

**Expectation
Maximalization**

The E.M. Algorithm

- We'll get back to unsupervised learning soon.
- But now we'll look at an even simpler case with hidden information.
- The EM algorithm
 - ❑ An excellent way of doing unsupervised clustering.
 - ❑ Many, many other uses, including inference of Hidden Markov Models.

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

	$\sim A$	A
$\sim B$	3/7	1/7
B	1/7	2/7

	$\sim A$	A
$\sim B$.429	.143
B	.143	.285

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

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

	$\sim A$	A
$\sim B$	4/8	1/8
B	1/8	2/8

	$\sim A$	A
$\sim B$.5	.125
B	.125	.25

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

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

$$\theta_0$$

	$\sim A$	A
$\sim B$.25	.25
B	.25	.25

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

	$\sim A$	A
$\sim B$.25	.25
B	.25	.25

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

Fill in With Distribution

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

	$\sim A$	A
$\sim B$	3.5/8	1/8
B	1.5/8	2/8

	$\sim A$	A
$\sim B$.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

$$\theta_1$$

	$\sim A$	A
$\sim B$.4375	.125
B	.1875	.25

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

Fill in With Distribution

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

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

θ_2

	$\sim A$	A
$\sim B$	3.7/8	1/8
B	1.3/8	2/8

	$\sim A$	A
$\sim B$.4625	.125
B	.1625	.25

Fill in With Distribution

A	B
1	1
1	1
0	0
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

	$\sim A$	A
$\sim B$.4625	.125
B	.1625	.25

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

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

	$\sim A$	A
$\sim B$	3.74/8	1/8
B	1.26/8	2/8

	$\sim A$	A
$\sim B$.4675	.125
B	.1575	.25

Increasing Log-Likelihood

 θ_0

	$\sim A$	A
$\sim B$.25	.25
B	.25	.25

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

 θ_1

	$\sim A$	A
$\sim B$.4375	.125
B	.1875	.25

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

 θ_2

	$\sim A$	A
$\sim B$.4625	.125
B	.1625	.25

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

 θ_3

	$\sim A$	A
$\sim B$.4675	.125
B	.1575	.25

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

Increasing Log-Likelihood

 θ_0

	$\sim A$	A
$\sim B$.25	.25
B	.25	.25

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

ignore: -9.498

best val: -9.481

 θ_1

	$\sim A$	A
$\sim B$.4375	.125
B	.1875	.25

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

 θ_2

	$\sim A$	A
$\sim B$.4625	.125
B	.1625	.25

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

 θ_3

	$\sim A$	A
$\sim B$.4675	.125
B	.1575	.25

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

Another (Simple) Example

Let events be “grades in a class”

$$w_1 = \text{Gets an A} \quad P(A) = \frac{1}{2}$$

$$w_2 = \text{Gets a B} \quad P(B) = \mu$$

$$w_3 = \text{Gets a C} \quad P(C) = 2\mu$$

$$w_4 = \text{Gets a D} \quad P(D) = \frac{1}{2} - 3\mu$$

(Note $0 \leq \mu \leq 1/6$)

Assume we want to estimate μ from data. In a given class there were

a A's

b B's

c C's

d D's

What's the maximum likelihood estimate of μ given a,b,c,d
?

Simple Example - contd

Let events be “grades in a class”

$w_1 =$ Gets an A	$P(A) = \frac{1}{2}$
$w_2 =$ Gets a B	$P(B) = \mu$
$w_3 =$ Gets a C	$P(C) = 2\mu$
$w_4 =$ Gets a D	$P(D) = \frac{1}{2} - 3\mu$

(Note $0 \leq \mu \leq 1/6$)

Assume we want to estimate μ from data. In a given class there were

- a A's
- b B's
- c C's
- d D's

What's the maximum likelihood estimate of μ given a,b,c,d ?

Trivial Statistics

$$P(A) = \frac{1}{2} \quad P(B) = \mu \quad P(C) = 2\mu \quad P(D) = \frac{1}{2} - 3\mu$$

$$P(a, b, c, d \mid \mu) = K \left(\frac{1}{2}\right)^a (\mu)^b (2\mu)^c \left(\frac{1}{2} - 3\mu\right)^d$$

$$\log P(a, b, c, d \mid \mu) = \log K + a \log \frac{1}{2} + b \log \mu + c \log 2\mu + d \log \left(\frac{1}{2} - 3\mu\right)$$

$$\text{FOR MAX LIKE } \mu, \text{ SET } \frac{\partial \text{LogP}}{\partial \mu} = 0$$

$$\frac{\partial \text{LogP}}{\partial \mu} = \frac{b}{\mu} + \frac{2c}{2\mu} - \frac{3d}{1/2 - 3\mu} = 0$$

$$\text{Gives max like } \mu = \frac{b + c}{6(b + c + d)}$$

So if class got

A	B	C	D
14	6	9	10

$$\text{Max like } \mu = \frac{1}{10}$$

Same Problem with Hidden Information

Someone tells us that

Number of High grades (A's + B's) = h

Number of C's = c

Number of D's = d

What is the max. like estimate of μ now?

REMEMBER

$$P(A) = \frac{1}{2}$$

$$P(B) = \mu$$

$$P(C) = 2\mu$$

$$P(D) = \frac{1}{2} - 3\mu$$

Same Problem with Hidden Information

Someone tells us that

Number of High grades (A's + B's) = h

Number of C's = c

Number of D's = d

REMEMBER

$$P(A) = \frac{1}{2}$$

$$P(B) = \mu$$

$$P(C) = 2\mu$$

$$P(D) = \frac{1}{2} - 3\mu$$

What is the max. like estimate of μ now?

We can answer this question circularly:

EXPECTATION

If we know the value of μ we could compute the expected value of a and b

Since the ratio $a:b$ should be the same as the ratio $\frac{1}{2} : \mu$

$$a = \frac{\frac{1}{2}}{\frac{1}{2} + \mu} h \quad b = \frac{\mu}{\frac{1}{2} + \mu} h$$

MAXIMIZATION

If we know the expected values of a and b we could compute the maximum likelihood value of μ

$$\mu = \frac{b + c}{6(b + c + d)}$$

E.M. for our Trivial Problem

We begin with a guess for μ

We iterate between EXPECTATION and MAXIMALIZATION to improve our estimates of μ and a and b .

REMEMBER

$$P(A) = \frac{1}{2}$$

$$P(B) = \mu$$

$$P(C) = 2\mu$$

$$P(D) = \frac{1}{2} - 3\mu$$

Define $\mu(t)$ the estimate of μ on the t 'th iteration

$b(t)$ the estimate of b on t 'th iteration

$\mu(0)$ = initial guess

$$b(t) = \frac{\mu(t)h}{\frac{1}{2} + \mu(t)} = E[b \mid \mu(t)]$$

$$\mu(t+1) = \frac{b(t) + c}{6(b(t) + c + d)}$$

= max like est of μ given $b(t)$



E-step



M-step

Continue iterating until converged.

Good news: Converging to local optimum is assured.

Bad news: "local" optimum.

E.M. Convergence

- Convergence proof based on fact that $\text{Prob}(\text{data} \mid \mu)$ must increase or remain same between each iteration [NOT OBVIOUS]
 - But it can never exceed 1 [OBVIOUS]
- So it must therefore converge [OBVIOUS]

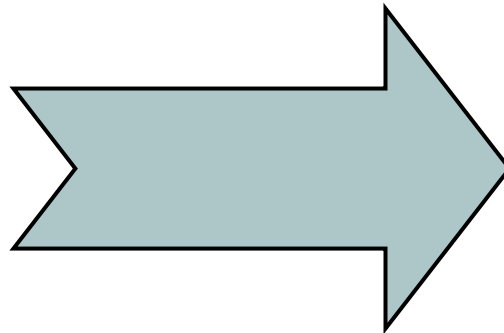
In our example,
suppose we had

$$h = 20$$

$$c = 10$$

$$d = 10$$

$$\mu(0) = 0$$



t	$\mu(t)$	b(t)
0	0	0
1	0.0833	2.857
2	0.0937	3.158
3	0.0947	3.185
4	0.0948	3.187
5	0.0948	3.187
6	0.0948	3.187

Convergence is generally linear: error decreases by a constant factor each time step.

Recap

Maximum Likelihood Estimation

- We have data points X_1, X_2, \dots, X_n drawn from some (finite or countable) set \mathcal{X}
- We have a parameter vector Θ
- We have a parameter space Ω
- We have a distribution $P(X | \Theta)$ for any $\Theta \in \Omega$, such that

$$\sum_{X \in \mathcal{X}} P(X | \Theta) = 1 \text{ and } P(X | \Theta) \geq 0 \text{ for all } X$$

- We assume that our data points X_1, X_2, \dots, X_n are drawn at random (independently, identically distributed) from a distribution $P(X | \Theta^*)$ for some $\Theta^* \in \Omega$

Log-Likelihood

- We have data points X_1, X_2, \dots, X_n drawn from some (finite or countable) set \mathcal{X}
- We have a parameter vector Θ , and a parameter space Ω
- We have a distribution $P(X | \Theta)$ for any $\Theta \in \Omega$

- The likelihood is

$$\text{Likelihood}(\Theta) = P(X_1, X_2, \dots, X_n | \Theta) = \prod_{i=1}^n P(X_i | \Theta)$$

- The log-likelihood is

$$L(\Theta) = \log \text{Likelihood}(\Theta) = \sum_{i=1}^n \log P(X_i | \Theta)$$

MLE: $\Theta_{ML} = \operatorname{argmax}_{\Theta \in \Omega} L(\Theta) = \operatorname{argmax}_{\Theta \in \Omega} \sum_i \log P(X_i | \Theta)$

Models with Hidden Variables: Formalization

- Now say we have two sets \mathcal{X} and \mathcal{Y} , and a joint distribution $P(X, Y \mid \Theta)$

- If we had **fully observed data**, (X_i, Y_i) pairs, then

$$L(\Theta) = \sum_i \log P(X_i, Y_i \mid \Theta)$$

- If we have **partially observed data**, X_i examples, then

$$\begin{aligned} L(\Theta) &= \sum_i \log P(X_i \mid \Theta) \\ &= \sum_i \log \sum_{Y \in \mathcal{Y}} P(X_i, Y \mid \Theta) \end{aligned}$$

- The **EM (Expectation Maximization) algorithm** is a method for finding

$$\Theta_{ML} = \operatorname{argmax}_{\Theta} \sum_i \log \sum_{Y \in \mathcal{Y}} P(X_i, Y \mid \Theta)$$

EM Formalization

- Let X be all *observed* variable values (over all examples)
- Let Z be all *unobserved* variable values
- Can't calculate MLE:

$$\theta \leftarrow \arg \max_{\theta} \log P(X, Z|\theta)$$

- EM seeks estimate:

$$\theta \leftarrow \arg \max_{\theta} E_{Z|X, \theta}[\log P(X, Z|\theta)]$$

Deriving the EM Algorithm

- Want to find θ to maximize $\Pr(D | \theta)$

- Instead, find θ, \tilde{P} to maximize

$$\begin{aligned}g(\theta, \tilde{P}) &= \sum_H \tilde{P}(H) \log(\Pr(D, H | \theta) / \tilde{P}(H)) \\ &= E_{\tilde{P}} \log \Pr(D, H | \theta) - \log \tilde{P}(H)\end{aligned}$$

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

EM Algorithm

- Pick initial θ_0
- Loop until apparently converged
 - $\tilde{P}_{t+1}(H) = \Pr(H | D, \theta_t)$
 - $\theta_{t+1} = \arg \max_{\theta} E_{\tilde{P}_{t+1}} \log \Pr(D, H | \theta)$
- Monotonically increasing likelihood
- Convergence is hard to determine due to plateaus
- Problems with local optima

MLE for CPTs

- Each conditional probability table θ_i part of our parameters
- Given table, have pdf

$$p(x | \theta) = \prod_{i=1}^n p(x_i | \pi_i, \theta_i)$$

- Have M variables:

$$X_U = \{x_1, \dots, x_M\}$$

- Have N x M dataset:

$$\mathcal{D} = \{X_{U,1}, \dots, X_{U,N}\}$$

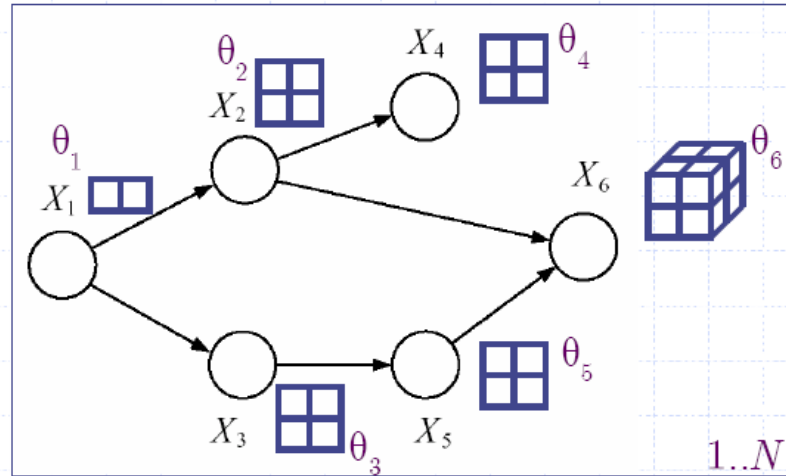
- Maximum likelihood:

$$\theta^* = \arg \max_{\theta} \log p(\mathcal{D} | \theta)$$

$$= \arg \max_{\theta} \sum_{n=1}^N \log p(X_{U,n} | \theta)$$

$$= \arg \max_{\theta} \sum_{n=1}^N \log \prod_{i=1}^M p(x_{i,n} | \pi_{i,n} \theta_i)$$

$$= \arg \max_{\theta} \sum_{n=1}^N \sum_{i=1}^M \log p(x_{i,n} | \pi_{i,n} \theta_i)$$

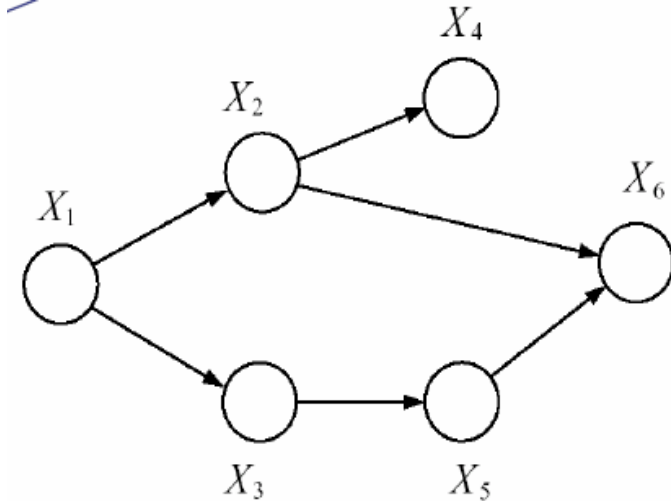


each θ_i appears independently,
can do ML for each CPT alone!
efficient storage
+
efficient learning

parents of child $i = pa_i = \pi_i$

MLE for CPTs

PATIENT	FLU	FEVER	SINUS	TEMP	SWELL	HEAD
1	Y	Y	N	L	Y	Y
2	N	N	N	M	N	Y
3	Y	N	Y	H	Y	N
4	Y	N	Y	M	N	N



$X_1 = \text{FLU}$, $X_2 = \text{FEVER}$, $X_3 = \text{SINUS}$, $X_4 = \text{TEMP}$,
 $X_5 = \text{SWELL}$, $X_6 = \text{HEAD}$

Let: $\theta(x_i, \pi_i) = p(x_i | \pi_i, \theta_i)$

Note: $\sum_{x_i} \theta(x_i, \pi_i) = 1$

MLE: $\theta(x_i, \pi_i) = \frac{m(x_i, \pi_i)}{m(\pi_i)}$

	$x_1 = 0$	$x_1 = 1$	
$x_3 = 0$	1	1	$m(x_3, x_1)$
$x_3 = 1$	0	2	

	$x_1 = 0$	$x_1 = 1$	
	1	3	$m(x_1)$

	$x_1 = 0$	$x_1 = 1$	
	1	1/3	$p(x_3 x_1)$
	0	2/3	

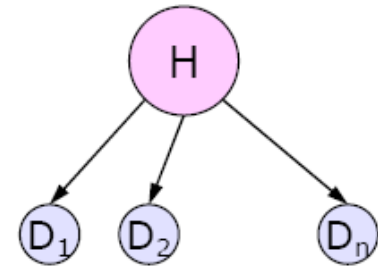
EM for Bayesian Nets

- 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

EM for BN Example

D_1	D_2	...	D_n	$\Pr(H^m D^m, \theta_t)$
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



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

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

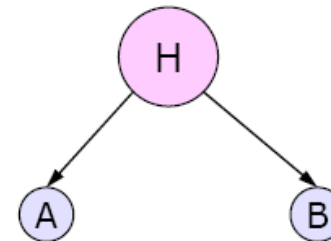
$$\Pr(D_2 | H) \approx 2.3 / 3.7 = .6216$$

Re-estimate

θ

EM for BN: Worked Example

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



$$\theta_1 = \Pr(H)$$

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

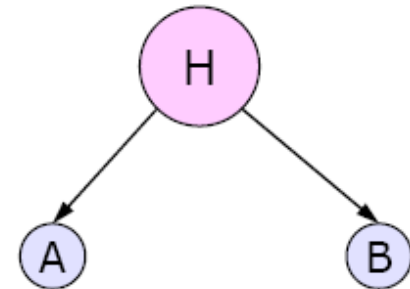
$$\theta_3 = \Pr(A | \neg H)$$

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

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

EM for BN: Initial Model

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



$$\Pr(H) = 0.4$$

$$\Pr(A|H) = 0.55$$

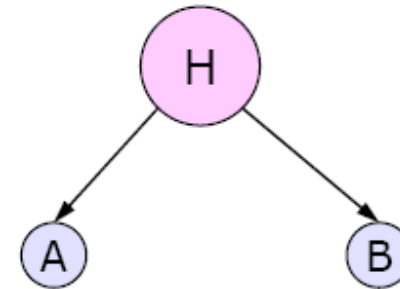
$$\Pr(A|\neg H) = 0.61$$

$$\Pr(B|H) = 0.43$$

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

Iteration 1: Fill in Data

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



$$\Pr(H) = 0.4$$

$$\Pr(A|H) = 0.55$$

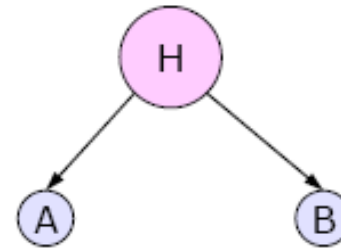
$$\Pr(A|\neg H) = 0.61$$

$$\Pr(B|H) = 0.43$$

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

Iteration 1: Re-estimate Params

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



$$\Pr(H) = 0.42$$

$$\Pr(A|H) = 0.35$$

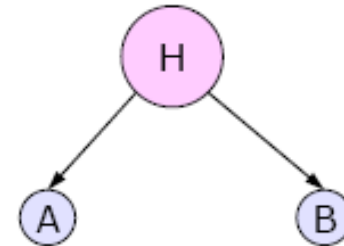
$$\Pr(A|\neg H) = 0.46$$

$$\Pr(B|H) = 0.34$$

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

Iteration 2: Fill in Data

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



$$\Pr(H) = 0.42$$

$$\Pr(A|H) = 0.35$$

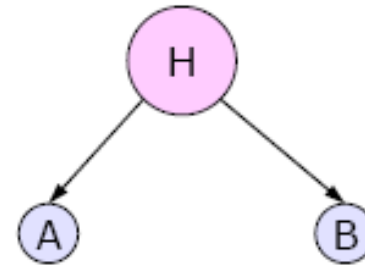
$$\Pr(A|\neg H) = 0.46$$

$$\Pr(B|H) = 0.34$$

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

Iteration 2: Re-estimate Params

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



$$\Pr(H) = 0.42$$

$$\Pr(A|H) = 0.31$$

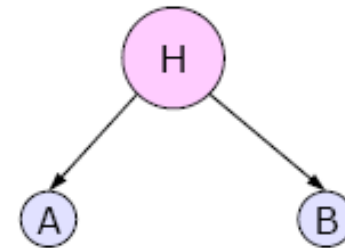
$$\Pr(A|\neg H) = 0.50$$

$$\Pr(B|H) = 0.30$$

$$\Pr(B|\neg H) = 0.50$$

Iteration 5

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



$$\Pr(H) = 0.46$$

$$\Pr(A|H) = 0.09$$

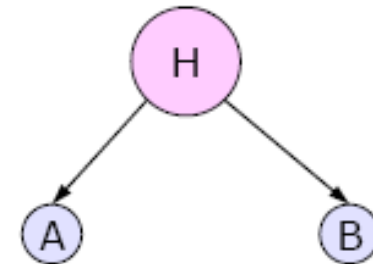
$$\Pr(A|\neg H) = 0.69$$

$$\Pr(B|H) = 0.09$$

$$\Pr(B|\neg H) = 0.69$$

Iteration 10

A	B	#	$\Pr(H^m D^m, \theta)$
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

