Learning CPTs from incomplete data

Examples $t = 1, 2, \ldots, T$

Hidden nodes $H^{(t)}$

Visible nodes $V^{(t)}$

EM algorithm

E-step: compute $P(X_i = x, \pi_{a_i} = \pi | V^{(t)})$

M-step: update CPTs

Update rule:

$$P(X_i = x | \pi_{a_i} = \pi) \leftarrow \frac{\sum_t P(X_i = x, \pi_{a_i} = \pi | V^{(t)})}{\sum_t P(\pi_{a_i} = \pi | V^{(t)})}$$

Iterate until convergence

Monotonic improvement in log-likelihood

$$L = \sum_x \log P(V^{(t)})$$

Example

$$\begin{align*}
\text{A} & \rightarrow \text{B} \\
\text{B} & \rightarrow \text{C}
\end{align*}$$

incomplete data set $\{ (a_t, c_t) \}_{t=1}^T$

E-step: compute $P(B = b | A = a_t, C = c_t)$ for all examples $t = 1, \ldots, T$

M-step: $P(B = b | A = a) \leftarrow \frac{\sum_t P(B = b, A = a_t, C = c_t)}{\sum_t P(A = a | A = a_t, C = c_t)}$
Simplify: \( P(b | a) \leftarrow \frac{\sum_t I(a, a_t) P(b | a_t, c_t)}{\sum_t I(a, a_t)} \)

\[ P(C = c | B = b) \leftarrow \frac{\sum_t P(B = b, C = c | A = a_t, C = c_t)}{\sum_t P(B = b | A = a_t, C = c_t)} \]

Simplify: \( P(c | b) \leftarrow \frac{\sum_t I(c, c_t) P(b | a_t, c_t)}{\sum_t P(b | a_t, c_t)} \)

**Noisy-OR**

\[ \text{Symptom } y \in \{0, 1\} \]
\[ \text{Diseases } x_i \in \{0, 1\} \]
\[ P(Y = 1 | x_1, \ldots, x_n) = 1 - \prod_{i=1}^n (1 - p_i)^{x_i} \text{ with } p_i \in [0, 1] \]

- From complete data \( \{(X_t, Y_t)\}_{t=1}^T \), how to estimate \( p_i \in [0, 1] \)?

**Note:** noisy-OR is a "parametric" model of CPT. No simple, closed-form ML estimate of \( p_i \).

- Alternative formulation

\[ P(Y = 1 | z_1, \ldots, z_n) = \text{OR}(z_1, z_2, \ldots, z_n) \]
\[ P(z_i = 1 | X = i) = p_i \]
\[ P(z_i = 1 | X = 0) = 0 \]

Equivalently:
\[ P(z_i = 0 | x_i) = (1 - p_i)^{x_i} = \begin{cases} 1 - p_i & \text{if } x_i = 1 \\ 1 & \text{if } x_i = 0 \end{cases} \]
What is $P(Y=1 \mid \tilde{x})$ in this model?

$$
P(Y=1 \mid \tilde{x}) = \sum_{\tilde{z} \in \{0,1\}^n} P(Y=1 \mid \tilde{z}, \tilde{x}) P(\tilde{z} \mid \tilde{x})$$

marginalization

$$
= \sum_{\tilde{z}} P(Y=1 \mid \tilde{z}, \tilde{x}) P(\tilde{z} \mid \tilde{x})
= \sum_{\tilde{z} \neq 0} P(\tilde{z} \mid \tilde{x})
= 1 - P(\tilde{z}=0 \mid \tilde{x})
= 1 - \prod_{i=1}^{n} P(z_i=0 \mid x_i)
= 1 - \prod_{i=1}^{n} (1-p_i) x_i
$$

Same as original noisy-OR BN!

Posterior probability

$$
P(z_i=1 \mid \tilde{x}, y) = \frac{P(y=0, z_i=1 \mid \tilde{x})}{P(y=0 \mid \tilde{x})} \cdot \frac{P(y=1, z_i=1 \mid \tilde{x})}{P(y=1 \mid \tilde{x})}
= \begin{cases} 
0 & \text{if } y=0 \\
\frac{p_i x_i}{1 - \prod_{i=1}^{n} (1-p_i) x_i} & \text{if } y=1
\end{cases}
= \frac{Y \pi_i x_i}{1 - \prod_{i=1}^{n} (1-p_i) x_i}
$$

(Conditional)

Log-likelihood of (in)complete data $\{(x_t, y_t)\}_{t=1}^{T}$

$$
\mathcal{L} = \sum_{t=1}^{T} \log P(y_t \mid \tilde{x}_t)
= \sum_{t=1}^{T} \left[ (1-y_t) \log P(y=0 \mid \tilde{x}_t) + y_t \log P(y=1 \mid \tilde{x}_t) \right]
= \sum_{t=1}^{T} \left[ (1-y_t) \sum_{i=1}^{n} x_i \log (1-p_i) + y_t \log \left[ 1 - \prod_{i=1}^{n} (1-p_i) x_i \right] \right]
$$
Note: complicated expression to optimize with respect to $\hat{p}_i$!

EM to the rescue!

Shorthand: let $T_i = \sum_{t=1}^{T} X_{it}$ count # times that $x_i=1$ (or $i^{th}$ disease is present)

EM update rule

$$p_i = P(z_i=1| x_i=1) \leftarrow \frac{\sum_{t} P(z_i=1, x_i=1 | x = \hat{x}_t, y = y_t)}{\sum_{t} P(x_i=1 | x = \hat{x}_t, y = y_t)}$$

Simplify

$$p_i \leftarrow \frac{\sum_{t} I(x_{it}, 1) P(z_i=1 | \hat{x}_t, y_t)}{\sum_{t} I(x_{it}, 1)}$$

$$p_i \leftarrow \frac{1}{T_i} \sum_{t} I(x_{it}, 1) \left[ \frac{y_t p_i x_{it}}{1 - \prod_{t=1}^{T} (1-p_i) x_{it}} \right]$$

$$p_i \leftarrow \frac{1}{T_i} \sum_{t} \frac{y_t p_i x_{it}}{1 - \prod_{t=1}^{T} (1-p_i) x_{it}}$$

This update, applied in parallel to $\{p_i\}_{i=1}^{N}$, will monotonically increase $L = \sum_{t} \log P(y_t | \hat{x}_t)$. 