Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

October 11, 2017

Today:

- Graphical models
- Bayes Nets:
  - Conditional independencies
  - Simple inference

Readings:

- Bishop chapter 8

Warning! Your HW6 code might take hours to run.

If it’s not perfect, you might need multiple runs.

Do not wait until the last minute to begin!

Warning! Your HW6 code might take hours to run.

If it’s not perfect, you might need multiple runs.

Do not wait until the last minute to begin!

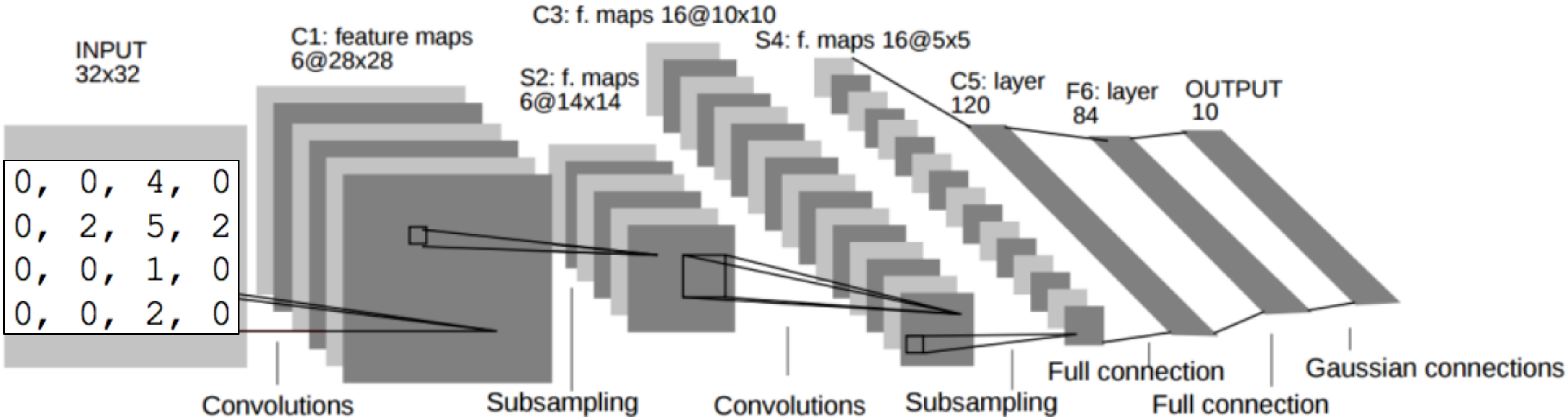
- **New HW6 due date: Friday night, Oct 13**
- **HW7 due date will also be pushed forward to Oct 20, but will still be handed out today**

Q: Why is Tom not replying to my personal emails??

A: 5 minutes/email * 500 students = 42 hours

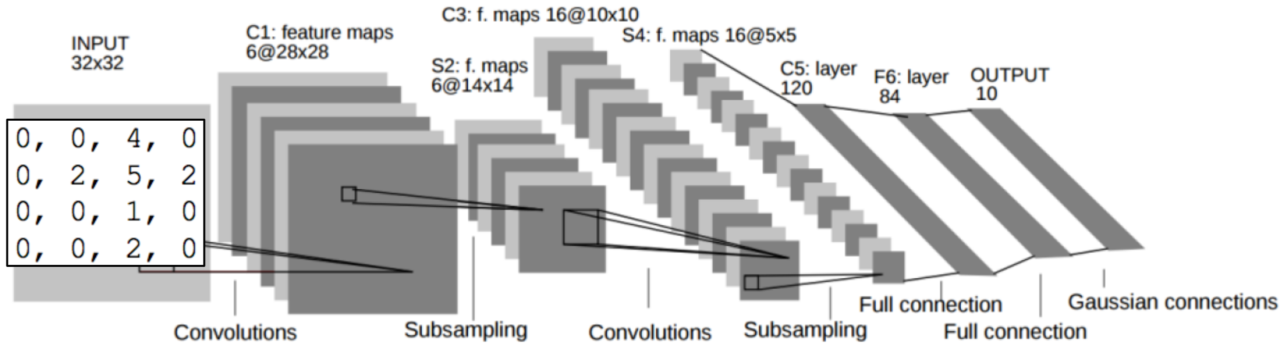
Reminder: My office hours are right after class, here.

How to solve HW5 without deriving any equations



- What is the (2,1,1) entry of the first convolution layer output?
- What is the loss function value for this given input and target output?
- What is the value of the derivative of your loss function with respect to the (2,1,1) entry of the first convolution layer output?

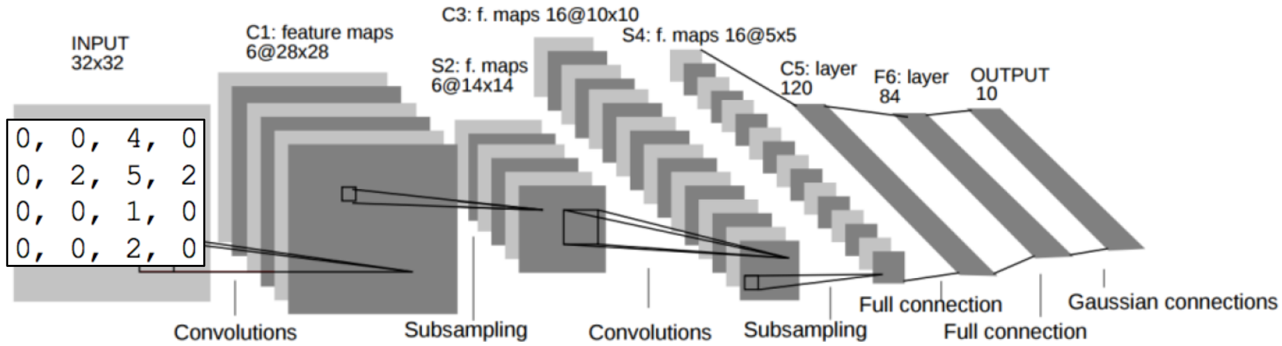
How to solve HW5 without deriving any equations



- What is the value of the derivative of your loss function with respect to the (2,1,1) entry of the first convolution layer output?

hint: $$\left(\frac{\partial f(x)}{\partial x}\right)_{x=x_0} = \lim_{\Delta \to 0} \frac{f(x_0+\Delta) - f(x_0)}{\Delta}$$

How to solve HW5 without deriving any equations



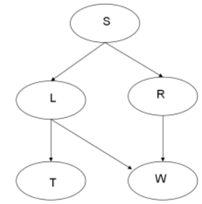
- What is the value of the derivative of your loss function with respect to the (2,1,1) entry of the first convolution layer output?

hint: $$\left(\frac{\partial f(x)}{\partial x}\right)_{x=x_0} = \lim_{\Delta \to 0} \frac{f(x_0+\Delta) - f(x_0)}{\Delta}$$

$$\left(\frac{\partial f(x)}{\partial x}\right)_{x=x_0} \approx \frac{f(x_0+\Delta) - f(x_0)}{\Delta}$$

for sufficiently tiny Δ (e.g., $\Delta = 0.0001 \cdot x_0$)

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 conditional probability distributions (CPD's)

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

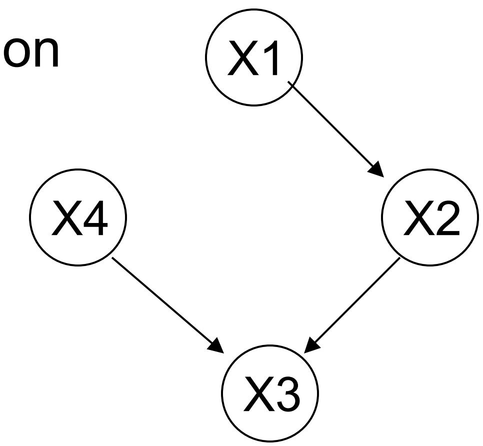
$$P(X_1 \dots X_n) = \prod_i P(X_i | Pa(X_i))$$

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

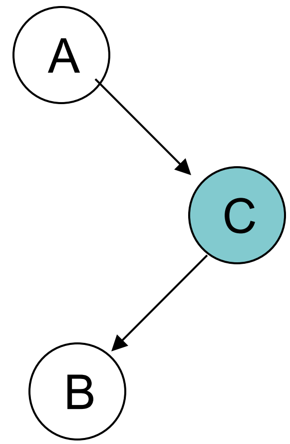
Conditional Independence, Revisited

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

- Does this rule give us all 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

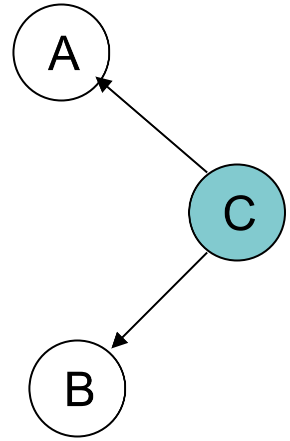


prove A cond indep of B given C?

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

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

Easy Network 2: Tail to Tail

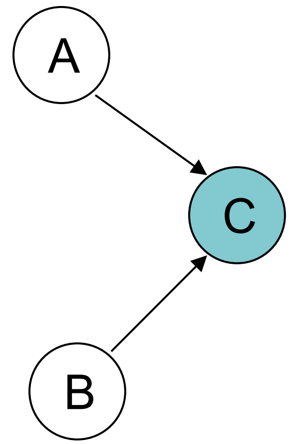


prove A cond indep of B given C? ie., $p(a,b|c) = p(a|c)\ p(b|c)$

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

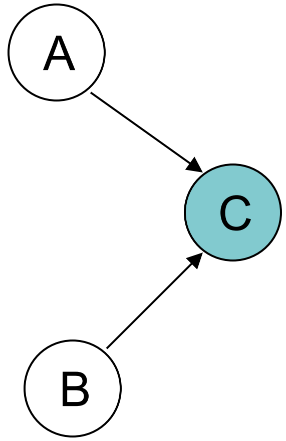
Easy Network 3: Head to Head

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



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) Does Not Equal 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.

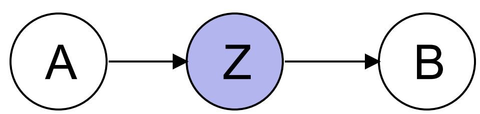
[Bishop, 8.2.2]

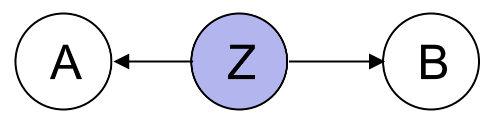
Suppose we have three <u>sets</u> of random variables: X, Y and Z

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

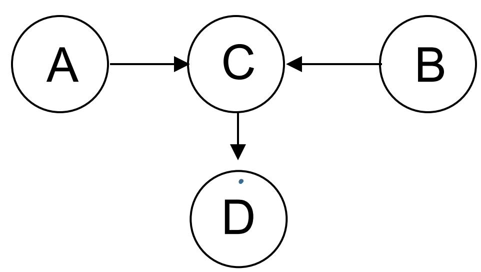
A path from variable X to variable Y is **<u>blocked</u>** if it includes a node such that *either* of the following holds:

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

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

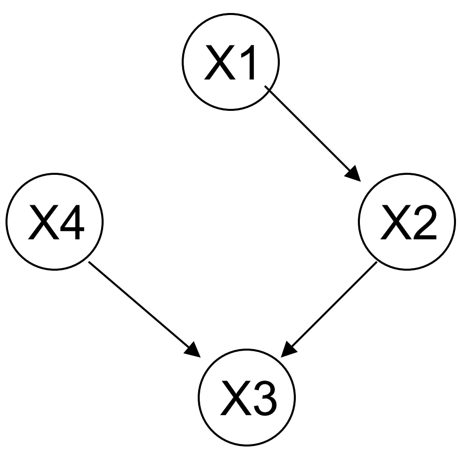
1. arrows on the path meet either head-to-tail or tail-to-tail at the node and this node is in Z

2. or, the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in Z

X1 indep of X3 given X2?

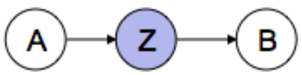
X3 indep of X1 given X2?

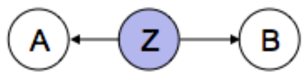
X4 indep of X1 given X2?

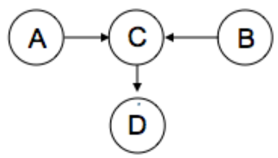


X and Y are **D-separated** by Z (and therefore conditionally indep, given Z) iff every path from any variable in X to any variable in Y is **blocked** by Z

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 

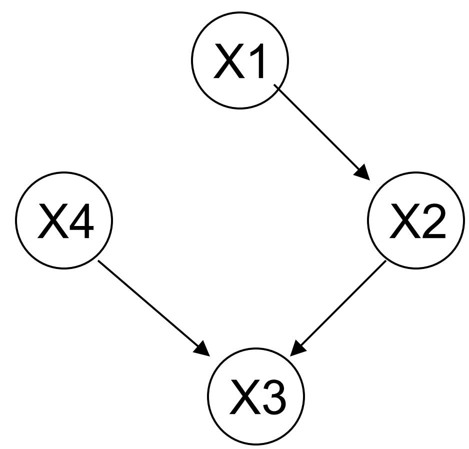


2.the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in Z 

X4 indep of X1 given X3?

X4 indep of X1 given {X3, X2}?

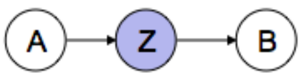
X4 indep of X1 given {}?



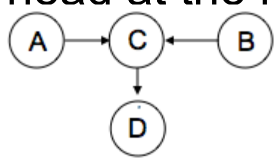
X and Y are **<u>D-separated</u>** by Z (and therefore conditionally indep, given Z) iff every path from any variable in X to any variable in Y is **<u>blocked</u>** by Z

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



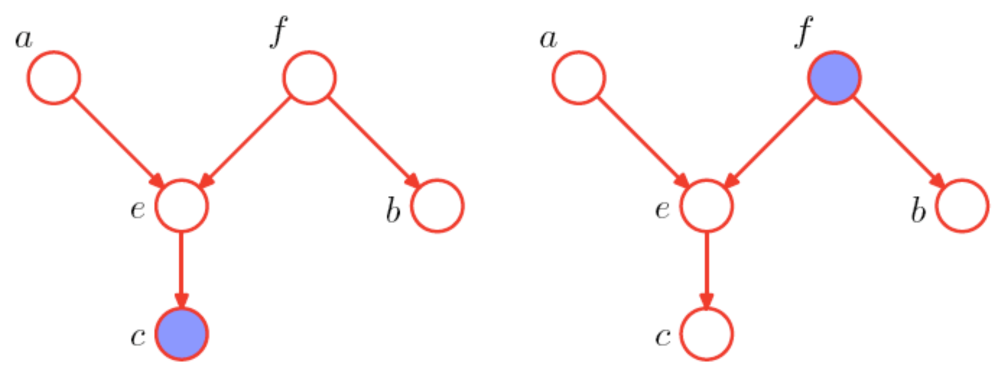
2.the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in Z



a indep of b given c?

a indep of b given f ?

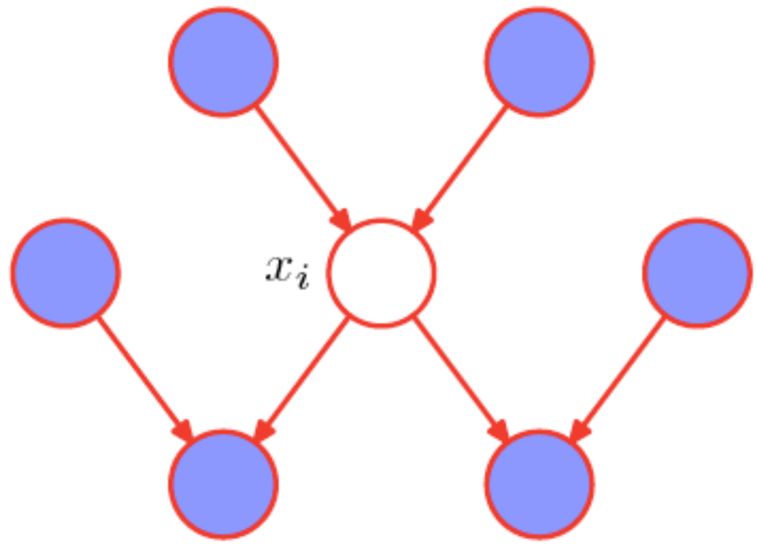
a indep of b given {} ?



(a) (b)

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.



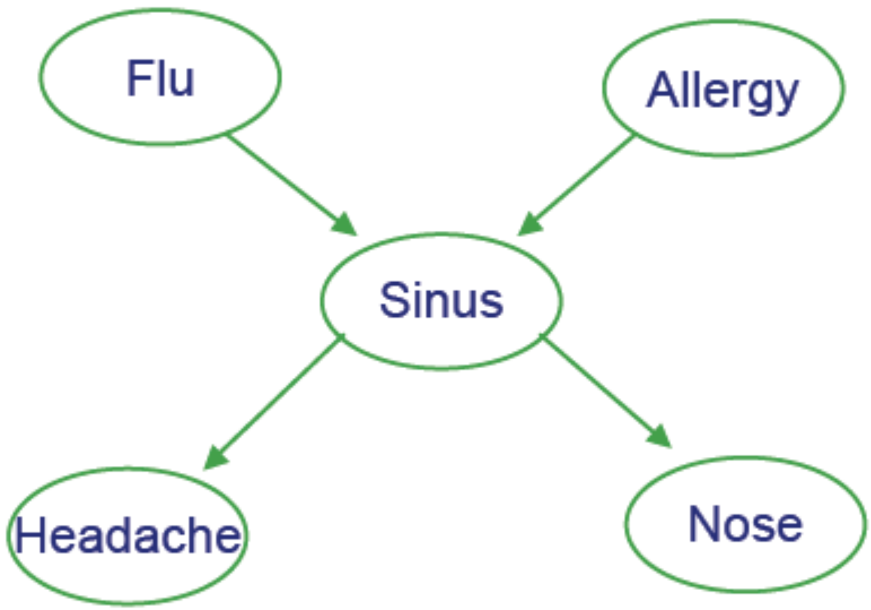
from [Bishop, 8.2]

Inference in Bayes Nets

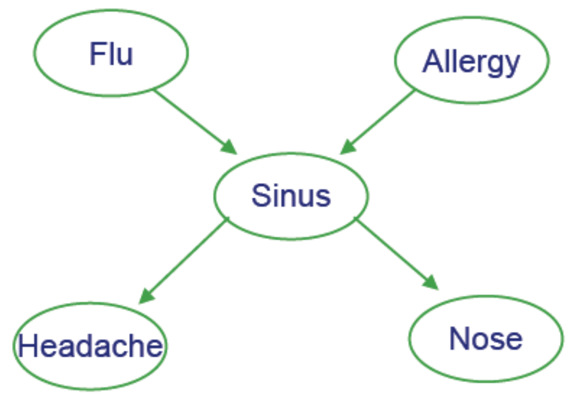
- In general, intractable (NP-complete)
- For certain cases, tractable

Example

- Flu and Allegies both cause Sinus problems
- Sinus problems cause Headaches and runny Nose



Prob. of joint assignment: easy



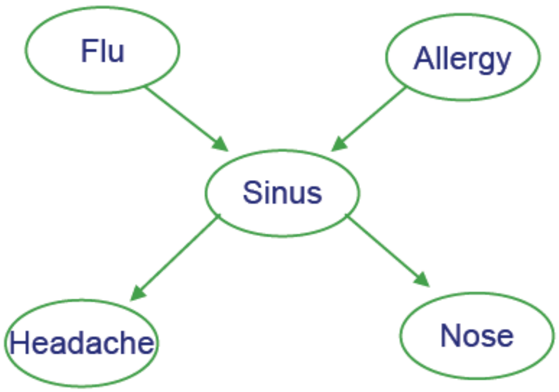
Suppose we are interested in joint assignment <F=f,A=a,S=s,H=h,N=n>

What is P(f,a,s,h,n)?

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

Marginal probabilities $P(X_i)$: not so easy

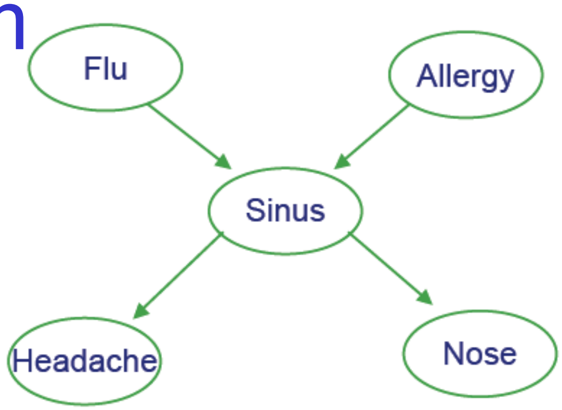
- How do we calculate P(N=n) ?



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

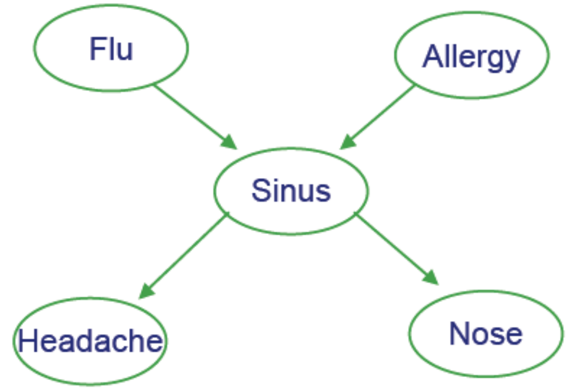
Generating a random sample from joint distribution: easy

How can we generate random samples drawn according to P(F,A,S,H,N)?



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

Generating a sample from joint distribution: easy



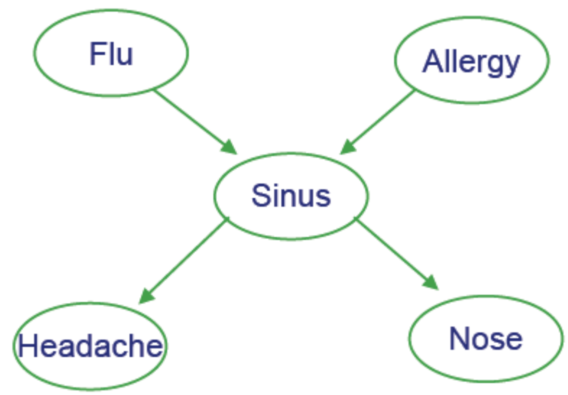
How can we generate random samples drawn according to P(F,A,S,H,N)?

To generate a random sample for roots of network (F or A):

1. let θ = P(F=1) # look up from CPD
2. r = random number drawn uniformly between 0 and 1
3. if $r<\theta$ then output 1, else 0

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

Generating a sample from joint distribution: easy



How can we generate random samples drawn according to P(F,A,S,H,N)?

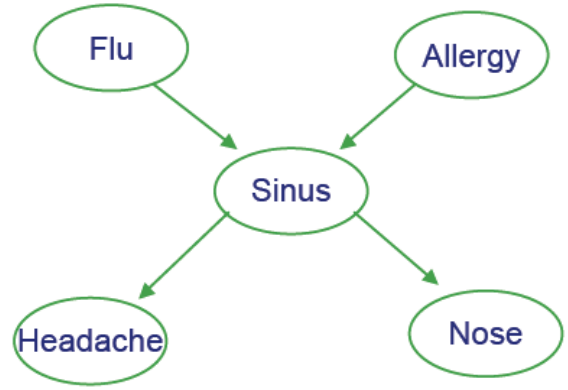
To generate a random sample for roots of network (F or A):

1. let θ = P(F=1) # look up from CPD
2. r = random number drawn uniformly between 0 and 1
3. if $r<\theta$ then output 1, else 0

To generate a random sample for S, given F,A:

1. let θ = P(S=1|F=f,A=a) # look up from CPD
2. r = random number drawn uniformly between 0 and 1
3. if $r<\theta$ then output 1, else 0

Generating a sample from joint distribution: easy



Note we can estimate marginals

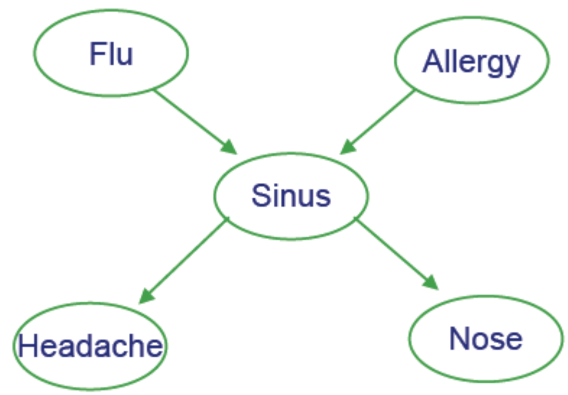
like P(N=n) by generating many samples

from joint distribution, then count the fraction of samples for which N=n

Similarly, for anything else we care about P(F=1|H=1, N=0)

→ weak but general method for estimating <u>any</u> probability term…

Generating a sample from joint distribution: easy

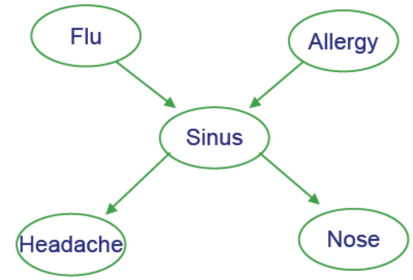


We can easily sample P(F,A,S,H,N)

Can we use this to get P(F,A,S,H | N)?

Directly sample P(F,A,S,H | N)?

Gibbs Sampling:



Goal: Directly sample conditional distributions

$P(X_1,\ldots,X_n \mid X_{n+1}, \ldots, X_m)$

Approach:

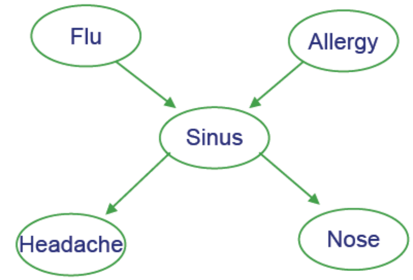
- start with arbitrary initial values for unobserved $X_1^{(0)},\ldots,X_n^{(0)}$ (and the observed $X_{n+1}, \ldots, X_m$)
- iterate for s=0 to a big number:

$$X_1^{s+1} \sim P(X_1 | X_2^s, X_3^s \ldots\ X_n^s, X_{n+1}, \ldots\ X_m)$$
$$X_2^{s+1} \sim P(X_2 | X_1^{s+1}, X_3^s \ldots\ X_n^s, X_{n+1}, \ldots\ X_m)$$
$$\ldots$$
$$X_n^{s+1} \sim P(X_n | X_1^{s+1}, X_2^{s+1}, \ldots\ X_{n-1}^{s+1}, X_{n+1}, \ldots\ X_m)$$

Eventually (after burn-in), the collection of samples will constitute a sample of the true $P(X_1,\ldots,X_n \mid X_{n+1}, \ldots, X_m)$

* but often use every 100th sample, since iters not independent

Gibbs Sampling:



Approach:

- start with arbitrary initial values for $X_1^{(0)},\ldots,X_n^{(0)}$ (and observed $X_{n+1}, \ldots, X_m$)
- iterate for s=0 to a big number:

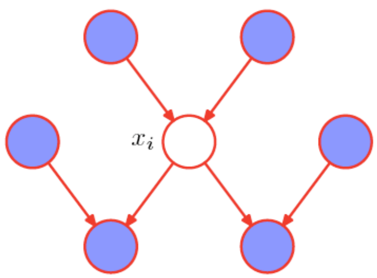
$$X_1^{s+1} \sim P(X_1 | X_2^s, X_3^s \ldots X_n^s, X_{n+1}, \ldots X_m)$$

$$X_2^{s+1} \sim P(X_2 | X_1^{s+1}, X_3^s \ldots X_n^s, X_{n+1}, \ldots X_m)$$

$$\ldots$$

$$X_n^{s+1} \sim P(X_n | X_1^{s+1}, X_2^{s+1}, \ldots X_{n-1}^{s+1}, X_{n+1}, \ldots X_m)$$

Only need Markov Blanket at each step!

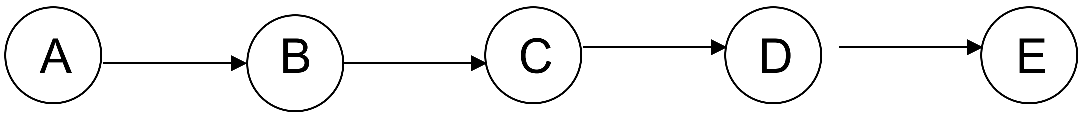


Gibbs is special case of Markov Chain Monte Carlo method

Prob. of marginals: not so easy

But sometimes the structure of the network allows us to be clever → avoid exponential work

eg., chain



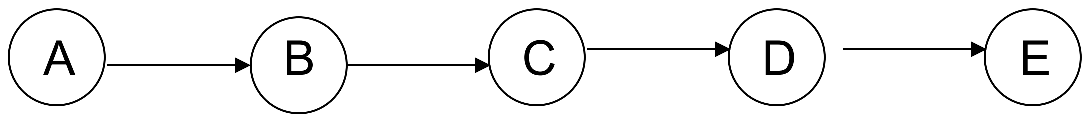
what is P(C=1|B=b, D=d)?

what is P(C=1) ?

Prob. of marginals: not so easy

But sometimes the structure of the network allows us to be clever → avoid exponential work

eg., chain



what is P(C=1) ?

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
- Can often use Monte Carlo methods
  - Generate many samples, then count up the results
  - Gibbs sampling (example of Markov Chain Monte Carlo)
- Many other approaches
  - Variational methods for tractable approximate solutions
  - Junction tree, Belief propagation, …

see Graphical Models course 10-708