CSE 150 - 2/3/10

* Bigram models of language

$$w_1 \rightarrow w_2 \rightarrow \cdots \rightarrow w_L$$

$w_\ell \in \{1, 2, \ldots, V\}$ $V = \#$ words in vocabulary

$$P_{ML}(w_{\ell+1}=j \mid w_\ell = i) = \frac{C_{ij}}{C_i} = \frac{\text{count}(i \to j)}{\text{count}(i)}$$

* ML Estimation from incomplete data

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

Hidden Nodes $H^{(t)}$

Visible Nodes $V^{(t)}$

Choos CPTs to maximize log-likelihood

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

How?

* EM Algorithm

Iterative porocedure to maximize $\sum_t \log P(V^{(t)})$ in terms of CPTs.

E-Step: compute pøsterior probabilities

$$P(X_i = x, pa_i = \pi \mid V = v^{(t)})$$ (run inference algorithm)

M-Step: update CPTs

$$P(X_i = x \mid pa_i = \pi) \leftarrow \frac{\sum_t P(X_i = x, pa_i = \pi \mid V = v^{(t)})}{\sum_t P(pa_i = \pi \mid V = v^{(t)})}$$

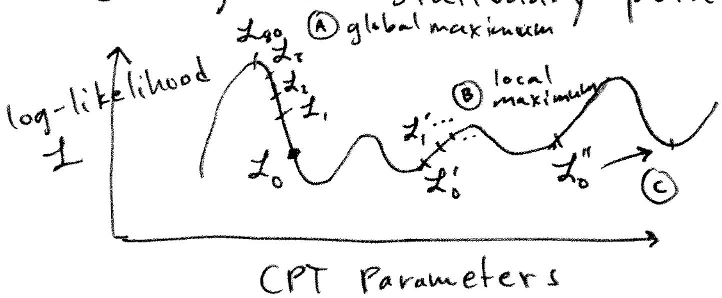
Intuition: expected statistics under $P(H|V)$ are filling in "missing values"

Iterate E & M steps until convergence.
Why iterate? RHS depends on current CPTs.

* Key Properties

- monotonic convergence
  Each iteration of EM improves the log-likelihood
  $$\mathcal{L} = \sum_t \log P(V^{(t)})$$
  If $\mathcal{L}_k$ is log-likelihood at k^{th} iteration, then $\mathcal{L}_k \geq \mathcal{L}_{k-1}$

- Converges to stationary point



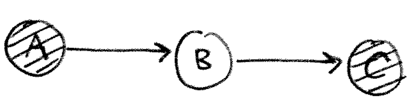
Ⓐ global maximum: most desirable, but not guaranteed

Ⓑ local maximum: usual outcome

Ⓒ local minimum: possible in theory, but never occurs in practice.

- No tuning parameters: no step sizes, learning rates, back-tracking, ...

Example



A and C are observed (visible nodes).
B is hidden.

* Posterior probability

$$P(B=b \mid A=a, C=c) = \frac{P(C=c \mid B=b, A=a)\, P(B=b \mid A=a)}{\sum_{b'} P(C=c \mid B=b', A=a)\, P(B=b' \mid A=a)} \quad \text{Bayes rule}$$

$$= \frac{P(C=c \mid B=b)\, P(B=b \mid A=a)}{\sum_{b'} P(C=c \mid B=b')\, P(B=b' \mid A=a)} \quad \text{conditional independence}$$

Shorthand: $P(b \mid a, c) \Leftrightarrow P(B=b \mid A=a, C=c)$

* Incomplete data set $\{(a_t, c_t)\}_{t=1}^{T}$ (I.I.D.)

$$\mathcal{L} = \sum_t \log P(A=a_t, C=c_t)$$

$$= \sum_t \log \sum_b P(A=a_t, B=b, C=c_t) \quad \textit{marginalization}$$

$$= \sum_t \log \sum_b \left[P(a_t)\, P(b|a_t)\, P(c_t|b)\right] \quad \textit{product rule, conditional independence, shorthand}$$

General EM Algorithm:

$$P(X_i=x \mid pa_i=\pi) \leftarrow \frac{\sum_t P(X_i=x, pa_i=\pi \mid V^{(t)})}{\sum_t P(pa_i=\pi \mid V^{(t)})}$$

Now apply to this example:

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

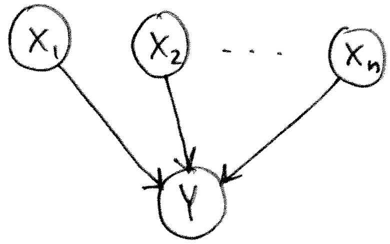
Simplify RHS: $$\frac{\sum_t I(a, a_t)\, P(b \mid a_t, c_t)}{\sum_t I(a, a_t)}$$

$$P(C=c \mid B=b) \leftarrow \frac{\sum_t P(B=b, C=c \mid A=a_t, C=c_t)}{\sum_t P(B=b \mid A=a_t, C=c_t)}$$

simplify:

$$P(C=c \mid B=b) \leftarrow \frac{\sum_t I(c, c_t)\, P(b \mid a_t, c_t)}{\sum_t P(b \mid a_t, c_t)}$$

Noisy-OR Model



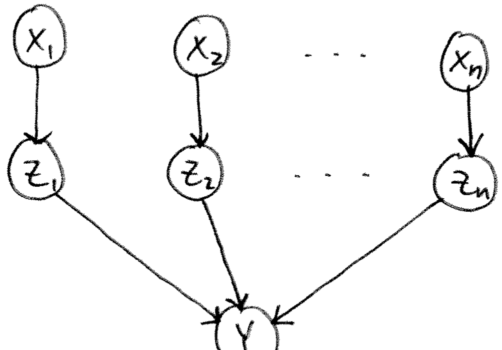
disease $X_i \in \{0,1\}$
symptom $Y \in \{0,1\}$

$$P(Y=1 \mid X_1, X_2, \ldots, X_n) = 1 - \prod_{i=1}^{n} (1-p_i)^{X_i} \quad \text{with } p_i \in [0,1]$$

* From complete data $\{(\vec{x}_t, y_t)\}_{t=1}^{T}$, how do we estimate $p_i \in [0,1]$?

Note: Noisy-OR is a "parametric" model of CPT.
No simply, closed-form ML estimate for $p_i \in [0,1]$.

* Alternative formulation:



$P(Y \mid Z_1, Z_2, \ldots, Z_n) = \mathrm{OR}(Z_1, Z_2, \ldots, Z_n)$
logical-OR (deterministic)

$P(Z_i = 1 \mid X = 1) = p_i$

$P(Z_i = 1 \mid X = 0) = 0$

Equivalently: $P(Z_i = 0 \mid X_i) = (1-p_i)^{X_i} = \begin{cases} 1-p_i, & X=1 \\ 1, & X=0 \end{cases}$

What is $P(Y=1 \mid \vec{X})$ in this new model?

$$P(Y=1 \mid \vec{X}) = \sum_{\vec{z} \in \{0,1\}^n} P(Y=1, \vec{z} \mid \vec{X}) \quad \text{marginalization}$$

$$= \sum_{\vec{z}} P(Y=1 \mid \vec{z}, \vec{x}) P(\vec{z} \mid \vec{X}) \quad \text{product rule}$$

$$= \sum_{\vec{z}} P(Y=1 \mid \vec{z}) \, P(\vec{z} \mid \vec{X}) \quad \text{conditional independence}$$

$$= \sum_{\vec{z} \neq \vec{0}} P(\vec{z} \mid \vec{X}) \quad \text{because } Y = \mathrm{OR}(\vec{z})$$

$$= 1 - P(\vec{z} = \vec{0} \mid \vec{x})$$ from normalization

$$= 1 - \prod_{i=1}^{n} P(z_i = 0 \mid x_i)$$ conditional independence

$$= 1 - \prod_{i=1}^{n} (1 - p_i)^{x_i}$$

$\boxed{\text{Same as original Noisy-OR BN!}}$

* Posterior Probability

$$P(z_i = 1 \mid \vec{x}, y) = \frac{\overbrace{P(y \mid \vec{x}, z_i = 1)}^{\substack{1 \text{ if } y=1 \\ 0 \text{ if } y=0}} \; \overbrace{P(z_i = 1 \mid \vec{x})}^{\substack{p_i \text{ if } x_i=1 \\ 0 \text{ if } x_i=0}}}{\underbrace{P(y \mid \vec{x})}_{1 - \prod_{i=1}^{n} (1-p_i)^{x_i} \text{ if } y=1}}$$

$$= \frac{y \, p_i \, x_i}{1 - \prod_{i=1}^{n} (1 - p_i)^{x_i}}$$