6.825 Techniques in Artificial Intelligence

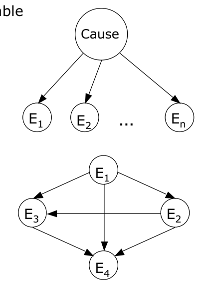
Learning With Hidden Variables

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

Hidden variables

Cause is unobservable

O(n) parameters



$O(2^n)$ parameters

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

Missing Data

A	B
1	1
1	1
0	0
0	0
0	0
0	H
0	1
1	0

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

Ignore it

A	B
1	1
1	1
0	0
0	0
0	0
0	H
0	1
1	0

Estimated Parameters

	~A	A
~B	3/7	1/7
B	1/7	2/7

	~A	A
~B	.429	.143
B	.143	.285

$$\begin{aligned}\log \Pr(D|M) &= \log(\Pr(D, H=0 \mid M) + \Pr(D, H=1 \mid M)) \\ &= 3\log .429 + 2\log .143 + 2\log .285 + \log(.429 + .143) \\ &= -9.498\end{aligned}$$

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

A	B
1	1
1	1
0	0
0	0
0	0
0	0
0	1
1	0

Estimated Parameters

	~A	A
~B	4/8	1/8
B	1/8	2/8

	~A	A
~B	.5	.125
B	.125	.25

$$\begin{aligned}\log \Pr(D|M) &= \log(\Pr(D, H=0 \mid M) + \Pr(D, H=1 \mid M) \\ &= 3\log .5 + 2\log .125 + 2\log .25 + \log(.5 + .125) \\ &= -9.481\end{aligned}$$

Fill in With Distribution

A	B
1	1
1	1
0	0
0	0
0	0
0	H
0	1
1	0

Guess a distribution over A,B and compute a distribution over H

θ_0

	~A	A
~B	.25	.25
B	.25	.25

$$\begin{aligned}\Pr(H|D,\theta_0) &= \Pr(H \mid D^6,\theta_0)\\ &= \Pr(B \mid \neg A,\theta_0)\\ &= \Pr(\neg A,B \mid \theta_0)/\Pr(\neg A \mid \theta_0)\\ &= .25/0.5\\ &= 0.5\end{aligned}$$

Fill in With Distribution

A	B
1	1
1	1
0	0
0	0
0	0
0	0, 0.5 1, 0.5
0	1
1	0

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

Maximum likelihood estimation using *expected counts*

θ_1

	~A	A
~B	3.5/8	1/8
B	1.5/8	2/8

	~A	A
~B	.4375	.125
B	.1875	.25

Fill in With Distribution

A	B
1	1
1	1
0	0
0	0
0	0
0	
0	1
1	0

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

θ_1

	~A	A
~B	.4375	.125
B	.1875	.25

$$\begin{aligned}\Pr(H|D,\theta_1) &= \Pr(\neg A,B \mid \theta_1)/\Pr(\neg A \mid \theta_1)\\ &= .1875/.625\\ &= 0.3\end{aligned}$$

Fill in With Distribution

A	B
1	1
1	1
0	0
0	0
0	0
0	0, 0.7 1, 0.3
0	1
1	0

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

θ_2

	~A	A
~B	3.7/8	1/8
B	1.3/8	2/8

	~A	A
~B	.4625	.125
B	.1625	.25

Fill in With Distribution

A	B
1	1
1	1
0	0
0	0
0	0
0	
0	1
1	0

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

θ_2

	~A	A
~B	.4625	.125
B	.1625	.25

$$\begin{aligned}\Pr(H|D,\theta_2) &= \Pr(\neg A,B \mid \theta_2)/\Pr(\neg A \mid \theta_2)\\ &= .1625/.625\\ &= 0.26\end{aligned}$$

Fill in With Distribution

A	B
1	1
1	1
0	0
0	0
0	0
0	0, 0.74 1, 0.26
0	1
1	0

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

θ_3

	~A	A
~B	3.74/8	1/8
B	1.26/8	2/8

	~A	A
~B	.4675	.125
B	.1575	.25

Increasing Log-Likelihood

θ_0

	~A	A
~B	.25	.25
B	.25	.25

$\log \Pr(D \mid \theta_0) = -10.3972$

ignore: -9.498
best val: -9.481

θ_1

	~A	A
~B	.4375	.125
B	.1875	.25

$\log \Pr(D \mid \theta_1) = -9.4760$

θ_2

	~A	A
~B	.4625	.125
B	.1625	.25

$\log \Pr(D \mid \theta_2) = -9.4524$

θ_3

	~A	A
~B	.4675	.125
B	.1575	.25

$\log \Pr(D \mid \theta_3) = -9.4514$

Deriving the EM Algorithm

- Want to find θ to maximize $\Pr(D \mid \theta)$
- Instead, find θ, $\tilde{P}$ to maximize

$$g(\theta, \tilde{P}) = \sum_H \tilde{P}(H) \log(\Pr(D, H \mid \theta) / \tilde{P}(H))$$
$$= E_{\tilde{P}} \log \Pr(D, H \mid \theta) - \log \tilde{P}(H)$$

- Alternate between
  - holding θ fixed and optimizing $\tilde{P}$
  - holding $\tilde{P}$ fixed and optimizing θ
- g has same local and global optima as $\Pr(D \mid \theta)$

EM Algorithm

- Pick initial θ_0
- Loop until apparently converged
  - $\tilde{P}_{t+1}(H) = \Pr(H \mid D, \theta_t)$
  - $\theta_{t+1} = \arg\max_{\theta} E_{\tilde{P}_{t+1}} \log \Pr(D, H \mid \theta)$
- 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

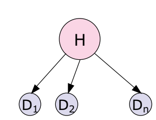
$$\tilde{}_{t+1}(H) = \Pr(H \mid D, \theta_t)$$
$$= \underset{m}{} \Pr(H^m \mid D^m, \theta_t)$$

- 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

Bayes net inference

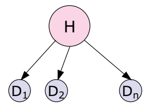
D_1	D_2	..	D_n	$\Pr(H^m \mid D^m, \theta_t)$
1	1		0	.9
0	1		0	.2
0	0		1	.1
1	0		1	.6
1	1		1	.2
1	1		1	.5
0	1		0	.3
0	0		0	.7
1	1		0	.2



EM for BN: Simple Case

D_1	D_2	...	D_n	$\Pr(H^m \mid D^m, \theta_t)$
1	1		0	.9
0	1		0	.2
0	0		1	.1
1	0		1	.6
1	1		1	.2
1	1		1	.5
0	1		0	.3
0	0		0	.7
1	1		0	.2

Bayes net inference



$$E\#(H) = \sum_m \Pr(H^m \mid D^m, \theta_t) = 3.7$$

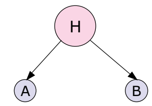
$$E\#(H \wedge D_2) = \sum_m \Pr(H^m \mid D^m, \theta_t) I(D_2^m) = .9 + .2 + .2 + .5 + .3 + .2 = 2.3$$

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

Re-estimate θ

EM for BN: Worked Example

A	B	#	$\Pr(H^m \mid D^m, \theta_t)$
0	0	6	
0	1	1	
1	0	1	
1	1	4	



$\theta_1 = \Pr(H)$

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

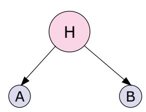
$\theta_3 = \Pr(A \mid \neg H)$

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

$\theta_5 = \Pr(B \mid \neg H)$

EM for BN: Initial Model

A	B	#	$\Pr(H^m \mid D^m, \theta_t)$
0	0	6	
0	1	1	
1	0	1	
1	1	4	



$\Pr(H) = 0.4$

$\Pr(A|H) = 0.55$

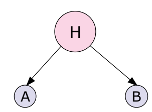
$\Pr(A|\neg H) = 0.61$

$\Pr(B|H) = 0.43$

$\Pr(B|\neg H) = 0.52$

Iteration 1: Fill in data

A	B	#	$\Pr(H^m \mid D^m, \theta_t)$
0	0	6	.48
0	1	1	.39
1	0	1	.42
1	1	4	.33



$\Pr(H) = 0.4$

$\Pr(A|H) = 0.55$

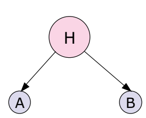
$\Pr(A|\neg H) = 0.61$

$\Pr(B|H) = 0.43$

$\Pr(B|\neg H) = 0.52$

Iteration 1: Re-estimate Params

A	B	#	$\Pr(H^m \mid D^m, \theta_t)$
0	0	6	.48
0	1	1	.39
1	0	1	.42
1	1	4	.33



$\Pr(H) = 0.42$

$\Pr(A|H) = 0.35$

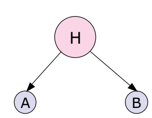
$\Pr(A|\neg H) = 0.46$

$\Pr(B|H) = 0.34$

$\Pr(B|\neg H) = 0.47$

Iteration 2: Fill in Data

A	B	#	$\Pr(H^m \mid D^m, \theta_t)$
0	0	6	.52
0	1	1	.39
1	0	1	.39
1	1	4	.28



$\Pr(H) = 0.42$

$\Pr(A|H) = 0.35$

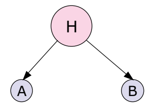
$\Pr(A|\neg H) = 0.46$

$\Pr(B|H) = 0.34$

$\Pr(B|\neg H) = 0.47$

Iteration 2: Re-estimate params

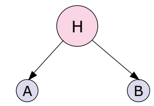
A	B	#	$\Pr(H^m \mid D^m, \theta_t)$
0	0	6	.52
0	1	1	.39
1	0	1	.28
1	1	4	.28



$$\Pr(H) = 0.42$$
$$\Pr(A|H) = 0.31$$
$$\Pr(A|\neg H) = 0.50$$
$$\Pr(B|H) = 0.30$$
$$\Pr(B|\neg H) = 0.50$$

Iteration 5

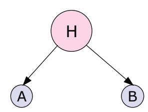
A	B	#	$\Pr(H^m \mid D^m, \theta_t)$
0	0	6	.79
0	1	1	.31
1	0	1	.31
1	1	4	.05



$$\Pr(H) = 0.46$$
$$\Pr(A|H) = 0.09$$
$$\Pr(A|\neg H) = 0.69$$
$$\Pr(B|H) = 0.09$$
$$\Pr(B|\neg H) = 0.69$$

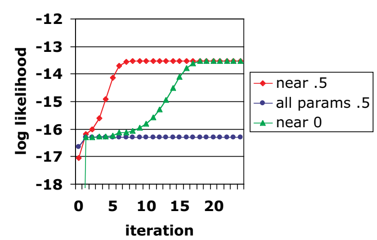
Iteration 10

A	B	#	$\Pr(H^m \mid D^m, \theta_t)$
0	0	6	.971
0	1	1	.183
1	0	1	.183
1	1	4	.001



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

Increasing Log Likelihood



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