Learning With Hidden Variables

- Why do we want hidden variables?
- Simple case of missing data
- EM algorithm
- Bayesian networks with hidden variables

Hidden variables

Cause is unobservable

Without the cause, all the evidence is dependent on each other
Missing Data

- Given two variables, no independence relations
- Some data are missing
- Estimate parameters in joint distribution
- Data must be missing at random

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>H</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Ignore it

Estimated Parameters

\[
\log \Pr(D|M) = \log(\Pr(D,H = 0 | M) + \Pr(D,H = 1 | M)) = 3\log .429 + 2\log .143 + 2\log .285 + \log(.429 + .143)
\]
\[
= -9.498
\]
Recitation Problem

Show the remaining steps required to get from this expression

$$\log \Pr(D|M) = \log(\Pr(D, H = 0 \mid M) + \Pr(D, H = 1 \mid M))$$

to a number for the log likelihood of the observed data given the model.

Explain any assumptions you might have had to make.

---

Fill in With Best Value

Estimated Parameters

<table>
<thead>
<tr>
<th></th>
<th>~A</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>~B</td>
<td>4/8</td>
<td>1/8</td>
</tr>
<tr>
<td>B</td>
<td>1/8</td>
<td>2/8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>~A</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>~B</td>
<td>.5</td>
<td>.125</td>
</tr>
<tr>
<td>B</td>
<td>.125</td>
<td>.25</td>
</tr>
</tbody>
</table>

$$\log \Pr(D|M) = \log(\Pr(D, H = 0 \mid M) + \Pr(D, H = 1 \mid M))$$

$$= 3 \log .5 + 2 \log .125 + 2 \log .25 + \log (.5 + .125)$$

$$= -9.481$$
Fill in With Distribution

Guess a distribution over \( A, B \) and compute a distribution over \( H \)

\[
\begin{array}{c|cc}
\theta_0 & \sim A & A \\
\hline
\sim B & 0.25 & 0.25 \\
B & 0.25 & 0.25 \\
\end{array}
\]

\[
\Pr(H|D, \theta_0) = \Pr(H|D^c, \theta_0) \\
= \Pr(B|\sim A, \theta_0) \\
= \Pr(\sim A, B|\theta_0)/\Pr(\sim A|\theta_0) \\
= 0.25/0.5 \\
= 0.5
\]

Fill in With Distribution

Use distribution over \( H \) to compute better distribution over \( A, B \)

Maximum likelihood estimation using expected counts

\[
\begin{array}{c|cc}
\theta_1 & \sim A & A \\
\hline
\sim B & 3.5/8 & 1/8 \\
B & 1.5/8 & 2/8 \\
\end{array}
\]

\[
\begin{array}{c|cc}
\theta_1 & \sim A & A \\
\hline
\sim B & 0.4375 & 0.125 \\
B & 0.1875 & 0.25 \\
\end{array}
\]
Fill in With Distribution

Use new distribution over AB to get a better distribution over H

\[ \theta_1 \]

\[
\begin{array}{c|cc}
 & \sim A & A \\
\hline
\sim B & .4375 & .125 \\
B & .1875 & .25 \\
\end{array}
\]

\[
\Pr(H|D,\theta_1) = \Pr(\sim A, B |\theta_1)/\Pr(\sim A |\theta_1)
\]

\[
= .1875/.625
\]

\[
= 0.3
\]

Lecture 18 • 9

Fill in With Distribution

Use distribution over H to compute better distribution over A,B

\[ \theta_2 \]

\[
\begin{array}{c|cc}
 & \sim A & A \\
\hline
\sim B & 3.7/8 & 1/8 \\
B & 1.3/8 & 2/8 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & \sim A & A \\
\hline
\sim B & .4625 & .125 \\
B & .1625 & .25 \\
\end{array}
\]

Lecture 18 • 10
### Fill in With Distribution

Use new distribution over AB to get a better distribution over H

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

$$\theta_2$$

<table>
<thead>
<tr>
<th></th>
<th>(\sim A)</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.4625</td>
<td>0.125</td>
</tr>
<tr>
<td>0</td>
<td>0.1625</td>
<td>0.25</td>
</tr>
</tbody>
</table>

$$\text{Pr}(H|D, \theta_2) = \frac{\text{Pr}(\sim A, B | \theta_2)}{\text{Pr}(\sim A | \theta_2)}$$

$$= \frac{0.1625}{0.625}$$

$$= 0.26$$

### Fill in With Distribution

Use distribution over H to compute better distribution over A,B

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$$\theta_3$$

<table>
<thead>
<tr>
<th></th>
<th>(\sim A)</th>
<th>A</th>
<th>(\sim A)</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.74/8</td>
<td>1/8</td>
<td>0.4675</td>
<td>0.125</td>
</tr>
<tr>
<td>0</td>
<td>1.26/8</td>
<td>2/8</td>
<td>0.1575</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Increasing Log-Likelihood

<table>
<thead>
<tr>
<th>( \theta_0 )</th>
<th>( \sim A )</th>
<th>( A )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sim B )</td>
<td>.25</td>
<td>.25</td>
</tr>
<tr>
<td>B</td>
<td>.25</td>
<td>.25</td>
</tr>
</tbody>
</table>

\[ \log \Pr(D \mid \theta_0) = -10.3972 \]

\[ \text{ignore: -9.498} \]

<table>
<thead>
<tr>
<th>( \theta_1 )</th>
<th>( \sim A )</th>
<th>( A )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sim B )</td>
<td>.4375</td>
<td>.125</td>
</tr>
<tr>
<td>B</td>
<td>.1875</td>
<td>.25</td>
</tr>
</tbody>
</table>

\[ \log \Pr(D \mid \theta_1) = -9.4760 \]

<table>
<thead>
<tr>
<th>( \theta_2 )</th>
<th>( \sim A )</th>
<th>( A )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sim B )</td>
<td>.4625</td>
<td>.125</td>
</tr>
<tr>
<td>B</td>
<td>.1625</td>
<td>.25</td>
</tr>
</tbody>
</table>

\[ \log \Pr(D \mid \theta_2) = -9.4524 \]

<table>
<thead>
<tr>
<th>( \theta_3 )</th>
<th>( \sim A )</th>
<th>( A )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sim B )</td>
<td>.4675</td>
<td>.125</td>
</tr>
<tr>
<td>B</td>
<td>.1575</td>
<td>.25</td>
</tr>
</tbody>
</table>

\[ \log \Pr(D \mid \theta_3) = -9.4514 \]

| \( \text{best val: -9.481} \) |

Deriving the EM Algorithm

- Want to find \( \theta \) to maximize \( \Pr(D \mid \theta) \)

- Instead, find \( \hat{\theta}, \hat{P} \) to maximize

\[
g(\theta, \hat{P}) = \sum_H \hat{P}(H) \log(\Pr(D,H \mid \theta) / \hat{P}(H))
= E_\hat{P} \log \Pr(D,H \mid \theta) - \log \hat{P}(H)
\]

- Alternate between

  - holding \( \theta \) fixed and optimizing \( \hat{P} \)
  - holding \( \hat{P} \) fixed and optimizing \( \theta \)

- \( g \) has same local and global optima as \( \Pr(D \mid \theta) \)
EM Algorithm

- Pick initial $\theta_0$
- Loop until apparently converged
  - $\tilde{P}_{t+1}(H) = \Pr(H | D, \theta_t)$
  - $\theta_{t+1} = \arg \max_{\theta} \mathbb{E}_{\tilde{P}_{t+1}} \log \Pr(D, H | \theta)$

- Monotonically increasing likelihood
- Convergence is hard to determine due to plateaus
- Problems with local optima

EM for Bayesian Networks

- D: observable variables
- H: values of hidden variables in each case
- Assume structure is known
- Goal: maximum likelihood estimation of CPTs

- Initialize CPTs to anything (with no 0’s)
- Fill in the data set with distribution over values for hidden vars
- Estimate CPTs using expected counts
Filling in the data

- Distribution over H factors over the M data cases
  \[ \tilde{P}_{t+1}(H) = \Pr(H | D, \theta_t) \]
  \[ = \prod_{m} \Pr(H^m | D^m, \theta_t) \]
- We really just need to compute a distribution over each individual hidden variable
- Each factor is a call to Bayes net inference

EM for BN: Simple Case

| D_1 | D_2 | ... | D_n | Pr(H^m | D^m, \theta_t) |
|-----|-----|-----|-----|-----------------------|
| 1   | 1   | 0   | 0   | .9                    |
| 0   | 1   | 0   | 0   | .2                    |
| 0   | 0   | 1   | 0   | .1                    |
| 1   | 0   | 1   | 0   | .6                    |
| 1   | 1   | 1   | 0   | .2                    |
| 1   | 1   | 1   | 0   | .5                    |
| 0   | 1   | 0   | 0   | .3                    |
| 0   | 0   | 0   | 0   | .7                    |
| 1   | 1   | 0   | 0   | .2                    |
EM for BN: Simple Case

![Bayes net inference]

<table>
<thead>
<tr>
<th>$D_1$</th>
<th>$D_2$</th>
<th>$D_n$</th>
<th>$\Pr(H^m \mid D^n, \theta)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>.9</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>.1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>.6</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.5</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>.3</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.7</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>.2</td>
</tr>
</tbody>
</table>

$E\#(H) = \sum_{m} \Pr(H^m \mid D^n, \theta)$

$= 3.7$

$E\#(H \times D_2) = \sum_{m} \Pr(H^m \mid D^n, \theta)I(D_2^m)$

$= .9 + .2 + .5 + .3 + .2$

$= 2.3$

$\Pr(D_2 \mid H) = 2.3/3.7 = .6216$

Re-estimate $\theta$

EM for BN: Worked Example

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>$\Pr(H^m \mid D^n, \theta)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

$\theta_1 = \Pr(H)$

$\theta_2 = \Pr(A \mid H)$

$\theta_3 = \Pr(A \mid \neg H)$

$\theta_4 = \Pr(B \mid H)$

$\theta_5 = \Pr(B \mid \neg H)$
EM for BN: Initial Model

| A | B | # | Pr(H'' | D'',θ) |
|---|---|---|----------------|
| 0 | 0 | 6 |                |
| 0 | 1 | 1 |                |
| 1 | 0 | 1 |                |
| 1 | 1 | 4 |                |

Pr(H) = 0.4
Pr(A|H) = 0.55
Pr(A¬H) = 0.61
Pr(B|H) = 0.43
Pr(B¬H) = 0.52

Iteration 1: Fill in data

| A | B | # | Pr(H'' | D'',θ) |
|---|---|---|----------------|
| 0 | 0 | 6 | .48            |
| 0 | 1 | 1 | .39            |
| 1 | 0 | 1 | .42            |
| 1 | 1 | 4 | .33            |

Pr(H) = 0.4
Pr(A|H) = 0.55
Pr(A¬H) = 0.61
Pr(B|H) = 0.43
Pr(B¬H) = 0.52
Iteration 1: Re-estimate Params

| 0 0 6 | .48 |
| 0 1 1 | .39 |
| 1 0 1 | .42 |
| 1 1 4 | .33 |

Pr(H) = 0.42
Pr(A|H) = 0.35
Pr(A¬H) = 0.46
Pr(B|H) = 0.34
Pr(B¬H) = 0.47

Iteration 2: Fill in Data

| 0 0 6 | .52 |
| 0 1 1 | .39 |
| 1 0 1 | .39 |
| 1 1 4 | .28 |

Pr(H) = 0.42
Pr(A|H) = 0.35
Pr(A¬H) = 0.46
Pr(B|H) = 0.34
Pr(B¬H) = 0.47
Iteration 2: Re-estimate params

| A | B | # | Pr(H | D,θ) |
|---|---|---|---------|
| 0 | 0 | 6 | 0.52    |
| 0 | 1 | 1 | 0.39    |
| 1 | 0 | 1 | 0.28    |
| 1 | 1 | 4 | 0.28    |

Pr(H) = 0.42
Pr(A|H) = 0.31
Pr(A¬H) = 0.50
Pr(B|H) = 0.30
Pr(B¬H) = 0.50

Iteration 5

| A | B | # | Pr(H | D,θ) |
|---|---|---|---------|
| 0 | 0 | 6 | 0.79    |
| 0 | 1 | 1 | 0.31    |
| 1 | 0 | 1 | 0.31    |
| 1 | 1 | 4 | 0.05    |

Pr(H) = 0.46
Pr(A|H) = 0.09
Pr(A¬H) = 0.69
Pr(B|H) = 0.09
Pr(B¬H) = 0.69
### Iteration 10

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>#</th>
<th>Pr($H^n \mid D^n, \theta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>6</td>
<td>.971</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>.183</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>.183</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td>.001</td>
</tr>
</tbody>
</table>

- $\Pr(H) = 0.52$
- $\Pr(A \mid H) = 0.03$
- $\Pr(A \mid \neg H) = 0.83$
- $\Pr(B \mid H) = 0.03$
- $\Pr(B \mid \neg H) = 0.83$

---

### Increasing Log Likelihood

- Near .5
- All params .5
- Near 0
EM in BN issues

- With multiple hidden nodes, take advantage of conditional independencies
- Lots of tricks to speed up computation of expected counts
- If structure is unknown, add search operators to add and delete hidden nodes
- There are clever ways of search with unknown structure and hidden nodes