FYI: Hemo-globin History
Outline

MLE: Maximum Likelihood Estimators

EM: the Expectation Maximization Algorithm

Next: Motif description & discovery

Learning From Data: MLE

Maximum Likelihood Estimators

Probability Basics, I

Sample Space

\{1, 2, \ldots, 6\}

Distribution

\begin{align*}
p_1, \ldots, p_6 & \geq 0; \quad \sum_{1 \leq i \leq 6} p_i = 1 \\
f(x) & > 0; \quad \int_{\mathbb{R}} f(x) \, dx = 1
\end{align*}

e.g.

\begin{align*}
p_1 = \cdots = p_6 = 1/6 \\
f(x) & = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\end{align*}

Probability Basics, II

Expectation

\begin{align*}
E(g) & = \sum_{1 \leq i \leq 6} g(i)p_i \\
E(g) & = \int_{\mathbb{R}} g(x)f(x) \, dx
\end{align*}

Population

\begin{align*}
\mu & = \sum_{1 \leq i \leq 6} ip_i \\
\mu & = \int_{\mathbb{R}} x f(x) \, dx
\end{align*}

\begin{align*}
\sigma^2 & = \sum_{1 \leq i \leq 6} (i-\mu)^2p_i \\
\sigma^2 & = \int_{\mathbb{R}} (x-\mu)^2 f(x) \, dx
\end{align*}

Sample

\begin{align*}
\bar{x} & = \sum_{1 \leq i \leq n} x_i/n \\
\bar{x}^2 & = \sum_{1 \leq i \leq n} (x_i - \bar{x})^2/n
\end{align*}
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 HHTTTTHTHTTTTHH of (possibly biased) coin flips, estimate \( \theta = \) probability of Heads

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., \( \Sigma x P(x \mid \theta) = 1 \)

Viewed as a function of \( \theta \) (fixed \( x \)), it’s a likelihood

E.g., if \( \theta = \) prob of heads in a sequence of coin flips then

\( P(HHTTHH \mid .6) > P(HHTTHH \mid .5) \)

I.e., event HHTTHH is more likely when \( \theta = .6 \) than \( \theta = .5 \)

And what \( \theta \) make HHTTHH most likely?

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: \( \frac{\partial}{\partial \theta} L(\vec{x} \mid \theta) = 0 \) or \( \frac{\partial}{\partial \theta} \log L(\vec{x} \mid \theta) = 0 \)
Example 1

$n$ coin flips, $x_1, x_2, ..., x_n$; $n_0$ tails, $n_1$ heads, $n_0 + n_1 = n$;
$\theta =$ probability of heads

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

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

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

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

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

$$\theta = (\mu, \sigma^2)$$

Ex. 2: $x_i \sim N(\mu, \sigma^2)$, $\sigma^2 = 1$, $\mu$ unknown

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

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

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

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

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

Ex 3: $x_i \sim N(\mu, \sigma^2)$, $\mu, \sigma^2$ both unknown

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

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

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

Sample mean is MLE of population mean, again
Ex. 3, (cont.)

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

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

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

A consistent, but biased estimate of population variance. (An example of overfitting.) Unbiased estimate is:

\[
\hat{\theta}_2^\prime = \sum_{1 \leq i \leq n} \frac{(x_i - \hat{\theta}_1)^2}{n - 1}
\]

Moral: MLE is a great idea, but not a magic bullet

Aside: Is it Biased? Why?

Is it? Yes. As an extreme, when \( n = 1 \), \( \hat{\theta}_2 = 0 \).

Why? A bit harder to see, but think about \( n = 2 \). Then \( \hat{\theta}_1 \) is exactly between the two sample points, the position that exactly minimizes the expression for \( \hat{\theta}_2 \). Any other choices for \( \theta_1, \theta_2 \) make the likelihood of the observed data slightly lower. But it’s actually pretty unlikely that two sample points would be chosen exactly equidistant from, and on opposite sides of the mean, so the MLE \( \hat{\theta}_2 \) systematically underestimates \( \theta_2 \).

More Complex Example

EM

The Expectation-Maximization Algorithm

This?

Or this?

(A modeling decision, not a math problem..., but if later, what math?)
A Real Example:
CpG content of human gene promoters

"A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters" Saxonov, Berg, and Brutlag, PNAS 2006;103:1412-1417

©2006 by National Academy of Sciences

Gaussian Mixture Models / Model-based Clustering

Parameters $\theta$
- means $\mu_1, \mu_2$
- variances $\sigma_1^2, \sigma_2^2$
- mixing parameters $\tau_1, \tau_2 = 1 - \tau_1$

P.D.F.
$f(x|\mu_1, \sigma_1^2) f(x|\mu_2, \sigma_2^2)$

Likelihood
$L(x_1, x_2, \ldots, x_n|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2) = \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i|\mu_j, \sigma_j^2)$

No closed-form max

Likelihood Surface

$x_i = -10.2, -10, -9.8, -0.2, 0, 0.2, 11.8, 12, 12.2$

$\sigma_1^2 = 1.0$
$\tau_1 = 0.5$
$\tau_2 = 0.5$
EM as Egg vs Chicken

IF $z_{ij}$ known, could estimate parameters $\theta$
  E.g., only points in cluster 2 influence $\mu_2$, $\sigma_2$

IF parameters $\theta$ known, could estimate $z_{ij}$
  E.g., if $|x_i - \mu_1|/\sigma_1 << |x_i - \mu_2|/\sigma_2$, then $z_{i1} >> z_{i2}$

But we know neither; (optimistically) iterate:
  E: calculate expected $z_{ij}$, given parameters
  M: calc “MLE” of parameters, given E($z_{ij}$)

Overall, a clever “hill-climbing” strategy

A What-If Puzzle

Likelihood
$$L(x_1, x_2, \ldots, x_n | \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2)$$
$$= \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i | \mu_j, \sigma_j^2)$$

Messy: no closed form solution known for finding $\theta$ maximizing $L$

But what if we knew the hidden data?

$$z_{ij} = \begin{cases} 
1 & \text{if } x_i \text{ drawn from } f_j \\
0 & \text{otherwise}
\end{cases}$$

Simple Version: “Classification EM”

If $z_{ij} < .5$, pretend it’s 0; $z_{ij} > .5$, pretend it’s 1
I.e., classify points as component 0 or 1
Now recalc $\theta$, assuming that partition
Then recalc $z_{ij}$, assuming that $\theta$
Then re-recalc $\theta$, assuming new $z_{ij}$, etc., etc.
“Full EM” is a bit more involved, but this is the crux.
Full EM

$x_i$’s are known; $\theta$ unknown. Goal is to find MLE $\theta$ of:

$$L(x_1, \ldots, x_n \mid \theta)$$

(hideen data likelihood)

Would be easy if $z_{ij}$’s were known, i.e., consider:

$$L(x_1, \ldots, x_n, z_{11}, z_{12}, \ldots, z_{n2} \mid \theta)$$

(complete data likelihood)

But $z_{ij}$’s aren’t known.

Instead, maximize expected likelihood of visible data

$$E(L(x_1, \ldots, x_n, z_{11}, z_{12}, \ldots, z_{n2} \mid \theta)),$$

where expectation is over distribution of hidden data ($z_{ij}$’s)

The E-step:

Find $E(Z_{ij})$, i.e. $P(Z_{ij}=1)$

Assume $\theta$ known & fixed

$A$ ($B$): the event that $x_i$ was drawn from $f_1$ ($f_2$)

$D$: the observed datum $x_i$

Expected value of $z_{i1}$ is $P(A|D)$

$$E = 0 \cdot P(0) + 1 \cdot P(1)$$

$$P(A) = \frac{P(D|A)P(A)}{P(D)}$$

$$P(D) = P(D|A)P(A) + P(D|B)P(B)$$

$$= f_1(x_i|\theta_1)\tau_1 + f_2(x_i|\theta_2)\tau_2$$

Repeat for each $x_i$

Complete Data Likelihood

Recall:

$$z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ drawn from } f_j \\ 0 & \text{otherwise} \end{cases}$$

so, correspondingly.

$$L(x_1, z_{ij} \mid \theta) = \begin{cases} \tau_1 f_1(x_1 \mid \theta) & \text{if } z_{11} = 1 \\ \tau_2 f_2(x_1 \mid \theta) & \text{otherwise} \end{cases}$$

Formulas with “if’s” are messy; can we blend more smoothly?

Yes, many possibilities. Idea 1:

$$L(x_1, z_{ij} \mid \theta) = z_{11} \cdot \tau_1 f_1(x_1 \mid \theta) + z_{12} \cdot \tau_2 f_2(x_1 \mid \theta)$$

Idea 2 (Better):

$$L(x_1, z_{ij} \mid \theta) = (\tau_1 f_1(x_1 \mid \theta))^{z_{11}} \cdot (\tau_2 f_2(x_1 \mid \theta))^{z_{12}}$$

M-step:

Find $\theta$ maximizing $E(\log(\text{Likelihood}))$

(For simplicity, assume $\sigma_1 = \sigma_2 = \sigma$; $\tau_1 = \tau_2 = .5 = \tau$)

$$L(\vec{z}, \vec{x} \mid \theta) = \prod_{1 \leq i \leq n} \left( \frac{\tau}{\sqrt{2\pi\sigma^2}} \exp \left( -\sum_{1 \leq j \leq 2} z_{ij} \frac{(x_i - \mu_j)^2}{2\sigma^2} \right) \right)$$

$$E[\log L(\vec{z}, \vec{x} \mid \theta)] = E \left[ \sum_{1 \leq i \leq n} \left( \log \tau - \frac{1}{2} \log 2\pi\sigma^2 - \sum_{1 \leq j \leq 2} z_{ij} \frac{(x_i - \mu_j)^2}{2\sigma^2} \right) \right]$$

$$= \sum_{1 \leq i \leq n} \left( \log \tau - \frac{1}{2} \log 2\pi\sigma^2 - \sum_{1 \leq j \leq 2} \frac{E[z_{ij}](x_i - \mu_j)^2}{2\sigma^2} \right)$$

Find $\theta$ maximizing this as before, using $E[z_{ij}]$ found in E-step. Result:

$$\mu_j = \frac{\sum_{i=1}^{n} E[z_{ij}]x_i}{\sum_{i=1}^{n} E[z_{ij}]}$$

(intuit: avg, weighted by subpop prob)
2 Component Mixture

\[ \sigma_1 = \sigma_2 = 1; \ \tau = 0.5 \]

<table>
<thead>
<tr>
<th></th>
<th>mu1</th>
<th>-20.00</th>
<th>-6.00</th>
<th>-5.00</th>
<th>-4.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>mu2</td>
<td>6.00</td>
<td>3.75</td>
<td>3.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x1</td>
<td>-6</td>
<td>z11</td>
<td>5.11E-12</td>
<td>1.00E+00</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>x2</td>
<td>-5</td>
<td>z21</td>
<td>2.61E-23</td>
<td>1.00E+00</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>x3</td>
<td>-4</td>
<td>z31</td>
<td>1.33E-34</td>
<td>9.98E-01</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>x4</td>
<td>0</td>
<td>z41</td>
<td>9.09E-80</td>
<td>1.52E-08</td>
<td>4.11E-03</td>
</tr>
<tr>
<td>x5</td>
<td>4</td>
<td>z51</td>
<td>6.19E-125</td>
<td>5.75E-19</td>
<td>2.64E-18</td>
</tr>
<tr>
<td>x6</td>
<td>5</td>
<td>z61</td>
<td>3.16E-136</td>
<td>1.43E-21</td>
<td>4.20E-22</td>
</tr>
<tr>
<td>x7</td>
<td>6</td>
<td>z71</td>
<td>1.62E-147</td>
<td>3.53E-24</td>
<td>6.69E-26</td>
</tr>
</tbody>
</table>

EM Summary

Fundamentally a max likelihood parameter estimation problem

Useful if analysis is more tractable when 0/1 hidden data z known

Iterate:
- E-step: estimate E(z) for each z, given \( \theta \)
- M-step: estimate \( \theta \) maximizing E(log likelihood)
given E(z) [where “E(logL)” is wrt random z ~ E(z) = p(z=1)]

EM Issues

Under mild assumptions (sect 11.6), EM is guaranteed to increase likelihood with every E-M iteration, hence will converge.

But may converge to local, not global, max.

(Recall the 4-bump surface...)

Issue is intrinsic (probably), since EM is often applied to NP-hard problems (including clustering, above, and motif-discovery, soon)

Nevertheless, widely used, often effective