Bayesian Networks III
D-separation and EM

Required reading:

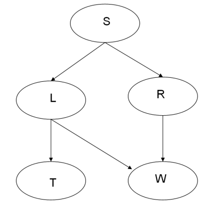
- Bishop online chapter 8, read all of section 8.2

Machine Learning 10-601

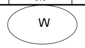
Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

February 23, 2009

Bayesian Networks Definition



Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



A Bayes network represents the joint probability distribution over a collection of random variables

A Bayes network is a directed acyclic graph and a set of CPD's

- Each node denotes a random variable
- Edges denote dependencies
- CPD for each node X_i defines $P(X_i \,/\, Pa(X_i))$
- The joint distribution over all variables is defined as

$$P(X_1 \ldots X_n) = \prod_i P(X_i | Pa(X_i))$$

Pa(X) = immediate parents of X in the graph

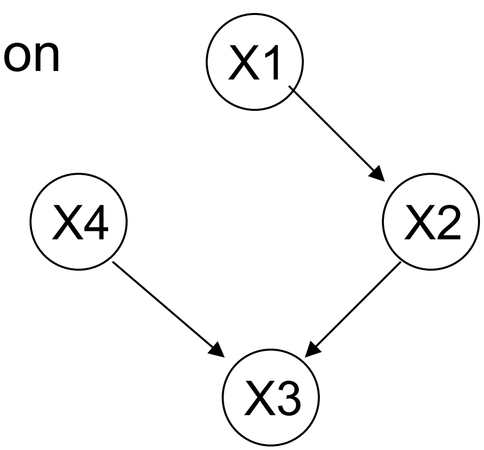
Inference in Bayes Nets

- In general, intractable (NP-complete)
- For certain cases, tractable
  - Assigning probability to fully observed set of variables
  - Or if just one variable unobserved
  - Or for singly connected graphs (ie., no undirected loops)
    - Variable elimination
    - Belief propagation
- For multiply connected graphs
    - Junction tree
    - Loopy belief propagation
- Sometimes use Monte Carlo methods
  - Generate many samples according to the Bayes Net distribution, then count up the results
- Variational methods for tractable approximate solutions

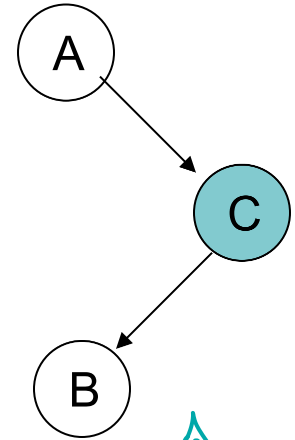
Conditional Independence, Revisited

- We said:
  - Each node is conditionally independent of its non-descendents, given its immediate parents.

- Does this rule give us <u>all</u> of the conditional independence relations implied by the Bayes network?
  - No!
  - E.g., X1 and X4 are conditionally indep given {X2, X3}
  - But X1 and X4 not conditionally indep given X3
  - For this, we need to understand D-separation



Easy Network 1: Head to Tail



prove A cond indep of B given C?

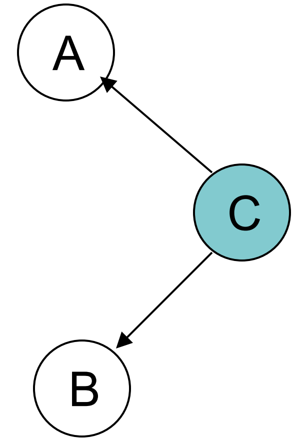
ie., p(a,b|c) = p(a|c) p(b|c)

$$p(ab|c) \equiv \frac{p(abc)}{p(c)} = \frac{p(a)\,p(c|a)\,p(b|c)}{p(c)}$$

$$p(ab|c) = p(a|c) \cdot p(b|c)$$

let's use p(a,b) as shorthand for p(A=a, B=b)

Easy Network 2: Tail to Tail

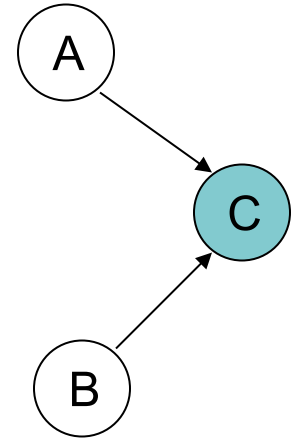


prove A cond indep of B given C?   ie., p(a,b|c) = p(a|c) p(b|c)

$$p(ab|c) = \frac{p(abc)}{p(c)} = \frac{\cancel{p(c)}\, p(a|c)\, p(b|c)}{\cancel{p(c)}}$$

let's use p(a,b) as shorthand for p(A=a, B=b)

Easy Network 3: Head to Head



prove A cond indep of B given C?   ie., p(a,b|c) = p(a|c) p(b|c)

$$p(ab|c) = \frac{\overbrace{p(a)\,p(b)\,p(c|ab)}^{=\,p(abc)}}{p(c)}$$

$$\neq p(a|c)\,p(b|c)$$

but

$$p(ab) = p(a)\,p(b)$$

$$p(a,b) = \sum_i p(a,b,c_i)$$

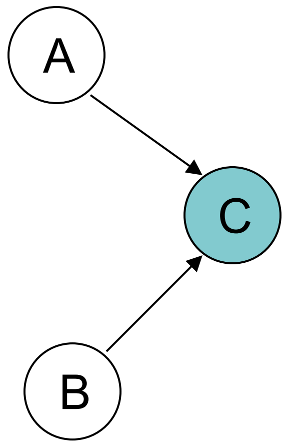
$$= \sum_i p(a)\,p(b)\,\boxed{p(c_i|ab)}$$

$$1 = \sum_i p(c_i|ab)$$

$$p(ab) = p(a)\,p(b)\underbrace{\sum_i p(c_i|ab)}_{=1}$$

let's use p(a,b) as shorthand for p(A=a, B=b)

Easy Network 3: Head to Head



prove A cond indep of B given C?   NO!

Summary:

- p(a,b)=p(a)p(b)
- p(a,b|c) NotEqual p(a|c)p(b|c)

Explaining away.

e.g.,

- A=earthquake
- B=breakIn
- C=motionAlarm

X and Y are conditionally independent given Z, **<u>if and only if</u>** X and Y are D-separated by Z.

[Bishop, 8.2.2]

Suppose we have three sets of random variables: X, Y and Z

X and Y are **<u>D-separated</u>** by Z (and therefore conditionally indep, given Z) iff every path from any variable in X to any variable in Y is **<u>blocked</u>**

A path from variable A to variable B is **blocked** if it includes a node such that either

1. arrows on the path meet either head-to-tail or tail-to-tail at the node and this node is in Z
2. the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in Z

X and Y are **<u>D-separated</u>** by Z (and therefore conditionally indep, given Z) iff every path from any variable in X to any variable in Y is **<u>blocked</u>**

A path from variable A to variable B is **blocked** if it includes a node such that either

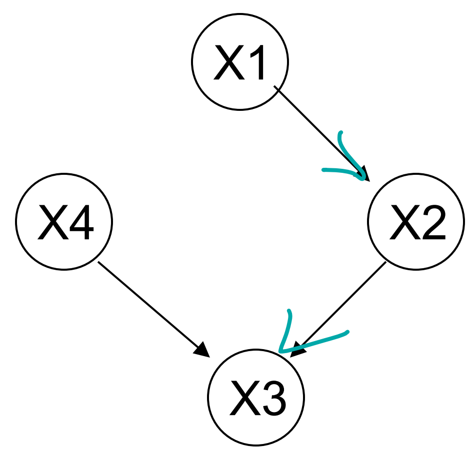
1. arrows on the path meet either head-to-tail or tail-to-tail at the node and this node is in Z
2. the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in Z

X4 ind of X1 given X3? No

X1 indep of X3 given X2? Yes

X3 indep of X1 given X2? Yes

X4 indep of X1 given X2? Yes



X and Y are **<u>D-separated</u>** by Z (and therefore conditionally indep, given Z) iff every path from any variable in X to any variable in Y is **<u>blocked</u>** by Z

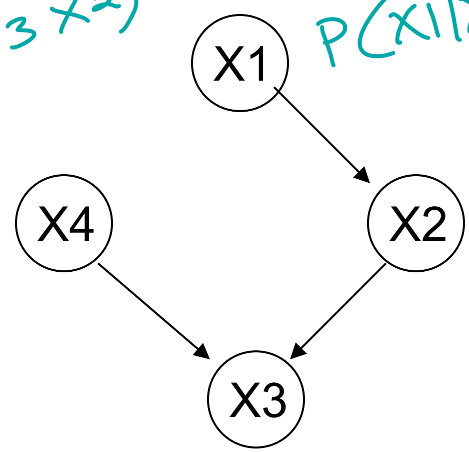
A path from variable A to variable B is **blocked** by Z if it includes a node such that either

1. arrows on the path meet either head-to-tail or tail-to-tail at the node and this node is in Z
2. the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in Z

X4 indep of X1 given X3? No

X4 indep of X1 given {X3, X2}? Y = $P(x_4\, x_1 | x_3\, x_2) = P(x_4 | x_3 x_2) \cdot P(x_1 | x_3 x_2)$

X4 indep of X1 given {}? Y



X and Y are **<u>D-separated</u>** by Z (and therefore conditionally indep, given Z) iff every path from any variable in X to any variable in Y is **<u>blocked</u>**

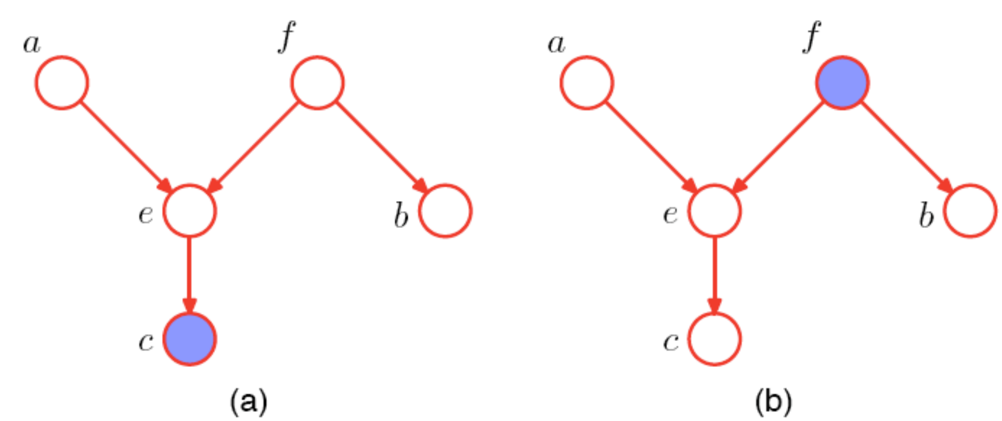
A path from variable A to variable B is **blocked** if it includes a node such that either

1. arrows on the path meet either head-to-tail or tail-to-tail at the node and this node is in Z
2. the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in Z

a indep of b given c?

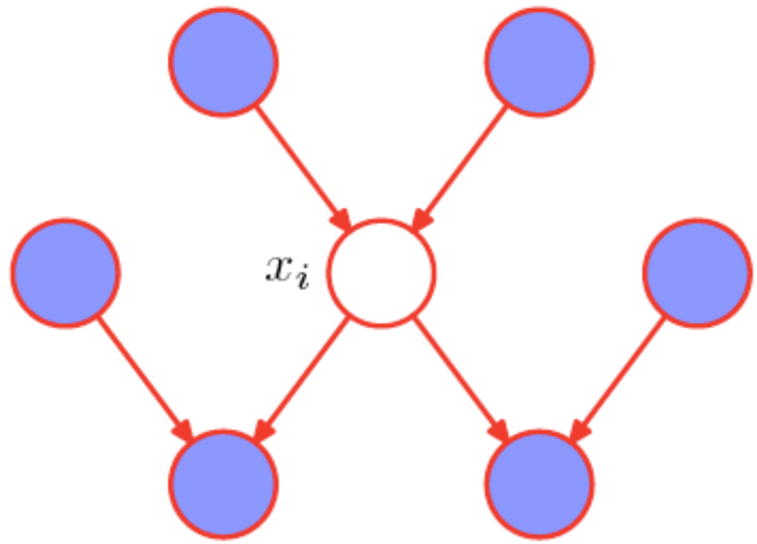
a indep of b given {} ?

a indep of b given f ?



Markov Blanket

The Markov blanket of a node $\mathbf{x}_i$ comprises the set of parents, children and co-parents of the node. It has the property that the conditional distribution of $\mathbf{x}_i$, conditioned on all the remaining variables in the graph, is dependent only on the variables in the Markov blanket.



from [Bishop, 8.2]

Java Bayes Net Applet

http://www.pmr.poli.usp.br/ltd/Software/javabayes/Home/applet.html

by **Fabio Gagliardi Cozman**

What You Should Know

- Bayes nets are convenient representation for encoding dependencies / conditional independence
- BN = Graph plus parameters of CPD's
  - Defines joint distribution over variables
  - Can calculate everything else from that
  - Though inference may be intractable
- Reading conditional independence relations from the graph
  - Each node is cond indep of non-descendents, given only its parents
  - D-separation
  - 'Explaining away'

Learning in Bayes Nets

- Four categories of learning problems
  - Graph structure may be known/unknown
  - Variable values may be observed/unobserved

- Easy case: known graph structure, training data is fully observed, learn CPD parameters

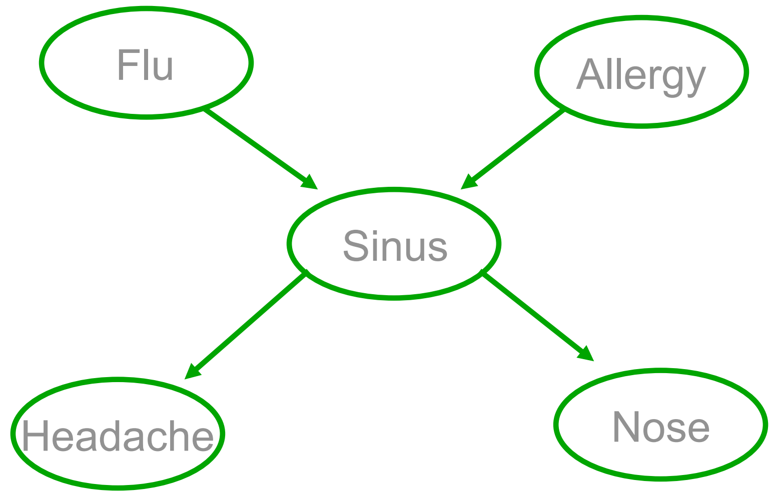
- Interesting case: known graph structure, training data is only partly observed, learn CPD parameters

Learning CPTs from Fully Observed Data

- Example: Consider learning the parameter

$$\theta_{s|ij} \equiv P(S=1|F=i, A=j)$$

- MLE (Max Likelihood Estimate) is

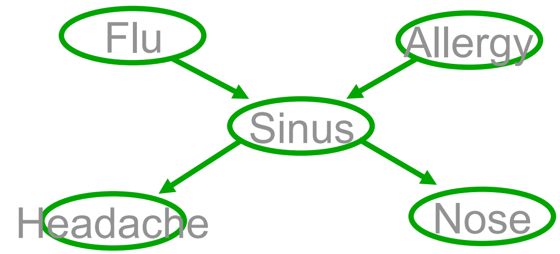


$$\theta_{s|ij} = \frac{\sum_{k=1}^{K} \delta(f_k = i, a_k = j, s_k = 1)}{\sum_{k=1}^{K} \delta(f_k = i, a_k = j)}$$

k[th] training example

- Remember why?

MLE estimate of $\theta_{s|ij}$ from fully observed data



- Maximum likelihood estimate

$$\theta \leftarrow \arg\max_{\theta} \log P(data|\theta)$$

data likelihood

- Our case:

$$P(data|\theta) = \prod_{k=1}^{K} P(f_k, a_k, s_k, h_k, n_k)$$

kth examp

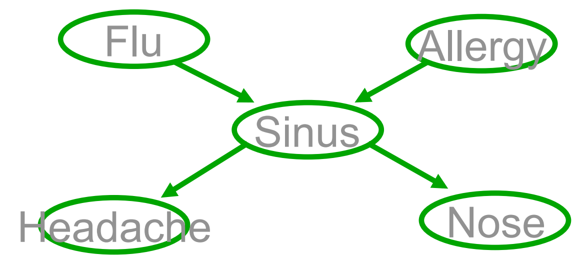
$$P(data|\theta) = \prod_{k=1}^{K} P(f_k)P(a_k)P(s_k|f_k a_k)P(h_k|s_k)P(n_k|s_k)$$

$$\log P(data|\theta) = \sum_{k=1}^{K} \log P(f_k) + \log P(a_k) + \log P(s_k|f_k a_k) + \log P(h_k|s_k) + \log P(n_k|s_k)$$

$$\frac{\partial \log P(data|\theta)}{\partial \theta_{s|ij}} = \sum_{k=1}^{K} \frac{\partial \log P(s_k|f_k a_k)}{\partial \theta_{s|ij}}$$

$$\theta_{s|ij} = \frac{\sum_{k=1}^{K} \delta(f_k = i, a_k = j, s_k = 1)}{\sum_{k=1}^{K} \delta(f_k = i, a_k = j)}$$

Estimate θ from partly observed data



- What if FAHN observed, but not S?
- Can't calculate MLE

$$\theta \leftarrow \arg\max_{\theta} \log \prod_k P(f_k, a_k, s_k, h_k, n_k|\theta)$$

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

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

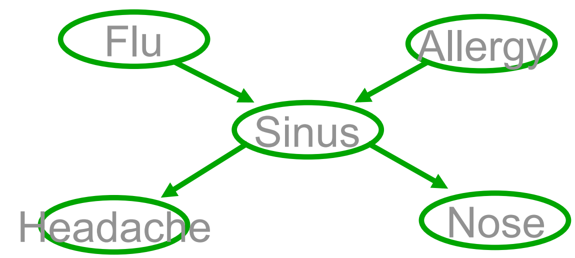
- WHAT TO DO?

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

EM

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

Estimate θ from partly observed data



- What if FAHN observed, but not S?
- Can't calculate MLE

$$\theta \leftarrow \arg\max_\theta \log \prod_k P(f_k, a_k, s_k, h_k, n_k|\theta)$$

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

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

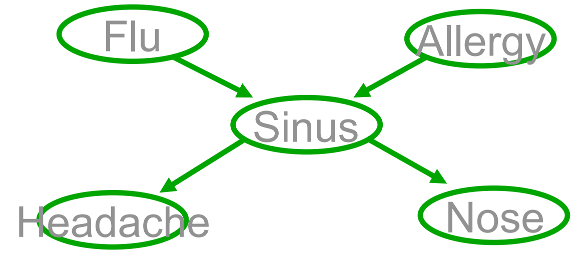
- EM seeks* to estimate:

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

* EM guaranteed to find local maximum

- EM seeks estimate:

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



- here, observed X={F,A,H,N}, unobserved Z={S}

$$\log P(X, Z|\theta) = \sum_{k=1}^{K} \log P(f_k) + \log P(a_k) + \log P(s_k|f_k a_k) + \log P(h_k|s_k) + \log P(n_k|s_k)$$

$$E_{X|Z,\theta}[log P(X, Z|\theta)] = \sum_{k=1}^{K} \sum_{i=0}^{1} P(s_k = i|f_k, a_k, h_k, n_k)$$

$$[log P(f_k) + \log P(a_k) + \log P(s_k|f_k a_k) + \log P(h_k|s_k) + \log P(n_k|s_k)]$$

EM Algorithm

EM is a general procedure for learning from partly observed data

Given observed variables X, unobserved Z (X={F,A,H,N}, Z={S})

Define $Q(\theta'|\theta) = E_{P(Z|X,\theta)}[\log P(X,Z|\theta')]$

Iterate until convergence:

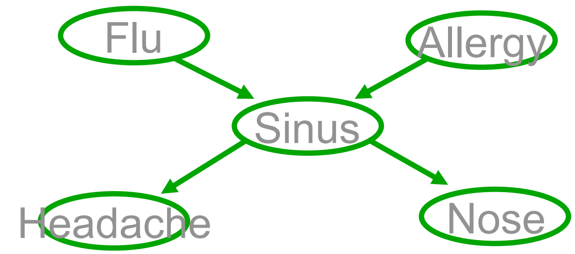
- E Step: Use X and current θ to calculate P(Z|X,θ)
- M Step: Replace current θ by

$$\theta \leftarrow \arg\max_{\theta'} Q(\theta'|\theta)$$

Guaranteed to find local maximum.
Each iteration increases $E_{P(Z|X,\theta)}[\log P(X,Z|\theta')]$

E Step: Use X, θ, to Calculate P(Z|X,θ)

observed X={F,A,H,N},
unobserved Z={S}



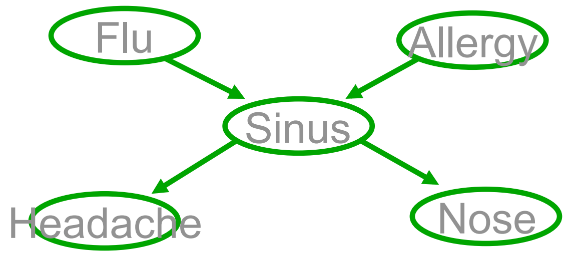
- How? Bayes net inference problem.

$$P(S_k = 1 | f_k a_k h_k n_k, \theta) = \frac{P(S_k=1\ f_k\ a_k h_k n_k)}{P(f_k\ a_k h_k\ n_k)}$$

$$P(S_k=1, f_k\ a_k h_k n_k) + P(S_k=0, f_k\ a_k h_k n_k)$$

E Step: Use X, θ, to Calculate P(Z|X,θ)

observed X={F,A,H,N},
unobserved Z={S}

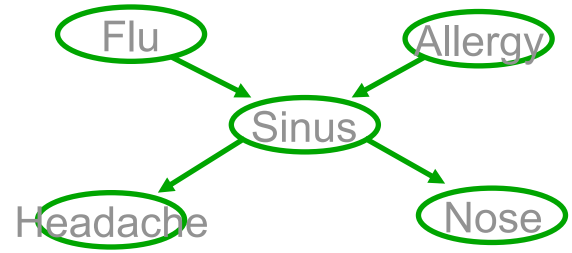


- How? Bayes net inference problem.

$$P(S_k = 1 | f_k a_k h_k n_k, \theta) =$$

$$P(S_k = 1 | f_k a_k h_k n_k, \theta) = \frac{P(S_k = 1, f_k a_k h_k n_k | \theta)}{P(S_k = 1, f_k a_k h_k n_k | \theta) + P(S_k = 0, f_k a_k h_k n_k | \theta)}$$

EM and estimating $\theta_{s|ij}$



observed X = {F,A,H,N}, unobserved Z={S}

E step: Calculate for each training example, k

$$P(S_k = 1 | f_k a_k h_k n_k, \theta) = E[s_k] = \frac{P(S_k = 1, f_k a_k h_k n_k | \theta)}{P(S_k = 1, f_k a_k h_k n_k | \theta) + P(S_k = 0, f_k a_k h_k n_k | \theta)}$$

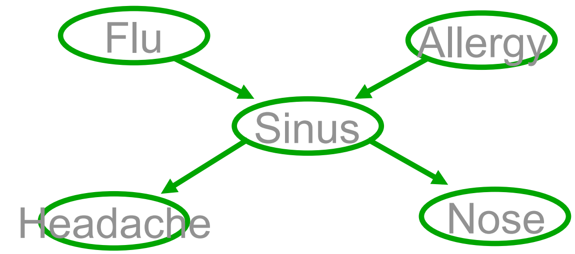
$P(S|AFHN)$

M step: update all relevant parameters. For example:

$$\theta_{s|ij} \leftarrow \frac{\sum_{k=1}^{K} \delta(f_k = i, a_k = j) \;\; E[s_k]}{\sum_{k=1}^{K} \delta(f_k = i, a_k = j)}$$

Recall MLE was: $\theta_{s|ij} = \frac{\sum_{k=1}^{K} \delta(f_k = i, a_k = j, s_k = 1)}{\sum_{k=1}^{K} \delta(f_k = i, a_k = j)}$

EM and estimating θ



More generally,
Given observed set X, unobserved set Z of boolean values

E step: Calculate for each training example, k

the expected value of each unobserved variable

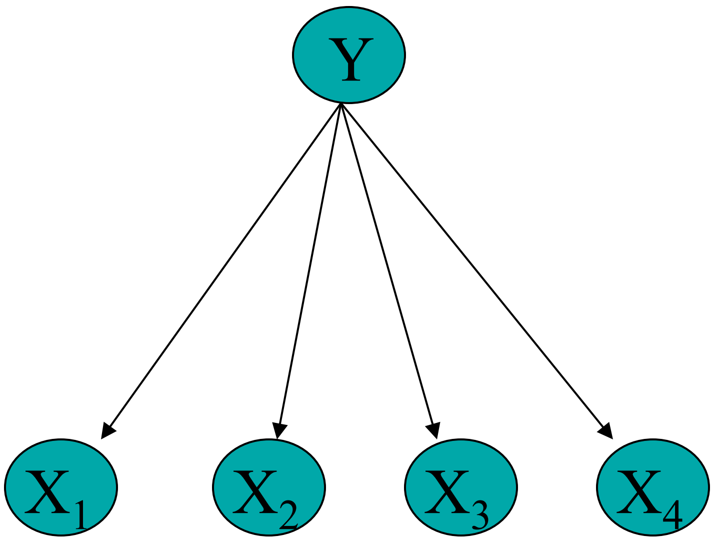
M step:

Calculate estimates similar to MLE, but replacing each count by its <u>expected count</u>

$$\delta(Y=1) \rightarrow E_{Z|X,\theta}[Y] \qquad \delta(Y=0) \rightarrow (1 - E_{Z|X,\theta}[Y])$$

Using Unlabeled Data to Help Train Naïve Bayes Classifier

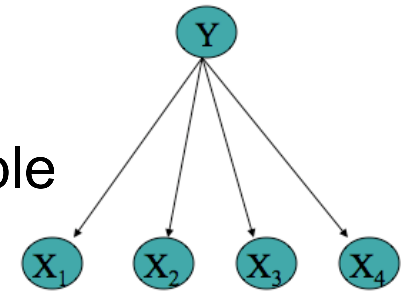
Learn P(Y|X)



Y	X1	X2	X3	X4
1	0	0	1	1
0	0	1	0	0
0	0	0	1	0
?	0	1	1	0
?	0	1	0	1

E step: Calculate for each training example, k

the expected value of each unobserved variable

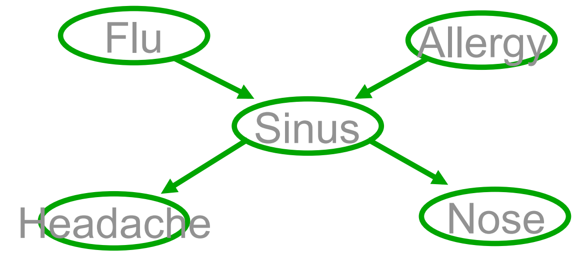


Exp val for Y_k given X_{1k} X_{2k} ... X_{nk}

k^{th} examp

$$P(Y|X_1, X_2 \dots X_n, \theta) = \frac{P(Y)\prod_i P(X_i|Y)}{P(X)}$$

EM and estimating θ



More generally,
Given observed set X, unobserved set Z of boolean values

E step: Calculate for each training example, k

the expected value of each unobserved variable

M step:

Calculate estimates similar to MLE, but
replacing each count by its <u>expected count</u>

$$\delta(Y=1) \rightarrow E_{Z|X,\theta}[Y] \qquad \delta(Y=0) \rightarrow (1 - E_{Z|X,\theta}[Y])$$

$P(x_i | Y=0)$ $\quad$ $P(x_i | Y=1)$