EM

The Expectation-Maximization Algorithm
Last lecture:
How to estimate $\mu$ given data

For this problem, we got a nice, closed form, solution, allowing calculation of the $\mu, \sigma$ that maximize the likelihood of the observed data.

We’re not always so lucky...
More Complex Example

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
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.

$$f(x|\mu_1, \sigma_1^2) \quad f(x|\mu_2, \sigma_2^2)$$

Likelihood

$$L(x_1, x_2, \ldots, x_n|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2) = \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i|\mu_j, \sigma_j^2)$$
$x_i = -10.2, -10, -9.8, -0.2, 0, 0.2, 11.8, 12, 12.2,$

$\mu_1, \mu_2$

$\sigma^2 = 1.0$

$\tau_1 = 0.5$

$\tau_2 = \frac{5}{33}$
\[ x_i = -10.2, -10, -9.8, -0.2, 0, 0.2, 11.8, 12, 12.2 \]

\[ \mu_1 \]

\[ \mu_2 \]

\[ \sigma^2 = 1.0 \]

\[ \tau_1 = 0.5 \]

\[ \tau_2 = \frac{5}{34} \]
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?

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

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

(hidden 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 $P(A|D)$

\[
E = 0 \cdot P(0) + 1 \cdot P(1)
\]

\[
P(A|D) = \frac{P(D|A)P(A)}{P(D)}
\]

\[
P(D) = P(D|A)P(A) + P(D|B)P(B)
\]

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

Repeat for each $x_i$
Complete Data Likelihood

Recall:

\[ z_{1j} = \begin{cases} 
1 & \text{if } x_1 \text{ drawn from } f_j \\
0 & \text{otherwise} 
\end{cases} \]

so, correspondingly,

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

Formulas with “if’s” are messy; can we blend more smoothly? Yes, many possibilities. Idea 1:

\[ L(x_1, z_{1j} \mid \theta) = z_{11} \cdot \tau_1 f_1(x_1 \mid \theta) + z_{12} \cdot \tau_2 f_2(x_1 \mid \theta) \]

Idea 2 (Better):

\[ L(x_1, z_{1j} \mid \theta) = (\tau_1 f_1(x_1 \mid \theta))^{z_{11}} \cdot (\tau_2 f_2(x_1 \mid \theta))^{z_{12}} \]
M-step:

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

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

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

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

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

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

$$\mu_j = \frac{\sum_{i=1}^n E[z_{ij}] x_i}{\sum_{i=1}^n E[z_{ij}]} \quad \text{(intuit: avg, weighted by subpop prob)}$$
2 Component Mixture

$\sigma_1 = \sigma_2 = 1; \quad \tau = 0.5$

<table>
<thead>
<tr>
<th></th>
<th>mu1</th>
<th>mu2</th>
<th>-6.00</th>
<th>-5.00</th>
<th>-4.99</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-20.00</td>
<td>6.00</td>
<td>0.00</td>
<td>3.75</td>
<td>3.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x1</th>
<th>-6</th>
<th>z11</th>
<th>5.11E-12</th>
<th>1.00E+00</th>
<th>1.00E+00</th>
</tr>
</thead>
<tbody>
<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>

Essentially converged in 2 iterations

(Excel spreadsheet on course web)
Applications

Clustering is a remarkably successful exploratory data analysis tool

- Web-search, information retrieval, gene-expression, ...
- Model-based approach above is one of the leading ways to do it

Gaussian mixture models widely used

- With many components, empirically match arbitrary distribution
- Often well-justified, due to “hidden parameters” driving the visible data

EM is extremely widely used for “hidden-data” problems

Hidden Markov Models
EM Summary

Fundamentally a maximum likelihood parameter estimation problem

Useful if hidden data, and 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 $E(\log \text{likelihood})$
given $E(z)$ [where “$E(\log L)$” is wrt random $z \sim E(z) = p(z=1)$]
EM Issues

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

But it may converge to a local, not global, max. (Recall the 4-bump surface...)

Issue is intrinsic (probably), since EM is often applied to problems (including clustering, above) that are NP-hard (next 3 weeks!)

Nevertheless, widely used, often effective