CSE 150A-250A AI: Probabilistic Models

Lecture 9

Fall 2025

Trevor Bonjour
Department of Computer Science and Engineering
University of California, San Diego

Slides adapted from previous versions of the course (Prof. Lawrence, Prof. Alvarado, Prof Berg-Kirkpatrick)

Agenda

Review

Incomplete Data

Expectation-Maximization Algorithm

Review

Learning in BNs with discrete nodes

- ML estimation for complete data:

$$P_{\mathrm{ML}}(X_i\!=\!x|\mathrm{pa}_i\!=\!\pi) \;=\; \frac{\mathrm{count}(X_i\!=\!x, \mathrm{pa}_i\!=\!\pi)}{\sum_{x'} \mathrm{count}(X_i\!=\!x', \mathrm{pa}_i\!=\!\pi)}$$

- For nodes with parents:

$$P_{\mathrm{ML}}(X_i\!=\!x|\mathrm{pa}_i\!=\!\pi) \;=\; \frac{\mathrm{count}(X_i\!=\!x, \mathrm{pa}_i\!=\!\pi)}{\mathrm{count}(\mathrm{pa}_i\!=\!\pi)}$$

- For root nodes:

$$P_{\mathrm{ML}}(X_i\!=\!x) \;=\; \frac{\mathrm{count}(X_i\!=\!x)}{T}$$

- *n*-gram models of word sequences:

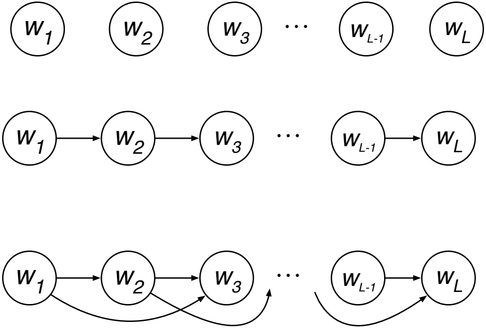
$$P(w_1, w_2, \ldots, w_L) = \prod_{\ell} P(w_\ell | \underbrace{w_{\ell-(n-1)}, \ldots, w_{\ell-1}}_{\text{previous words}}))$$

- As belief networks:

$n = 1$
unigram

$n = 2$
bigram

$n = 3$
trigram

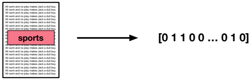


bi

tri

Naive Bayes model for document classification

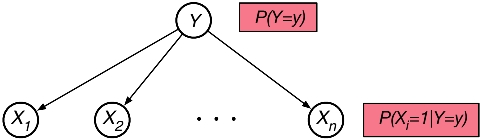
- Random variables



$Y \in \{1, 2, \ldots, m\}$ topic of document

$X_i \in \{0, 1\}$ i^{th} word appears?

- Belief network



- Naive Bayes assumption

$$P(X_1, \ldots, X_n | Y) = \prod_{i=1}^{n} P(X_i | Y)$$

Incomplete Data

Learning from incomplete data with tabular CPTs

Learning from incomplete data with tabular CPTs

$\boxed{\text{ASSUMPTIONS}}$

Learning from incomplete data with tabular CPTs

ASSUMPTIONS

1. The DAG is fixed (and known) over a finite set of discrete random variables $\{X_1, X_2, \ldots, X_n\}$.

Learning from incomplete data with tabular CPTs

$\boxed{\text{ASSUMPTIONS}}$

1. The DAG is fixed (and known) over a finite set of discrete random variables $\{X_1, X_2, \ldots, X_n\}$.

2. CPTs enumerate $P(X_i = x | \text{pa}(X_i) = \pi)$ as lookup tables; each must be estimated for all values of x and π.

Learning from incomplete data with tabular CPTs

$\boxed{\text{ASSUMPTIONS}}$

1. The DAG is fixed (and known) over a finite set of discrete random variables $\{X_1, X_2, \ldots, X_n\}$.
2. CPTs enumerate $P(X_i = x | \text{pa}(X_i) = \pi)$ as lookup tables; each must be estimated for all values of x and π.
3. The data is IID, but only consists of T partially complete instantiations of the nodes in the BN.

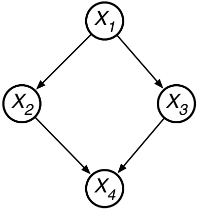
Toy example

Toy example

- Fixed DAG over binary random variables

Toy example

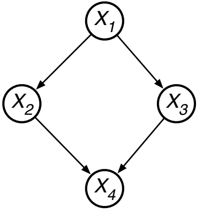
- Fixed DAG over binary random variables



$$
\begin{aligned}
X_1 &\in \{0,1\} \\
X_2 &\in \{0,1\} \\
X_3 &\in \{0,1\} \\
X_4 &\in \{0,1\}
\end{aligned}
$$

Toy example

- Fixed DAG over binary random variables

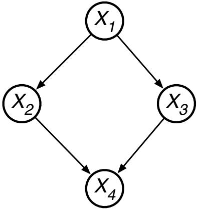


$$\begin{aligned} X_1 &\in \{0,1\} \\ X_2 &\in \{0,1\} \\ X_3 &\in \{0,1\} \\ X_4 &\in \{0,1\} \end{aligned}$$

- Incomplete data set

Toy example

- Fixed DAG over binary random variables



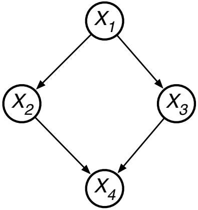
$$
\begin{aligned}
X_1 &\in \{0,1\} \\
X_2 &\in \{0,1\} \\
X_3 &\in \{0,1\} \\
X_4 &\in \{0,1\}
\end{aligned}
$$

- Incomplete data set

example	X_1	X_2	X_3	X_4
1	1	?	0	1
2	0	1	?	0
3	?	?	?	1
⋮	⋮	⋮	⋮	⋮
T	?	1	1	0

Toy example

- Fixed DAG over binary random variables



$$
\begin{aligned}
X_1 &\in \{0,1\} \\
X_2 &\in \{0,1\} \\
X_3 &\in \{0,1\} \\
X_4 &\in \{0,1\}
\end{aligned}
$$

- Incomplete data set

example	X_1	X_2	X_3	X_4
1	1	?	0	1
2	0	1	?	0
3	?	?	?	1
⋮	⋮	⋮	⋮	⋮
T	?	1	1	0

How to choose the CPTs so that the BN maximizes the probability of this data set?

A more interesting example ...

A more interesting example ...

A more interesting example ...

How to build a movie recommendation system?

A more interesting example ...

$\boxed{\text{How to build a movie recommendation system?}}$

- Collect a data set of movie ratings:

A more interesting example ...

How to build a movie recommendation system?

- Collect a data set of movie ratings:

$$\begin{bmatrix} + & - & + & - & ? & ? & + \\ - & ? & ? & + & + & ? & ? \\ + & + & + & + & + & + & + \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ - & - & - & - & - & ? & - \\ ? & ? & + & ? & ? & ? & - \end{bmatrix} \quad \boxed{\begin{array}{ll} + & \text{liked} \\ - & \text{disliked} \\ ? & \text{not seen} \end{array}} \quad \text{(user-item matrix)}$$

A more interesting example ...

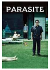
$\boxed{\text{How to build a movie recommendation system?}}$

- Collect a data set of movie ratings:

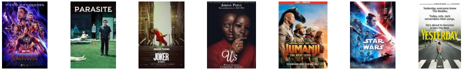
$$\begin{bmatrix} + & - & + & - & ? & ? & + \\ - & ? & ? & + & + & ? & ? \\ + & + & + & + & + & + & + \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ - & - & - & - & - & ? & - \\ ? & ? & + & ? & ? & ? & - \end{bmatrix} \qquad \boxed{\begin{array}{ll} + & \text{liked} \\ - & \text{disliked} \\ ? & \text{not seen} \end{array}} \qquad \textbf{(user-item matrix)}$$

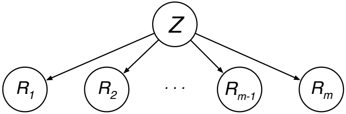
- Build a model of user profiles and fill in the missing ratings.
  But what model to build?

Naive Bayes model with incomplete data

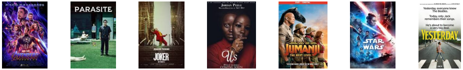
 

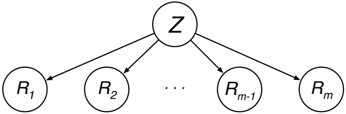
Naive Bayes model with incomplete data





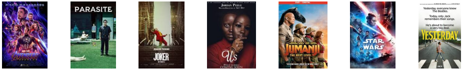
Naive Bayes model with incomplete data

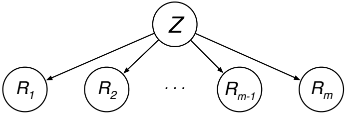




- Movie recommender system

Naive Bayes model with incomplete data

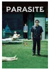


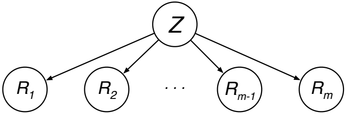


- Movie recommender system

$$Z \quad \in \quad \{1, 2, \ldots, k\} \quad \text{type of movie-goer}$$

Naive Bayes model with incomplete data

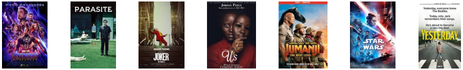



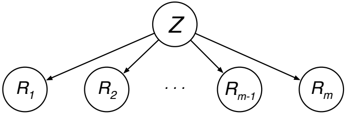



- Movie recommender system

$$
\begin{aligned}
Z &\in \{1, 2, \ldots, k\} && \text{type of movie-goer} \\
R_i &\in \{0, 1\} && \text{rating for } i^{\text{th}} \text{ movie}
\end{aligned}
$$

Naive Bayes model with incomplete data



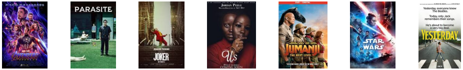


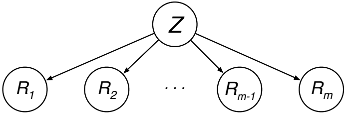
- Movie recommender system

$$\begin{aligned} Z &\in \{1, 2, \ldots, k\} && \text{type of movie-goer} \\ R_i &\in \{0, 1\} && \text{rating for } i^{\text{th}} \text{ movie} \end{aligned}$$

- Incomplete data set

Naive Bayes model with incomplete data





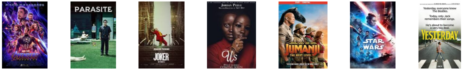
- Movie recommender system

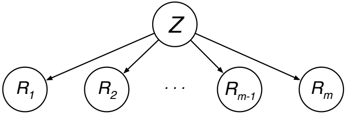
$$\begin{array}{lll l} Z & \in & \{1,2,\ldots,k\} & \text{type of movie-goer} \\ R_i & \in & \{0,1\} & \text{rating for } i^{\text{th}} \text{ movie} \end{array}$$

- Incomplete data set

student	Z	R_1	R_2	R_3	R_4	$\cdots$
1	?	0	1	1	?	$\cdots$
2	?	1	?	0	1	$\cdots$
3	?	0	0	?	1	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	?	?	1	0	?	$\cdots$

Naive Bayes model with incomplete data





- Movie recommender system

$$\begin{array}{lcll} Z & \in & \{1, 2, \ldots, k\} & \text{type of movie-goer} \\ R_i & \in & \{0, 1\} & \text{rating for } i^{\text{th}} \text{ movie} \end{array}$$

- Incomplete data set

student	Z	R_1	R_2	R_3	R_4	$\cdots$
1	?	0	1	1	?	$\cdots$
2	?	1	?	0	1	$\cdots$
3	?	0	0	?	1	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	?	?	1	0	?	$\cdots$

Note that the variable Z is never observed.

Learning from incomplete data

- Notation

Learning from incomplete data

- Notation

$$H_t \quad = \quad \text{set of hidden (latent) variables for } t^{\text{th}} \text{ example}$$

Learning from incomplete data

- Notation

$$\begin{aligned} H_t &= \text{set of hidden (latent) variables for } t^{\text{th}} \text{ example} \\ V_t &= \text{set of visible (observed) variables for } t^{\text{th}} \text{ example} \end{aligned}$$

Learning from incomplete data

- Notation

$$\begin{aligned} H_t &= \text{set of hidden (latent) variables for } t^{\text{th}} \text{ example} \\ V_t &= \text{set of visible (observed) variables for } t^{\text{th}} \text{ example} \end{aligned}$$

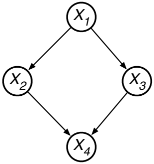
- Illustration

Learning from incomplete data

- Notation

$$\begin{aligned} H_t &= \text{set of hidden (latent) variables for } t^{\text{th}} \text{ example} \\ V_t &= \text{set of visible (observed) variables for } t^{\text{th}} \text{ example} \end{aligned}$$

- Illustration

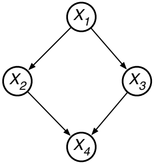


Learning from incomplete data

- Notation

$$H_t = \text{set of hidden (latent) variables for } t^{\text{th}} \text{ example}$$
$$V_t = \text{set of visible (observed) variables for } t^{\text{th}} \text{ example}$$

- Illustration



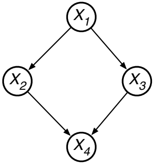
example	X_1	X_2	X_3	X_4
1	1	?	0	1
2	0	1	?	0
3	?	?	?	1
⋮	⋮	⋮	⋮	⋮

Learning from incomplete data

- Notation

$$H_t \quad = \quad \text{set of hidden (latent) variables for } t^{\text{th}} \text{ example}$$
$$V_t \quad = \quad \text{set of visible (observed) variables for } t^{\text{th}} \text{ example}$$

- Illustration



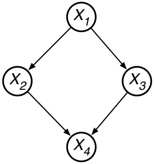
example	X_1	X_2	X_3	X_4
1	1	?	0	1
2	0	1	?	0
3	?	?	?	1
⋮	⋮	⋮	⋮	⋮

$$H_1 \quad = \quad \{X_2\} \qquad\qquad V_1 \quad = \quad \{X_1, X_3, X_4\}$$

- Notation

$$H_t = \text{set of hidden (latent) variables for } t^{\text{th}} \text{ example}$$
$$V_t = \text{set of visible (observed) variables for } t^{\text{th}} \text{ example}$$

- Illustration



example	X_1	X_2	X_3	X_4
1	1	?	0	1
2	0	1	?	0
3	?	?	?	1
⋮	⋮	⋮	⋮	⋮

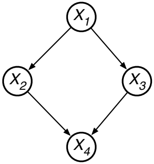
$$H_1 = \{X_2\} \qquad V_1 = \{X_1, X_3, X_4\}$$
$$H_2 = \{X_3\} \qquad V_2 = \{X_1, X_2, X_4\}$$

Learning from incomplete data

- Notation

$$H_t = \text{set of hidden (latent) variables for } t^{\text{th}} \text{ example}$$
$$V_t = \text{set of visible (observed) variables for } t^{\text{th}} \text{ example}$$

- Illustration



example	X_1	X_2	X_3	X_4
1	1	?	0	1
2	0	1	?	0
3	?	?	?	1
⋮	⋮	⋮	⋮	⋮

$$H_1 = \{X_2\} \qquad V_1 = \{X_1, X_3, X_4\}$$
$$H_2 = \{X_3\} \qquad V_2 = \{X_1, X_2, X_4\}$$
$$H_3 = \{X_1, X_2, X_3\} \qquad V_3 = \{X_4\}$$

Computing the log-likelihood with incomplete data

Computing the log-likelihood with incomplete data

$$\mathcal{L} \quad = \quad \log P(\text{data})$$

Computing the log-likelihood with incomplete data

$$
\begin{aligned}
\mathcal{L} &= \log P(\text{data}) \\
&= \log \prod_{t=1}^{T} P\left(V_t = v_t\right) \qquad \boxed{\text{data is IID}}
\end{aligned}
$$

Computing the log-likelihood with incomplete data

$$\begin{aligned}
\mathcal{L} &= \log P(\text{data}) \\
&= \log \prod_{t=1}^{T} P\left(V_t = v_t\right) && \boxed{\text{data is IID}} \\
&= \sum_{t=1}^{T} \log P\left(V_t = v_t\right) && \boxed{\log ab = \log a + \log b}
\end{aligned}$$

Computing the log-likelihood with incomplete data

$$
\begin{aligned}
\mathcal{L} &= \log P(\text{data}) \\
&= \log \prod_{t=1}^{T} P\left(V_t = v_t\right) \qquad \boxed{\text{data is IID}} \\
&= \sum_{t=1}^{T} \log P\left(V_t = v_t\right) \qquad \boxed{\log ab = \log a + \log b}
\end{aligned}
$$

$\propto H_t$

Q. What should we do next? (to express the full joint)

A. Use product rule

B. Express $P\left(V_t = v_t\right)$ using conditional independence

C. Use marginalization

D. Use Bayes Rule

E. None of them

Computing the log-likelihood with incomplete data

$$
\begin{aligned}
\mathcal{L} &= \log P(\text{data}) \\
&= \log \prod_{t=1}^{T} P\left(V_t = v_t\right) \qquad \boxed{\text{data is IID}} \\
&= \sum_{t=1}^{T} \log P\left(V_t = v_t\right) \qquad \boxed{\log ab = \log a + \log b} \\
&= \sum_{t=1}^{T} \log \sum_{h} P\left(H_t = h, V_t = v_t\right) \qquad \boxed{\text{marginalization}}
\end{aligned}
$$

Computing the log-likelihood with incomplete data

$$\begin{aligned}
\mathcal{L} &= \log P(\text{data}) \\
&= \log \prod_{t=1}^{T} P\left(V_t = v_t\right) \qquad \boxed{\text{data is IID}} \\
&= \sum_{t=1}^{T} \log P\left(V_t = v_t\right) \qquad \boxed{\log ab = \log a + \log b} \\
&= \sum_{t=1}^{T} \log \sum_{h} P\left(H_t = h, V_t = v_t\right) \qquad \boxed{\text{marginalization}} \\
&= \sum_{t=1}^{T} \log \sum_{h} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)\Big|_{\{H_t = h, V_t = v_t\}} \qquad \boxed{\text{joint}}
\end{aligned}$$

Computing the log-likelihood with incomplete data

$$
\begin{aligned}
\mathcal{L} &= \log P(\text{data}) \\
&= \log \prod_{t=1}^{T} P\left(V_t = v_t\right) && \boxed{\text{data is IID}} \\
&= \sum_{t=1}^{T} \log P\left(V_t = v_t\right) && \boxed{\log ab = \log a + \log b} \\
&= \sum_{t=1}^{T} \log \sum_{h} P\left(H_t = h, V_t = v_t\right) && \boxed{\text{marginalization}} \\
&= \sum_{t=1}^{T} \log \sum_{h} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)\Big|_{\{H_t = h, V_t = v_t\}} && \boxed{\text{joint}} \\
&= \sum_{t=1}^{T} \log \sum_{h} \prod_{i=1}^{n} P(X_i = x_i | \text{pa}_i = \pi_i)\Big|_{\{H_t = h, V_t = v_t\}} && \boxed{\text{product rule}}
\end{aligned}
$$

Computing the log-likelihood with incomplete data

$$
\begin{aligned}
\mathcal{L} &= \log P(\text{data}) \\
&= \log \prod_{t=1}^{T} P\left(V_t = v_t\right) && \boxed{\text{data is IID}} \\
&= \sum_{t=1}^{T} \log P\left(V_t = v_t\right) && \boxed{\log ab = \log a + \log b} \\
&= \sum_{t=1}^{T} \log \sum_{h} P\left(H_t = h, V_t = v_t\right) && \boxed{\text{marginalization}} \\
&= \sum_{t=1}^{T} \log \sum_{h} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)\Big|_{\{H_t = h, V_t = v_t\}} && \boxed{\text{joint}} \\
&= \sum_{t=1}^{T} \log \sum_{h} \prod_{i=1}^{n} P(X_i = x_i | \text{pa}_i = \pi_i)\Big|_{\{H_t = h, V_t = v_t\}} && \boxed{\text{product rule}}
\end{aligned}
$$

Complete versus incomplete data

Complete versus incomplete data

- Complete data

Complete versus incomplete data

- Complete data

$$\mathcal{L} = \sum_{i,\pi,x} \mathrm{count}(X_i = x, \mathrm{pa}_i = \pi) \log P(X_i = x | \mathrm{pa}_i = \pi)$$

Complete versus incomplete data

- Complete data

$$\mathcal{L} = \sum_{i,\pi,x} \text{count}(X_i{=}x, \text{pa}_i{=}\pi) \log P(X_i{=}x|\text{pa}_i{=}\pi)$$

The CPTs at different nodes are decoupled!
We can compute ML estimates in closed form.

Complete versus incomplete data

- Complete data

$$\mathcal{L} = \sum_{i,\pi,x} \text{count}(X_i = x, \text{pa}_i = \pi) \log P(X_i = x | \text{pa}_i = \pi)$$

  The CPTs at different nodes are decoupled!
  We can compute ML estimates in closed form.

- Incomplete data

Complete versus incomplete data

- **Complete data**

$$\mathcal{L} \;=\; \sum_{i,\pi,x} \mathrm{count}(X_i = x, \mathrm{pa}_i = \pi) \log P(X_i = x | \mathrm{pa}_i = \pi)$$

The CPTs at different nodes are decoupled!
We can compute ML estimates in closed form.

- **Incomplete data**

$$\mathcal{L} \;=\; \sum_{t=1}^{T} \log \sum_{h} \prod_{i=1}^{n} P(X_i = x_i | \mathrm{pa}_i = \pi_i)\Big|_{\{H_t = h, V_t = v_t\}}$$

Complete versus incomplete data

- **Complete data**

$$\mathcal{L} = \sum_{i,\pi,x} \mathrm{count}(X_i = x, \mathrm{pa}_i = \pi) \log P(X_i = x | \mathrm{pa}_i = \pi)$$

The CPTs at different nodes are decoupled!
We can compute ML estimates in closed form.

- **Incomplete data**

$$\mathcal{L} = \sum_{t=1}^{T} \log \sum_{h} \prod_{i=1}^{n} P(X_i = x_i | \mathrm{pa}_i = \pi_i) \Big|_{\{H_t = h, V_t = v_t\}}$$

The CPTs are potentially all coupled.
How to proceed?

Expectation-Maximization Algorithm

EM algorithm in a nutshell

EM algorithm in a nutshell

- If only the data weren't incomplete ...

EM algorithm in a nutshell

- If only the data weren't incomplete ...

student	Z	R_1	R_2	$\cdots$
1	?	0	1	$\cdots$
2	?	1	?	$\cdots$
3	?	0	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	?	?	?	$\cdots$

If the data were complete, we could easily estimate the CPTs. What can we do instead?

EM algorithm in a nutshell

- If only the data weren't incomplete ...

student	Z	R_1	R_2	$\cdots$
1	?	0	1	$\cdots$
2	?	1	?	$\cdots$
3	?	0	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	?	?	?	$\cdots$

If the data were complete, we could easily estimate the CPTs. What can we do instead?

- Here's a crazy idea ...

EM algorithm in a nutshell

- If only the data weren't incomplete ...

student	Z	R_1	R_2	$\cdots$
1	?	0	1	$\cdots$
2	?	1	?	$\cdots$
3	?	0	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	?	?	?	$\cdots$

If the data were complete, we could easily estimate the CPTs. What can we do instead?

- Here's a crazy idea ...

Randomly initialize the CPTs with nonzero elements.

EM algorithm in a nutshell

- If only the data weren't incomplete ...

student	Z	R_1	R_2	$\cdots$
1	?	0	1	$\cdots$
2	?	1	?	$\cdots$
3	?	0	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	?	?	?	$\cdots$

If the data were complete, we could easily estimate the CPTs. What can we do instead?

- Here's a crazy idea ...

Randomly initialize the CPTs with nonzero elements.
Use these CPTs to infer values for the missing data.

EM algorithm in a nutshell

- If only the data weren't incomplete ...

student	Z	R_1	R_2	$\cdots$
1	?	0	1	$\cdots$
2	?	1	?	$\cdots$
3	?	0	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	?	?	?	$\cdots$

If the data were complete, we could easily estimate the CPTs. What can we do instead?

- Here's a crazy idea ...

Randomly initialize the CPTs with nonzero elements.
Use these CPTs to infer values for the missing data.
Re-estimate CPTs from the newly completed data.

EM algorithm in a nutshell

- If only the data weren't incomplete ...

student	Z	R_1	R_2	$\cdots$
1	?	0	1	$\cdots$
2	?	1	?	$\cdots$
3	?	0	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	?	?	?	$\cdots$

If the data were complete, we could easily estimate the CPTs. What can we do instead?

- Here's a crazy idea ...

Randomly initialize the CPTs with nonzero elements.
Use these CPTs to infer values for the missing data.
Re-estimate CPTs from the newly completed data.
Iterate the last two steps until convergence?

EM algorithm in a nutshell

- If only the data weren't incomplete ...

student	Z	R_1	R_2	$\cdots$
1	?	0	1	$\cdots$
2	?	1	?	$\cdots$
3	?	0	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	?	?	?	$\cdots$

If the data were complete, we could easily estimate the CPTs. What can we do instead?

- Here's a crazy idea ...

Randomly initialize the CPTs with nonzero elements.
Use these CPTs to infer values for the missing data.
Re-estimate CPTs from the newly completed data.
Iterate the last two steps until convergence?

Amazingly, this is how EM works (more or less) ...

- Initialize the CPTs

EM algorithm — overview

- Initialize the CPTs

  Assign random probabilities to all $P(X_i = x | \mathrm{pa}_i = \pi)$.

EM algorithm — overview

- Initialize the CPTs

  Assign random probabilities to all $P(X_i = x | \mathrm{pa}_i = \pi)$.

  Avoid zero probabilities (which cannot be unlearned).

- Initialize the CPTs

  Assign random probabilities to all $P(X_i = x | \mathrm{pa}_i = \pi)$.

  Avoid zero probabilities (which cannot be unlearned).

  Different initializations may yield different results.

EM algorithm — overview

- Initialize the CPTs

  Assign random probabilities to all $P(X_i = x | \mathrm{pa}_i = \pi)$.

  Avoid zero probabilities (which cannot be unlearned).

  Different initializations may yield different results.

- Iterate until convergence

EM algorithm — overview

- **Initialize the CPTs**

  Assign random probabilities to all $P(X_i = x|\mathrm{pa}_i = \pi)$.

  Avoid zero probabilities (which cannot be unlearned).

  Different initializations may yield different results.

- **Iterate until convergence**

  [E-Step] Compute posterior probabilities $P(H_t = h|V_t = v_t)$.

EM algorithm — overview

- **Initialize the CPTs**

  Assign random probabilities to all $P(X_i = x | \mathrm{pa}_i = \pi)$.

  Avoid zero probabilities (which cannot be unlearned).

  Different initializations may yield different results.

- **Iterate until convergence**

  [E-Step] Compute posterior probabilities $P(H_t = h | V_t = v_t)$.

  [M-Step] Update CPTs based on these probabilities.

E-step (Inference)

E-step (Inference)

To fill in missing data, we must compute posterior probabilities.

E-step (Inference)

To fill in missing data, we must compute posterior probabilities. But which probabilities, specifically, do we need?

E-step (Inference)

To fill in missing data, we must compute posterior probabilities. But which probabilities, specifically, do we need?

At root nodes:

To fill in missing data, we must compute posterior probabilities. But which probabilities, specifically, do we need?

At root nodes: $P(X_i = x | V_t = v_t)$

E-step (Inference)

To fill in missing data, we must compute posterior probabilities. But which probabilities, specifically, do we need?

At root nodes: $P(X_i = x | V_t = v_t)$

At other nodes:

E-step (Inference)

To fill in missing data, we must compute posterior probabilities. But which probabilities, specifically, do we need?

At root nodes: $P(X_i = x | V_t = v_t)$

At other nodes: $P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)$

E-step (Inference)

To fill in missing data, we must compute posterior probabilities. But which probabilities, specifically, do we need?

$$
\begin{aligned}
\textbf{At root nodes:} &\quad P(X_i = x | V_t = v_t) \\
\textbf{At other nodes:} &\quad P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)
\end{aligned}
$$

These probabilities must be computed over a quadruple loop:

E-step (Inference)

To fill in missing data, we must compute posterior probabilities. But which probabilities, specifically, do we need?

At root nodes: $P(X_i = x | V_t = v_t)$
At other nodes: $P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)$

These probabilities must be computed over a quadruple loop:

examples $V_t \qquad t \in \{1, 2, \ldots, T\}$

E-step (Inference)

To fill in missing data, we must compute posterior probabilities. But which probabilities, specifically, do we need?

$$\begin{aligned} \textbf{At root nodes:} \quad & P(X_i\!=\!x|V_t\!=\!v_t) \\ \textbf{At other nodes:} \quad & P(X_i\!=\!x, \mathrm{pa}_i\!=\!\pi|V_t\!=\!v_t) \end{aligned}$$

These probabilities must be computed over a quadruple loop:

$$\begin{aligned} \textbf{examples } V_t \qquad & t \in \{1, 2, \ldots, T\} \\ \textbf{nodes } X_i \qquad & i \in \{1, 2, \ldots, n\} \end{aligned}$$

E-step (Inference)

To fill in missing data, we must compute posterior probabilities. But which probabilities, specifically, do we need?

At root nodes: $P(X_i = x | V_t = v_t)$

At other nodes: $P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)$

These probabilities must be computed over a quadruple loop:

examples V_t	$t \in \{1, 2, \ldots, T\}$
nodes X_i	$i \in \{1, 2, \ldots, n\}$
values of $X_i = x$	e.g., $x \in \{0, 1\}$

E-step (Inference)

To fill in missing data, we must compute posterior probabilities. But which probabilities, specifically, do we need?

At root nodes: $P(X_i = x | V_t = v_t)$

At other nodes: $P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)$

These probabilities must be computed over a quadruple loop:

examples V_t	$t \in \{1, 2, \ldots, T\}$
nodes X_i	$i \in \{1, 2, \ldots, n\}$
values of $X_i = x$	e.g., $x \in \{0, 1\}$
values of $\mathrm{pa}_i = \pi$	e.g., $\pi \in \{0, 1\}^k$

E-step (Inference)

To fill in missing data, we must compute posterior probabilities. But which probabilities, specifically, do we need?

$$\begin{aligned} \textbf{At root nodes:} \quad & P(X_i = x | V_t = v_t) \\ \textbf{At other nodes:} \quad & P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t) \end{aligned}$$

These probabilities must be computed over a quadruple loop:

$$\begin{array}{rl} \text{examples } V_t & t \in \{1, 2, \ldots, T\} \\ \text{nodes } X_i & i \in \{1, 2, \ldots, n\} \\ \text{values of } X_i = x & \text{e.g., } x \in \{0, 1\} \\ \text{values of } \mathrm{pa}_i = \pi & \text{e.g., } \pi \in \{0, 1\}^k \end{array}$$

The # of computations grows linearly in the size of the BN,

E-step (Inference)

To fill in missing data, we must compute posterior probabilities. But which probabilities, specifically, do we need?

$$\textbf{At root nodes:}\quad P(X_i\!=\!x|V_t\!=\!v_t)$$
$$\textbf{At other nodes:}\quad P(X_i\!=\!x, \mathrm{pa}_i\!=\!\pi|V_t\!=\!v_t)$$

These probabilities must be computed over a quadruple loop:

examples V_t	$t \in \{1, 2, \ldots, T\}$
nodes X_i	$i \in \{1, 2, \ldots, n\}$
values of $X_i\!=\!x$	e.g., $x \in \{0, 1\}$
values of $\mathrm{pa}_i\!=\!\pi$	e.g., $\pi \in \{0, 1\}^k$

The # of computations grows linearly in the size of the BN, and also in the amount of data (as expected).

M-step (Learning)

M-step (Learning)

Next we use these posterior probabilities to update CPTs:

M-step (Learning)

Next we use these posterior probabilities to update CPTs:

- At root nodes

M-step (Learning)

Next we use these posterior probabilities to update CPTs:

- At root nodes

Post

$$P(X_i = x) \longleftarrow \frac{1}{T}\sum_{t=1}^{T} P(X_i = x | V_t = v_t)$$

M-step (Learning)

Next we use these posterior probabilities to update CPTs:

- At root nodes

$$P(X_i = x) \quad \longleftarrow \quad \frac{1}{T} \sum_{t=1}^{T} P(X_i = x | V_t = v_t)$$

- At nodes with parents

M-step (Learning)

Next we use these posterior probabilities to update CPTs:

- At root nodes

$$P(X_i = x) \quad \longleftarrow \quad \frac{1}{T}\sum_{t=1}^{T} P(X_i = x | V_t = v_t)$$

- At nodes with parents

$$P(X_i = x | \mathrm{pa}_i = \pi) \quad \longleftarrow \quad \frac{\sum_{t=1} P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_{t=1}^{T} P(\mathrm{pa}_i = \pi | V_t = v_t)}$$

M-step (Learning)

Next we use these posterior probabilities to update CPTs:

- At root nodes

$$P(X_i = x) \longleftarrow \frac{1}{T} \sum_{t=1}^{T} P(X_i = x | V_t = v_t)$$

- At nodes with parents

$$P(X_i = x | \mathrm{pa}_i = \pi) \longleftarrow \frac{\sum_{t=1} P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_{t=1}^{T} P(\mathrm{pa}_i = \pi | V_t = v_t)}$$

Note that these are updates ($\longleftarrow$), not equalities ($=$).
The right hand sides depend on the current CPTs.

M-step (Learning)

Next we use these posterior probabilities to update CPTs:

- At root nodes

$$P(X_i = x) \quad \longleftarrow \quad \frac{1}{T}\sum_{t=1}^{T} P(X_i = x | V_t = v_t)$$

- At nodes with parents

$$P(X_i = x | \mathrm{pa}_i = \pi) \quad \longleftarrow \quad \frac{\sum_{t=1} P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_{t=1}^{T} P(\mathrm{pa}_i = \pi | V_t = v_t)}$$

Note that these are updates ($\longleftarrow$), not equalities ($=$).
The right hand sides depend on the current CPTs.

Formulas are great, but what about intuition?

Analogy to ML for complete data

Analogy to ML for complete data

- Indicator functions

Analogy to ML for complete data

- Indicator functions

$$I(x, x') = \begin{cases} 1 & \text{if } x = x' \\ 0 & \text{otherwise} \end{cases}$$

Analogy to ML for complete data

- Indicator functions

$$I(x, x') = \begin{cases} 1 & \text{if } x = x' \\ 0 & \text{otherwise} \end{cases}$$

- Counts

Analogy to ML for complete data

- Indicator functions

$$I(x, x') = \begin{cases} 1 & \text{if } x = x' \\ 0 & \text{otherwise} \end{cases}$$

- Counts

$$\mathrm{count}(X_i = x) = \sum_{t=1}^{T} I(x_{it}, x)$$

Analogy to ML for complete data

- Indicator functions

$$I(x, x') = \begin{cases} 1 & \text{if } x = x' \\ 0 & \text{otherwise} \end{cases}$$

- Counts

$$\mathrm{count}(X_i = x) = \sum_{t=1}^{T} I(x_{it}, x)$$

$$\mathrm{count}(\mathrm{pa}_i = \pi) = \sum_{t=1}^{T} I(\mathrm{pa}_{it}, \pi)$$

Analogy to ML for complete data

- Indicator functions

$$I(x, x') = \begin{cases} 1 & \text{if } x = x' \\ 0 & \text{otherwise} \end{cases}$$

- Counts

$$\mathrm{count}(X_i = x) = \sum_{t=1}^{T} I(x_{it}, x)$$

$$\mathrm{count}(\mathrm{pa}_i = \pi) = \sum_{t=1}^{T} I(\mathrm{pa}_{it}, \pi)$$

$$\mathrm{count}(X_i = x, \mathrm{pa}_i = \pi) = \sum_{t=1}^{T} I(x_{it}, x)\, I(\mathrm{pa}_{it}, \pi)$$

ML estimates for complete data

ML estimates for complete data

- At root nodes

ML estimates for complete data

- At root nodes

$$P_{\mathrm{ML}}(X_i = x) \quad = \quad \frac{\mathrm{count}(X_i = x)}{T}$$

ML estimates for complete data

- At root nodes

$$P_{\mathrm{ML}}(X_i = x) = \frac{\mathrm{count}(X_i = x)}{T}$$

$$P_{\mathrm{ML}}(X_i = x) = \frac{1}{T}\sum_{t=1}^{T} I(x_{it}, x)$$

ML estimates for complete data

- At root nodes

$$P_{\mathrm{ML}}(X_i\!=\!x) \quad = \quad \frac{\mathrm{count}(X_i\!=\!x)}{T}$$

$$P_{\mathrm{ML}}(X_i\!=\!x) \quad = \quad \frac{1}{T}\sum_{t=1}^{T} I(x_{it}, x)$$

- At nodes with parents

ML estimates for complete data

- At root nodes

$$P_{\mathrm{ML}}(X_i\!=\!x) \;=\; \frac{\mathrm{count}(X_i\!=\!x)}{T}$$

$$P_{\mathrm{ML}}(X_i\!=\!x) \;=\; \frac{1}{T}\sum_{t=1}^{T} I(x_{it}, x)$$

- At nodes with parents

$$P_{\mathrm{ML}}(X_i\!=\!x|\mathrm{pa}_i\!=\!\pi) \;=\; \frac{\mathrm{count}(X_i\!=\!x, \mathrm{pa}_i\!=\!\pi)}{\mathrm{count}(\mathrm{pa}_i\!=\!\pi)}$$

ML estimates for complete data

- At root nodes

$$P_{\mathrm{ML}}(X_i\!=\!x) \quad = \quad \frac{\mathrm{count}(X_i\!=\!x)}{T}$$

$$P_{\mathrm{ML}}(X_i\!=\!x) \quad = \quad \frac{1}{T}\sum_{t=1}^{T} I(x_{it}, x)$$

- At nodes with parents

$$P_{\mathrm{ML}}(X_i\!=\!x|\mathrm{pa}_i\!=\!\pi) \quad = \quad \frac{\mathrm{count}(X_i\!=\!x, \mathrm{pa}_i\!=\!\pi)}{\mathrm{count}(\mathrm{pa}_i\!=\!\pi)}$$

$$P_{\mathrm{ML}}(X_i\!=\!x|\mathrm{pa}_i\!=\!\pi) \quad = \quad \frac{\sum_{t=1}^{T} I(x_{it}, x)\, I(\mathrm{pa}_{it}, \pi)}{\sum_{t=1}^{T} I(\mathrm{pa}_{it}, \pi)}$$

Intuition for EM updates — by analogy

Intuition for EMI updates — by analogy

- At root nodes

Intuition for EM updates — by analogy

- At root nodes

$$P_{\mathrm{ML}}(X_i = x) \quad = \quad \frac{1}{T}\sum_t I(x_{it}, x) \qquad \boxed{\text{ML for complete data}}$$

Intuition for EM updates — by analogy

- At root nodes

$$P_{\mathrm{ML}}(X_i = x) \quad = \quad \frac{1}{T}\sum_t I(x_{it}, x) \qquad \boxed{\text{ML for complete data}}$$

$$P(X_i = x) \quad \leftarrow \quad \frac{1}{T}\sum_t P(X_i = x | V_t = v_t) \qquad \boxed{\text{EM update}}$$

Intuition for EM updates — by analogy

- At root nodes

$$P_{\mathrm{ML}}(X_i\!=\!x) \quad = \quad \frac{1}{T}\sum_t I(x_{it}, x) \qquad \boxed{\text{ML for complete data}}$$

$$P(X_i\!=\!x) \quad \leftarrow \quad \frac{1}{T}\sum_t P(X_i\!=\!x|V_t\!=\!v_t) \qquad \boxed{\text{EM update}}$$

- At nodes with parents

Intuition for EM updates — by analogy

- At root nodes

$$P_{\mathrm{ML}}(X_i = x) \quad = \quad \frac{1}{T}\sum_t I(x_{it}, x) \qquad \boxed{\text{ML for complete data}}$$

$$P(X_i = x) \quad \leftarrow \quad \frac{1}{T}\sum_t P(X_i = x | V_t = v_t) \qquad \boxed{\text{EM update}}$$

- At nodes with parents

$$P_{\mathrm{ML}}(X_i = x | \mathrm{pa}_i = \pi) \quad = \quad \frac{\sum_t I(x_{it}, x)\, I(\mathrm{pa}_{it}, \pi)}{\sum_t I(\mathrm{pa}_{it}, \pi)} \qquad \boxed{\text{ML for complete data}}$$

Intuition for EM updates — by analogy

- At root nodes

$$P_{\mathrm{ML}}(X_i = x) \quad = \quad \frac{1}{T}\sum_t I(x_{it}, x) \qquad \boxed{\text{ML for complete data}}$$

$$P(X_i = x) \quad \leftarrow \quad \frac{1}{T}\sum_t P(X_i = x | V_t = v_t) \qquad \boxed{\text{EM update}}$$

- At nodes with parents

$$P_{\mathrm{ML}}(X_i = x | \mathrm{pa}_i = \pi) \quad = \quad \frac{\sum_t I(x_{it}, x)\, I(\mathrm{pa}_{it}, \pi)}{\sum_t I(\mathrm{pa}_{it}, \pi)} \qquad \boxed{\text{ML for complete data}}$$

$$P(X_i = x | \mathrm{pa}_i = \pi) \quad \leftarrow \quad \frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi | V_t = v_t)} \qquad \boxed{\text{EM update}}$$

Intuition for EM updates — by analogy

- At root nodes

$$P_{\mathrm{ML}}(X_i = x) \quad = \quad \frac{1}{T}\sum_t I(x_{it}, x) \qquad \boxed{\text{ML for complete data}}$$

$$P(X_i = x) \quad \leftarrow \quad \frac{1}{T}\sum_t P(X_i = x | V_t = v_t) \qquad \boxed{\text{EM update}}$$

- At nodes with parents

$$P_{\mathrm{ML}}(X_i = x | \mathrm{pa}_i = \pi) \quad = \quad \frac{\sum_t I(x_{it}, x)\, I(\mathrm{pa}_{it}, \pi)}{\sum_t I(\mathrm{pa}_{it}, \pi)} \qquad \boxed{\text{ML for complete data}}$$

$$P(X_i = x | \mathrm{pa}_i = \pi) \quad \leftarrow \quad \frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi | V_t = v_t)} \qquad \boxed{\text{EM update}}$$

- Special case

Intuition for EM updates — by analogy

- At root nodes

$$P_{\mathrm{ML}}(X_i\!=\!x) \quad = \quad \frac{1}{T}\sum_t I(x_{it}, x) \qquad \boxed{\text{ML for complete data}}$$

$$P(X_i\!=\!x) \quad \leftarrow \quad \frac{1}{T}\sum_t P(X_i\!=\!x|V_t\!=\!v_t) \qquad \boxed{\text{EM update}}$$

- At nodes with parents

$$P_{\mathrm{ML}}(X_i\!=\!x|\mathrm{pa}_i\!=\!\pi) \quad = \quad \frac{\sum_t I(x_{it}, x)\, I(\mathrm{pa}_{it}, \pi)}{\sum_t I(\mathrm{pa}_{it}, \pi)} \qquad \boxed{\text{ML for complete data}}$$

$$P(X_i\!=\!x|\mathrm{pa}_i\!=\!\pi) \quad \leftarrow \quad \frac{\sum_t P(X_i\!=\!x, \mathrm{pa}_i\!=\!\pi|V_t\!=\!v_t)}{\sum_t P(\mathrm{pa}_i\!=\!\pi|V_t\!=\!v_t)} \qquad \boxed{\text{EM update}}$$

- Special case

  Consider a CPT whose nodes are fully observed.

Intuition for EM updates — by analogy

- At root nodes

$$P_{\mathrm{ML}}(X_i = x) = \frac{1}{T}\sum_t I(x_{it}, x) \qquad \boxed{\text{ML for complete data}}$$

$$P(X_i = x) \leftarrow \frac{1}{T}\sum_t P(X_i = x | V_t = v_t) \qquad \boxed{\text{EM update}}$$

- At nodes with parents

$$P_{\mathrm{ML}}(X_i = x | \mathrm{pa}_i = \pi) = \frac{\sum_t I(x_{it}, x)\, I(\mathrm{pa}_{it}, \pi)}{\sum_t I(\mathrm{pa}_{it}, \pi)} \qquad \boxed{\text{ML for complete data}}$$

$$P(X_i = x | \mathrm{pa}_i = \pi) \leftarrow \frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi | V_t = v_t)} \qquad \boxed{\text{EM update}}$$

- Special case

Consider a CPT whose nodes are fully observed.
EM updates in this case reduce to ML estimates for complete data.

EM updates

EM updates

$P(X_i = x)$	$\longleftarrow$	$\frac{1}{T}\sum_t P(X_i = x \vert V_t = v_t)$	root nodes
$P(X_i = x \vert \mathrm{pa}_i = \pi)$	$\longleftarrow$	$\frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi \vert V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi \vert V_t = v_t)}$	nodes with parents

EM updates

$P(X_i = x)$	$\longleftarrow$	$\frac{1}{T} \sum_t P(X_i = x \mid V_t = v_t)$	root nodes
$P(X_i = x \mid \mathrm{pa}_i = \pi)$	$\longleftarrow$	$\frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi \mid V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi \mid V_t = v_t)}$	nodes with parents

Intuitively:

EM updates

$P(X_i = x)$	$\longleftarrow$	$\dfrac{1}{T}\sum_t P(X_i = x \mid V_t = v_t)$	root nodes
$P(X_i = x \mid \mathrm{pa}_i = \pi)$	$\longleftarrow$	$\dfrac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi \mid V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi \mid V_t = v_t)}$	nodes with parents

Intuitively:

When the data is complete, we estimate the CPTs from observed counts.

EM updates

$P(X_i = x)$	$\longleftarrow$	$\frac{1}{T}\sum_t P(X_i = x \mid V_t = v_t)$	root nodes
$P(X_i = x \mid \mathrm{pa}_i = \pi)$	$\longleftarrow$	$\frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi \mid V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi \mid V_t = v_t)}$	nodes with parents

Intuitively:

When the data is complete, we estimate the CPTs from observed counts.

When the data is incomplete, we re-estimate the CPTs from expected counts.

EM updates

$P(X_i = x)$	$\longleftarrow$	$\frac{1}{T}\sum_t P(X_i = x \| V_t = v_t)$	root nodes
$P(X_i = x \| \mathrm{pa}_i = \pi)$	$\longleftarrow$	$\frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi \| V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi \| V_t = v_t)}$	nodes with parents

Intuitively:

When the data is complete, we estimate the CPTs from observed counts.

When the data is incomplete, we re-estimate the CPTs from expected counts.

These expected counts are computed from the posterior distributions $P(h|v_t)$.

Key properties of EM

Key properties of EM

- No learning rate

Key properties of EM

- No learning rate

  The updates do not require the tuning of a learning rate ($\eta > 0$), as in most gradient-based methods.

Key properties of EM

- No learning rate

  The updates do not require the tuning of a learning rate ($\eta > 0$), as in most gradient-based methods.

- Monotonic convergence

Key properties of EM

- **No learning rate**

  The updates do not require the tuning of a learning rate ($\eta > 0$), as in most gradient-based methods.

- **Monotonic convergence**

  The updated CPTs from EM always increase the incomplete-data log-likelihood $\mathcal{L} = \sum_t \log P(V_t = v_t)$.

Check In

Q. How much of EM did you understand?

A. (Nearly) All of it 15%

B. Some of it, but I have some doubts 40%

C. Maybe a little, but I'm pretty confused 35%

D. Almost none of it; I'm totally lost

E. Zzz...

Log-likelihood

Log-likelihood

- Incomplete data set

Log-likelihood

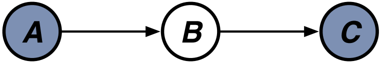
- Incomplete data set

t	A	B	C
1	a_1	?	c_1
2	a_2	?	c_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	a_T	?	c_T

Log-likelihood

- Incomplete data set

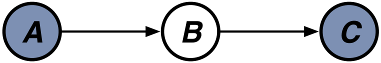
t	A	B	C
1	a_1	?	c_1
2	a_2	?	c_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	a_T	?	c_T



Log-likelihood

- Incomplete data set

t	A	B	C
1	a_1	?	c_1
2	a_2	?	c_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	a_T	?	c_T

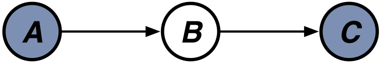


How to choose the CPTs
to maximize the log-likelihood
of this (incomplete) data?

Log-likelihood

- Incomplete data set

t	A	B	C
1	a_1	?	c_1
2	a_2	?	c_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	a_T	?	c_T



How to choose the CPTs
to maximize the log-likelihood
of this (incomplete) data?

- Log-likelihood

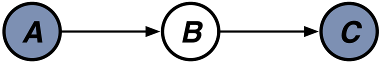
Log-likelihood

- Incomplete data set

t	A	B	C
1	a_1	?	c_1
2	a_2	?	c_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	a_T	?	c_T

- Log-likelihood

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

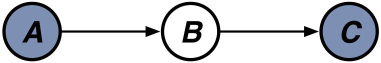


How to choose the CPTs
to maximize the log-likelihood
of this (incomplete) data?

Log-likelihood

- Incomplete data set

t	A	B	C
1	a_1	?	c_1
2	a_2	?	c_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	a_T	?	c_T



How to choose the CPTs
to maximize the log-likelihood
of this (incomplete) data?

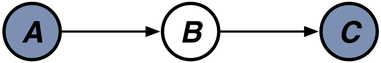
- Log-likelihood

$$\begin{aligned}\mathcal{L} &= \sum_t \log P(a_t, c_t) \\ &= \sum_t \log \sum_b P(a_t, b, c_t) \qquad \boxed{\text{marginalization}}\end{aligned}$$

Log-likelihood

- Incomplete data set

t	A	B	C
1	a_1	?	c_1
2	a_2	?	c_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	a_T	?	c_T



How to choose the CPTs
to maximize the log-likelihood
of this (incomplete) data?

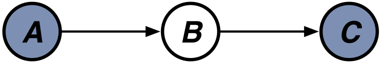
- Log-likelihood

$$
\begin{aligned}
\mathcal{L} &= \sum_t \log P(a_t, c_t) \\
&= \sum_t \log \sum_b P(a_t, b, c_t) \qquad \boxed{\text{marginalization}} \\
&= \sum_t \log \sum_b P(a_t)\, P(b|a_t)\, P(c_t|a_t, b) \qquad \boxed{\text{product rule}}
\end{aligned}
$$

Log-likelihood

- Incomplete data set

t	A	B	C
1	a_1	?	c_1
2	a_2	?	c_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	a_T	?	c_T

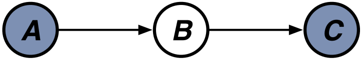


How to choose the CPTs
to maximize the log-likelihood
of this (incomplete) data?

- Log-likelihood

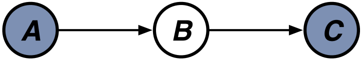
$$
\begin{aligned}
\mathcal{L} &= \sum_t \log P(a_t, c_t) \\
&= \sum_t \log \sum_b P(a_t, b, c_t) && \boxed{\text{marginalization}} \\
&= \sum_t \log \sum_b P(a_t)\, P(b|a_t)\, P(c_t|a_t, b) && \boxed{\text{product rule}} \\
&= \sum_t \log \sum_b P(a_t)\, P(b|a_t)\, P(c_t|b) && \boxed{\text{conditional independence}}
\end{aligned}
$$

Example



Suppose that A and C are observed and B is hidden.

Example



Suppose that A and C are observed and B is hidden.

Q. Which parameters of this network can you estimate directly from the data (in one step—no iteration required)?

A. $P(A)$

B. $P(B|A)$

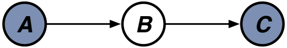
C. $P(C|B)$

D. Both A. and C.

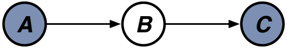
E. None of them

EM update for $P(A)$

EM update for $P(A)$

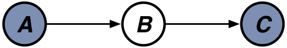


EM update for $P(A)$



- General form

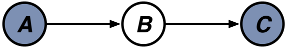
EM update for $P(A)$



- General form

$$P(X_i = x) \longleftarrow \frac{1}{T}\sum_t P(X_i = x | V_t = v_t) \qquad \boxed{\text{root node}}$$

EM update for $P(A)$

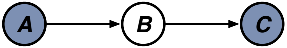


- General form

$$P(X_i = x) \longleftarrow \frac{1}{T} \sum_t P(X_i = x | V_t = v_t) \qquad \boxed{\text{root node}}$$

- Update for this CPT

EM update for $P(A)$



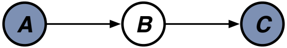
- General form

$$P(X_i = x) \longleftarrow \frac{1}{T} \sum_t P(X_i = x | V_t = v_t) \qquad \boxed{\text{root node}}$$

- Update for this CPT

$$P(A = a)$$

EM update for $P(A)$



- General form

$$P(X_i\!=\!x) \longleftarrow \frac{1}{T}\sum_t P(X_i\!=\!x|V_t\!=\!v_t) \qquad \boxed{\text{root node}}$$

- Update for this CPT

$$P(A\!=\!a) \longleftarrow \frac{1}{T}\sum_t P(A\!=\!a|A\!=\!a_t, C\!=\!c_t)$$

?

EM update for $P(A)$

$$A \longrightarrow B \longrightarrow C$$

- General form

$$P(X_i = x) \longleftarrow \frac{1}{T}\sum_t P(X_i = x | V_t = v_t) \qquad \boxed{\text{root node}}$$

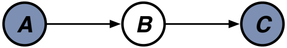
- Update for this CPT

$$P(A = a) \quad \longleftarrow \quad \frac{1}{T}\sum_t P(A = a | A = a_t, C = c_t)$$

Simplify:

$$P(A = a) \quad \longleftarrow$$

EM update for $P(A)$



- General form

$$P(X_i = x) \longleftarrow \frac{1}{T} \sum_t P(X_i = x | V_t = v_t)$$

root node

- Update for this CPT

$$P(A = a) \quad \longleftarrow \quad \frac{1}{T} \sum_t P(A = a | A = a_t, C = c_t)$$

Simplify:

$$P(A = a) \quad \longleftarrow \quad \frac{1}{T} \sum_t I(a, a_t)$$

EM update for $P(A)$

$$A \longrightarrow B \longrightarrow C$$

- General form

$$P(X_i\!=\!x) \;\longleftarrow\; \frac{1}{T}\sum_t P(X_i\!=\!x|V_t\!=\!v_t) \qquad \boxed{\text{root node}}$$

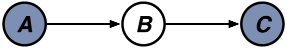
- Update for this CPT

$$P(A\!=\!a) \quad\longleftarrow\quad \frac{1}{T}\sum_t P(A\!=\!a|A\!=\!a_t, C\!=\!c_t)$$

Simplify:

$$P(A\!=\!a) \quad\longleftarrow\quad \frac{1}{T}\sum_t I(a, a_t) \;=\; \frac{1}{T}\,\mathrm{count}(A\!=\!a)$$

EM update for $P(A)$



- General form

$$P(X_i = x) \longleftarrow \frac{1}{T}\sum_t P(X_i = x | V_t = v_t) \qquad \boxed{\text{root node}}$$

- Update for this CPT

$$P(A = a) \quad \longleftarrow \quad \frac{1}{T}\sum_t P(A = a | A = a_t, C = c_t)$$

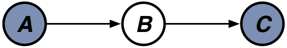
Simplify:

$$P(A = a) \quad \longleftarrow \quad \frac{1}{T}\sum_t I(a, a_t) \;=\; \frac{1}{T}\,\mathrm{count}(A = a)$$

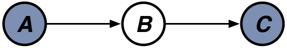
The update reduces to the ML estimate for complete data—as it must, because A is observed and has no unobserved parents.

EM update for $P(B|A)$

EM update for $P(B|A)$

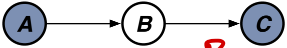


EM update for $P(B|A)$



- General form

EM update for $P(B|A)$

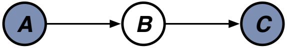


- General form

B A

$$P(X_i = x | \mathrm{pa}_i = \pi) \longleftarrow \frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi | V_t = v_t)}$$

EM update for $P(B|A)$

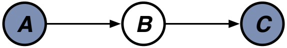


- General form

$$P(X_i = x | \mathrm{pa}_i = \pi) \longleftarrow \frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi | V_t = v_t)}$$

- Update for this CPT

EM update for $P(B|A)$



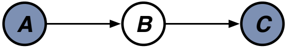
- General form

$$P(X_i = x | \mathrm{pa}_i = \pi) \longleftarrow \frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi | V_t = v_t)}$$

- Update for this CPT

$$P(B = b | A = a)$$

EM update for $P(B|A)$



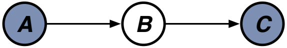
- General form

$$P(X_i = x | \mathrm{pa}_i = \pi) \longleftarrow \frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi | V_t = v_t)}$$

- Update for this CPT

$$P(B = b | A = a) \quad \longleftarrow \quad \frac{\sum_t P(B = b, A = a | A = a_t, C = c_t)}{}$$

EM update for $P(B|A)$



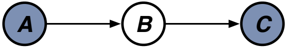
- General form

$$P(X_i = x | \mathrm{pa}_i = \pi) \longleftarrow \frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi | V_t = v_t)}$$

- Update for this CPT

$$P(B = b | A = a) \quad \longleftarrow \quad \frac{\sum_t P(B = b, A = a | A = a_t, C = c_t)}{\sum_t P(A = a | A = a_t, C = c_t)}$$

EM update for $P(B|A)$



- General form

$$P(X_i = x | \mathrm{pa}_i = \pi) \longleftarrow \frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi | V_t = v_t)}$$

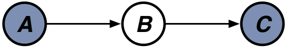
- Update for this CPT

$$P(B = b | A = a) \quad \longleftarrow \quad \frac{\sum_t P(B = b, A = a | A = a_t, C = c_t)}{\sum_t P(A = a | A = a_t, C = c_t)}$$

Simplify:

$$P(B = b | A = a) \quad \longleftarrow$$

EM update for $P(B|A)$



- General form

$$P(X_i = x | \mathrm{pa}_i = \pi) \longleftarrow \frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi | V_t = v_t)}$$

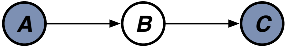
- Update for this CPT

$$P(B = b | A = a) \quad \longleftarrow \quad \frac{\sum_t P(B = b, A = a | A = a_t, C = c_t)}{\sum_t P(A = a | A = a_t, C = c_t)}$$

Simplify:

$$P(B = b | A = a) \quad \longleftarrow \quad \frac{\sum_t I(a, a_t)\, P(B = b | A = a_t, C = c_t)}{}$$

EM update for $P(B|A)$



- General form

$$P(X_i = x|\mathrm{pa}_i = \pi) \longleftarrow \frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi | V_t = v_t)}$$

- Update for this CPT

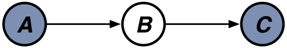
CPT

$$P(B = b|A = a) \longleftarrow \frac{\sum_t P(B = b, A = a | A = a_t, C = c_t)}{\sum_t P(A = a | A = a_t, C = c_t)}$$

Simplify:

$$P(B = b|A = a) \longleftarrow \frac{\sum_t I(a, a_t) P(B = b | A = a_t, C = c_t)}{\sum_t I(a, a_t)}$$

EM update for $P(B|A)$



- General form

$$P(X_i = x | \mathrm{pa}_i = \pi) \longleftarrow \frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi | V_t = v_t)}$$

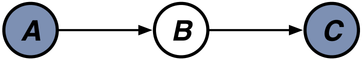
- Update for this CPT

$$P(B = b | A = a) \longleftarrow \frac{\sum_t P(B = b, A = a | A = a_t, C = c_t)}{\sum_t P(A = a | A = a_t, C = c_t)}$$

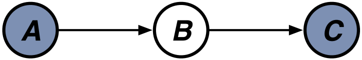
Simplify:

$$P(B = b | A = a) \longleftarrow \frac{\sum_t I(a, a_t) \overbrace{P(B = b | A = a_t, C = c_t)}^{\text{computed from Bayes rule}}}{\sum_t I(a, a_t)}$$

Example



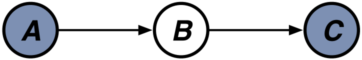
Example



Suppose that A and C are observed and B is hidden.

- Inference

Example

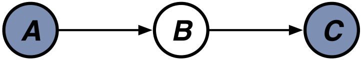


Suppose that A and C are observed and B is hidden.

- Inference

  $P(B\!=\!b|A\!=\!a, C\!=\!c)$

Example

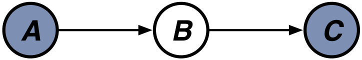


Suppose that A and C are observed and B is hidden.

- Inference

$$P(B{=}b|A{=}a, C{=}c) \quad = \quad \frac{P(C{=}c|B{=}b, A{=}a)\, P(B{=}b|A{=}a)}{P(C{=}c|A{=}a)} \qquad \boxed{\text{BR}}$$

Example



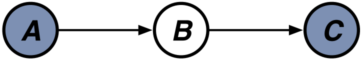
Suppose that A and C are observed and B is hidden.

- Inference

$$P(B{=}b|A{=}a, C{=}c) = \frac{P(C{=}c|B{=}b, A{=}a)\, P(B{=}b|A{=}a)}{P(C{=}c|A{=}a)} \quad \boxed{\text{BR}}$$

$$= \frac{P(C{=}c|B{=}b)\, P(B{=}b|A{=}a)}{P(C{=}c|A{=}a)} \quad \boxed{\text{CI}}$$

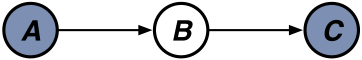
Example



Suppose that A and C are observed and B is hidden.

- Inference

$$
\begin{aligned}
P(B=b|A=a, C=c) &= \frac{P(C=c|B=b, A=a)\, P(B=b|A=a)}{P(C=c|A=a)} && \boxed{\text{BR}} \\
&= \frac{P(C=c|B=b)\, P(B=b|A=a)}{P(C=c|A=a)} && \boxed{\text{CI}} \\
&= \frac{P(C=c|B=b)\, P(B=b|A=a)}{\sum_{b'} P(C=c|B=b')\, P(B=b'|A=a)} && \boxed{\text{normalized}}
\end{aligned}
$$



Suppose that A and C are observed and B is hidden.

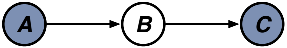
- Inference

$$\begin{aligned}
P(B=b|A=a, C=c) &= \frac{P(C=c|B=b, A=a)\, P(B=b|A=a)}{P(C=c|A=a)} && \boxed{\text{BR}} \\
&= \frac{P(C=c|B=b)\, P(B=b|A=a)}{P(C=c|A=a)} && \boxed{\text{CI}} \\
&= \frac{P(C=c|B=b)\, P(B=b|A=a)}{\sum_{b'} P(C=c|B=b')\, P(B=b'|A=a)} && \boxed{\text{normalized}}
\end{aligned}$$

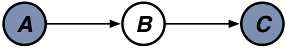
This is the only non-trivial posterior probability that we'll need for the EM updates in this example.

EM update for $P(C|B)$

EM update for $P(C|B)$

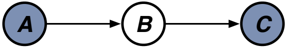


EM update for $P(C|B)$



- General form

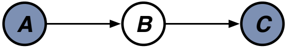
EM update for $P(C|B)$



- General form

$$P(X_i = x | \mathrm{pa}_i = \pi) \longleftarrow \frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi | V_t = v_t)}$$

EM update for $P(C|B)$

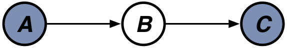


- General form

$$P(X_i = x | \mathrm{pa}_i = \pi) \longleftarrow \frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi | V_t = v_t)}$$

- Update for this CPT

EM update for $P(C|B)$



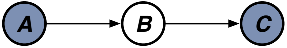
- General form

$$P(X_i = x | \mathrm{pa}_i = \pi) \longleftarrow \frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi | V_t = v_t)}$$

- Update for this CPT

$$P(C = c | B = b)$$

EM update for $P(C|B)$



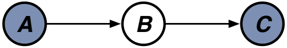
- General form

$$P(X_i = x | \mathrm{pa}_i = \pi) \longleftarrow \frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi | V_t = v_t)}$$

- Update for this CPT

$$P(C = c | B = b) \quad \longleftarrow \quad \frac{\sum_t P(C = c, B = b | A = a_t, C = c_t)}{}$$

EM update for $P(C|B)$



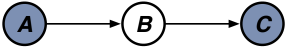
- General form

$$P(X_i\!=\!x|\mathrm{pa}_i\!=\!\pi) \;\longleftarrow\; \frac{\sum_t P(X_i\!=\!x, \mathrm{pa}_i\!=\!\pi|V_t\!=\!v_t)}{\sum_t P(\mathrm{pa}_i\!=\!\pi|V_t\!=\!v_t)}$$

- Update for this CPT

$$P(C\!=\!c|B\!=\!b) \quad\longleftarrow\quad \frac{\sum_t P(C\!=\!c, B\!=\!b|A\!=\!a_t, C\!=\!c_t)}{\sum_t P(B\!=\!b|A\!=\!a_t, C\!=\!c_t)}$$

EM update for $P(C|B)$



- General form

$$P(X_i\!=\!x|\mathrm{pa}_i\!=\!\pi) \;\longleftarrow\; \frac{\sum_t P(X_i\!=\!x, \mathrm{pa}_i\!=\!\pi|V_t\!=\!v_t)}{\sum_t P(\mathrm{pa}_i\!=\!\pi|V_t\!=\!v_t)}$$

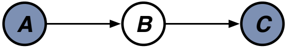
- Update for this CPT

$$P(C\!=\!c|B\!=\!b) \quad\longleftarrow\quad \frac{\sum_t P(C\!=\!c, B\!=\!b|A\!=\!a_t, C\!=\!c_t)}{\sum_t P(B\!=\!b|A\!=\!a_t, C\!=\!c_t)}$$

Simplify:

$$P(C\!=\!c|B\!=\!b) \quad\longleftarrow$$

EM update for $P(C|B)$



- General form

$$P(X_i = x | \mathrm{pa}_i = \pi) \longleftarrow \frac{\sum_t P(X_i = x, \mathrm{pa}_i = \pi | V_t = v_t)}{\sum_t P(\mathrm{pa}_i = \pi | V_t = v_t)}$$

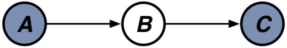
- Update for this CPT

$$P(C = c | B = b) \quad \longleftarrow \quad \frac{\sum_t P(C = c, B = b | A = a_t, C = c_t)}{\sum_t P(B = b | A = a_t, C = c_t)}$$

Simplify:

$$P(C = c | B = b) \quad \longleftarrow \quad \frac{\sum_t I(c, c_t)\, P(B = b | A = a_t, C = c_t)}{}$$

EM update for $P(C|B)$



- General form

$$P(X_i\!=\!x|\mathrm{pa}_i\!=\!\pi) \longleftarrow \frac{\sum_t P(X_i\!=\!x, \mathrm{pa}_i\!=\!\pi|V_t\!=\!v_t)}{\sum_t P(\mathrm{pa}_i\!=\!\pi|V_t\!=\!v_t)}$$

- Update for this CPT

$$P(C\!=\!c|B\!=\!b) \longleftarrow \frac{\sum_t P(C\!=\!c, B\!=\!b|A\!=\!a_t, C\!=\!c_t)}{\sum_t P(B\!=\!b|A\!=\!a_t, C\!=\!c_t)}$$

Simplify:

$$P(C\!=\!c|B\!=\!b) \longleftarrow \frac{\sum_t I(c, c_t)\, P(B\!=\!b|A\!=\!a_t, C\!=\!c_t)}{\sum_t P(B\!=\!b|A\!=\!a_t, C\!=\!c_t)}$$

Summary of EM algorithm

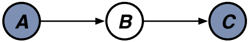
Summary of EM algorithm

- **E-step** (Inference)

Summary of EM algorithm

- **E-step** (Inference)

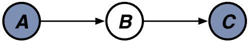
$$P(b|a_t, c_t) = \frac{P(c_t|b)\, P(b|a_t)}{\sum_{b'} P(c_t|b')\, P(b'|a_t)}$$



Summary of EM algorithm

- **E-step** (Inference)

$$P(b|a_t, c_t) = \frac{P(c_t|b)\,P(b|a_t)}{\sum_{b'} P(c_t|b')\,P(b'|a_t)}$$

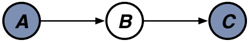


- **M-step** (Learning)

Summary of EM algorithm

- **E-step** (Inference)

$$P(b|a_t, c_t) = \frac{P(c_t|b)\,P(b|a_t)}{\sum_{b'} P(c_t|b')\,P(b'|a_t)}$$



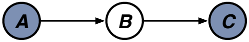
- **M-step** (Learning)

$$P(a) = \frac{1}{T}\,\mathrm{count}(A\!=\!a)$$

Summary of EM algorithm

- **E-step** (Inference)

$$P(b|a_t, c_t) = \frac{P(c_t|b)\, P(b|a_t)}{\sum_{b'} P(c_t|b')\, P(b'|a_t)}$$



- **M-step** (Learning)

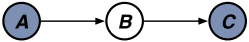
$$P(a) = \frac{1}{T}\,\mathrm{count}(A{=}a)$$

$$P(b|a) \longleftarrow \frac{\sum_t I(a, a_t)\, P(b|a_t, c_t)}{\sum_t I(a, a_t)}$$

Summary of EM algorithm

- **E-step** (Inference)

$$P(b|a_t, c_t) \;=\; \frac{P(c_t|b)\,P(b|a_t)}{\sum_{b'} P(c_t|b')\,P(b'|a_t)}$$



- **M-step** (Learning)

$$P(a) \;=\; \frac{1}{T}\,\mathrm{count}(A{=}a)$$

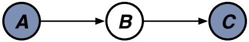
$$P(b|a) \;\longleftarrow\; \frac{\sum_t I(a, a_t)\, P(b|a_t, c_t)}{\sum_t I(a, a_t)}$$

$$P(c|b) \;\longleftarrow\; \frac{\sum_t I(c, c_t)\, P(b|a_t, c_t)}{\sum_t P(b|a_t, c_t)}$$

Summary of EM algorithm

- **E-step** (Inference)

$$P(b|a_t, c_t) = \frac{P(c_t|b)\, P(b|a_t)}{\sum_{b'} P(c_t|b')\, P(b'|a_t)}$$



- **M-step** (Learning)

$$P(a) = \frac{1}{T}\,\mathrm{count}(A = a)$$

$$P(b|a) \longleftarrow \frac{\sum_t I(a, a_t)\, P(b|a_t, c_t)}{\sum_t I(a, a_t)}$$

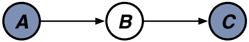
$$P(c|b) \longleftarrow \frac{\sum_t I(c, c_t)\, P(b|a_t, c_t)}{\sum_t P(b|a_t, c_t)}$$

- **Convergence**

Summary of EM algorithm

- **E-step** (Inference)

$$P(b|a_t, c_t) = \frac{P(c_t|b)\, P(b|a_t)}{\sum_{b'} P(c_t|b')\, P(b'|a_t)}$$



- **M-step** (Learning)

$$P(a) = \frac{1}{T}\,\mathrm{count}(A{=}a)$$

$$P(b|a) \longleftarrow \frac{\sum_t I(a, a_t)\, P(b|a_t, c_t)}{\sum_t I(a, a_t)}$$

$$P(c|b) \longleftarrow \frac{\sum_t I(c, c_t)\, P(b|a_t, c_t)}{\sum_t P(b|a_t, c_t)}$$

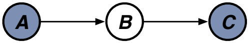
- **Convergence**

  There are no learning rates to tune.

Summary of EM algorithm

- **E-step** (Inference)

$$P(b|a_t, c_t) \;=\; \frac{P(c_t|b)\,P(b|a_t)}{\sum_{b'} P(c_t|b')\,P(b'|a_t)}$$



- **M-step** (Learning)

$$P(a) \;=\; \frac{1}{T}\,\mathrm{count}(A{=}a)$$

$$P(b|a) \;\longleftarrow\; \frac{\sum_t I(a, a_t)\,P(b|a_t, c_t)}{\sum_t I(a, a_t)}$$

$$P(c|b) \;\longleftarrow\; \frac{\sum_t I(c, c_t)\,P(b|a_t, c_t)}{\sum_t P(b|a_t, c_t)}$$

- **Convergence**

There are no learning rates to tune.
Each update increases the incomplete data log-likelihood:

$$\mathcal{L} \;=\; \sum_t \log \sum_b P(a_t)\,P(b|a_t)\,P(c_t|b)$$

Check In

Q. How much of EM did you understand?

A. (Nearly) All of it 25% ← 15

B. Some of it, but I have some doubts 45% ← 40

C. Maybe a little, but I'm pretty confused 27% ← 35

D. Almost none of it; I'm totally lost

That's all folks!