + n_1 = n$; 

$\theta = \text{probability of heads}$

$$L(x_1, x_2, \ldots, x_n \mid \theta) = (1 - \theta)^{n_0} \theta^{n_1}$$

$$\log L(x_1, x_2, \ldots, x_n \mid \theta) = n_0 \log(1 - \theta) + n_1 \log \theta$$

$$\frac{\partial}{\partial \theta} \log L(x_1, x_2, \ldots, x_n \mid \theta) = \frac{-n_0}{1-\theta} + \frac{n_1}{\theta}$$

Setting to zero and solving:

$$\hat{\theta} = \frac{n_1}{n}$$

(Also verify it’s max, not min, & not better on boundary)
Parameter Estimation

• Assuming sample $x_1, x_2, ..., x_n$ is from a parametric distribution $f(x|\theta)$, estimate $\theta$.

• E.g.: Given $n$ normal samples, estimate mean & variance

\[
f(x) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-(x-\mu)^2/(2\sigma^2)}
\]

\[
\theta = (\mu, \sigma^2)
\]
Ex2: I got data; a little birdie tells me it’s normal, and promises $\sigma^2 = 1$
Which is more likely: (a) this?

\[ \mu \text{ unknown, } \sigma^2 = 1 \]
Which is more likely: (b) or this?

$\mu$ unknown, $\sigma^2 = 1$
Which is more likely: (c) or this?

$\mu$ unknown, $\sigma^2 = 1$

Observed Data
Which is more likely: (c) or this?

$\mu$ unknown, $\sigma^2 = 1$

Looks good by eye, but how do I optimize my estimate of $\mu$?
Ex. 2: \( x_i \sim N(\mu, \sigma^2), \sigma^2 = 1, \mu \text{ unknown} \)

\[
L(x_1, x_2, \ldots, x_n | \theta) = \prod_{1 \leq i \leq n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - \theta)^2}{2}}
\]

\[
\ln L(x_1, x_2, \ldots, x_n | \theta) = \sum_{1 \leq i \leq n} -\frac{1}{2} \ln 2\pi - \frac{(x_i - \theta)^2}{2}
\]

\[
\frac{d}{d\theta} \ln L(x_1, x_2, \ldots, x_n | \theta) = \sum_{1 \leq i \leq n} (x_i - \theta)
\]

And verify it’s max, not min & not better on boundary

\[
= \left( \sum_{1 \leq i \leq n} x_i \right) - n\theta = 0
\]

\[
\hat{\theta} = \left( \sum_{1 \leq i \leq n} x_i \right) / n = \bar{x}
\]

Sample mean is MLE of population mean
Hmm …, density ≠ probability

• So why is “likelihood” function equal to product of densities??
  
  • a) for maximizing likelihood, we really only care about relative likelihoods, and density captures that

• and/or

  • b) if density at $x$ is $f(x)$, for any small $\delta > 0$, the probability of a sample within $\pm \delta/2$ of $x$ is $\approx \delta f(x)$, but $\delta$ is constant wrt $\theta$, so it just drops out of

    \[
    \frac{d}{d\theta} \log L(\ldots) = 0.
    \]
Ex3: I got data; a little birdie tells me it’s normal (but does not tell me $\sigma^2$)
Which is more likely: (a) this?

$\mu, \sigma^2$ both unknown
Which is more likely: (b) or this?

$\mu, \sigma^2$ both unknown
Which is more likely: (c) or this?

$\mu, \sigma^2$ both unknown
Which is more likely: (d) or this?

\[ \mu, \sigma^2 \text{ both unknown} \]
Which is more likely: (d) or this?

μ, σ² both unknown

Looks good by eye, but how do I optimize my estimates of μ & σ²?

Observed Data

μ ± 0.5
Ex 3: \[ x_i \sim N(\mu, \sigma^2) \], \( \mu, \sigma^2 \) both unknown

\[
\ln L(x_1, x_2, \ldots, x_n | \theta_1, \theta_2) = \sum_{1 \leq i \leq n} -\frac{1}{2} \ln 2\pi \theta_2 - \frac{(x_i - \theta_1)^2}{2\theta_2}
\]

\[
\frac{\partial}{\partial \theta_1} \ln L(x_1, x_2, \ldots, x_n | \theta_1, \theta_2) = \sum_{1 \leq i \leq n} \frac{(x_i - \theta_1)}{\theta_2} = 0
\]

\[
\hat{\theta}_1 = \left( \sum_{1 \leq i \leq n} x_i \right) / n = \bar{x}
\]

Sample mean is MLE of population mean, again

In general, a problem like this results in 2 equations in 2 unknowns. Easy in this case, since \( \theta_2 \) drops out of the \( \partial / \partial \theta_1 = 0 \) equation.
Ex. 3, (cont.)

\[ \ln L(x_1, x_2, \ldots, x_n | \theta_1, \theta_2) = \sum_{1 \leq i \leq n} \left( -\frac{1}{2} \ln 2\pi \theta_2 - \frac{(x_i - \theta_1)^2}{2\theta_2} \right) \]

\[ \frac{\partial}{\partial \theta_2} \ln L(x_1, x_2, \ldots, x_n | \theta_1, \theta_2) = \sum_{1 \leq i \leq n} \left( -\frac{1}{2} \frac{2\pi}{2\pi \theta_2} + \frac{(x_i - \theta_1)^2}{2\theta_2^2} \right) = 0 \]

\[ \hat{\theta}_2 = \left( \sum_{1 \leq i \leq n} (x_i - \hat{\theta}_1)^2 \right) / n = \hat{s}^2 \]

*Sample variance is MLE of population variance*
Summary

• MLE is *one* way to estimate *parameters* from *data*

• You choose the *form* of the model (normal, binomial, ...)

• Math chooses the *value(s)* of parameter(s)

• Has the intuitively appealing property that the parameters maximize the *likelihood* of the observed data; basically just assumes your sample is “representative”