Graphical Models
and
Bayesian Networks II

Required reading:

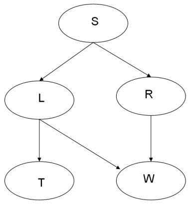
- Bishop online chapter 8, read all of section 8.2

Machine Learning 10-601

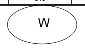
Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

February 18, 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

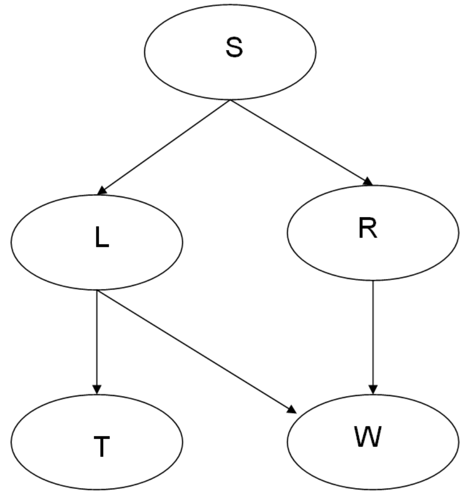
- Each node denotes a random variable
- Edges denote dependencies
- CPD for each node X_i defines $P(X_i \mid Pa(X_i))$
- The joint distribution over all variables is defined as

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

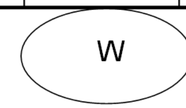
Pa(X) = immediate parents of X in the graph

Bayesian Networks

- CPD for each node X_i describes $P(X_i \mid Pa(X_i))$



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



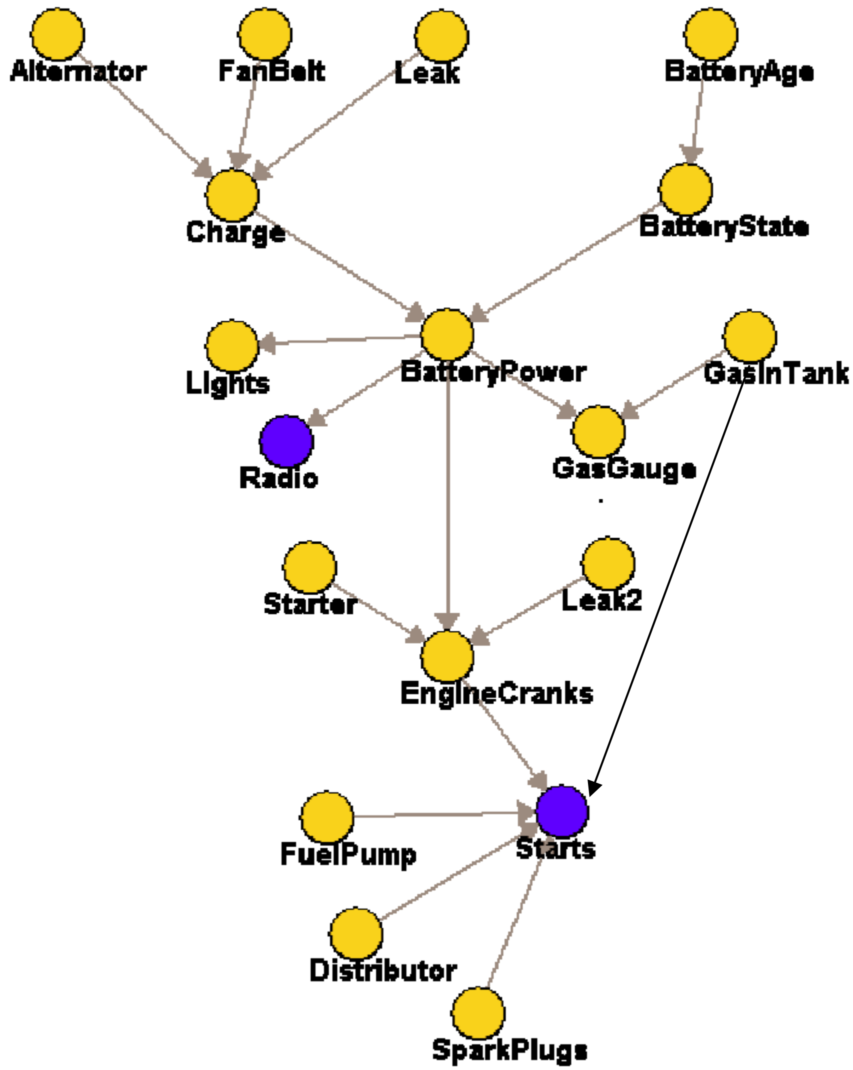
Chain rule of probability:

$$P(S,L,R,T,W) = P(S)P(L|S)P(R|S,L)P(T|S,L,R)P(W|S,L,R,T)$$

But in a Bayes net: $P(X_1 \ldots X_n) = \prod_i P(X_i|Pa(X_i))$

Questions we might ask of Bayes Net

$P(Y \mid x_1, \ldots x_n)$



Inference:

P(BattPower=t | Radio=t, Starts=f)

Most probable explanation:

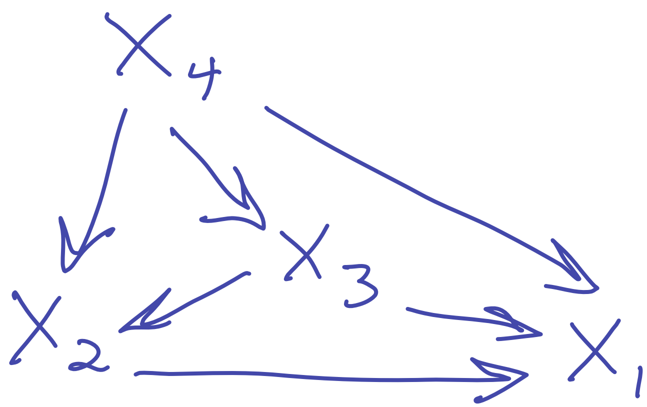
What is most likely value of <Leak, BatteryPower> given Starts=f?

Active data collection:

What is most useful variable to observe next, to improve our knowledge of node X?

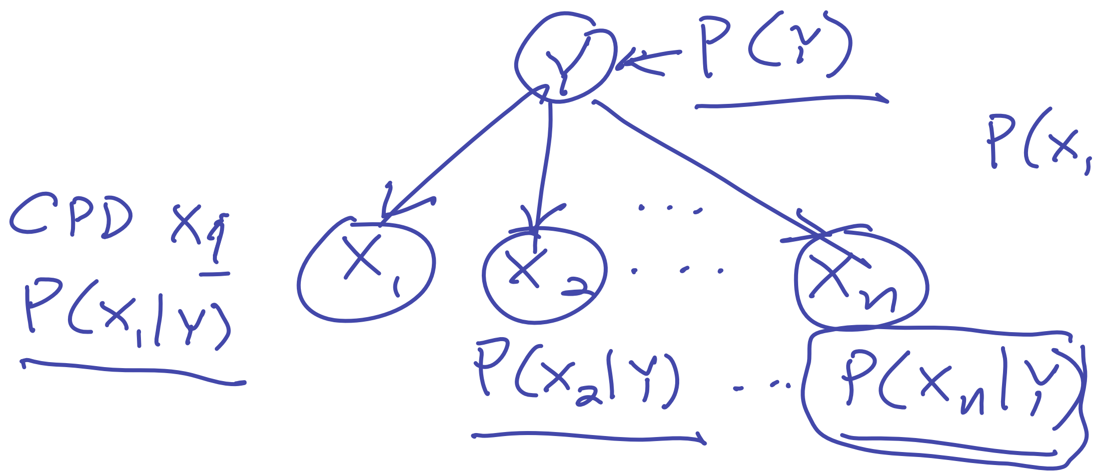
What is the Bayes Network for X1, X2, X3, X4 with NO assumed conditional independencies?

$$P(x_1 x_2 x_3 x_4) = \underline{P(x_1 \mid x_2 x_3 x_4)} \underbrace{P(x_2 x_3 x_4)}_{\underline{P(x_2 \mid x_3 x_4)\, P(x_3 \mid x_4)\, P(x_4)}}$$



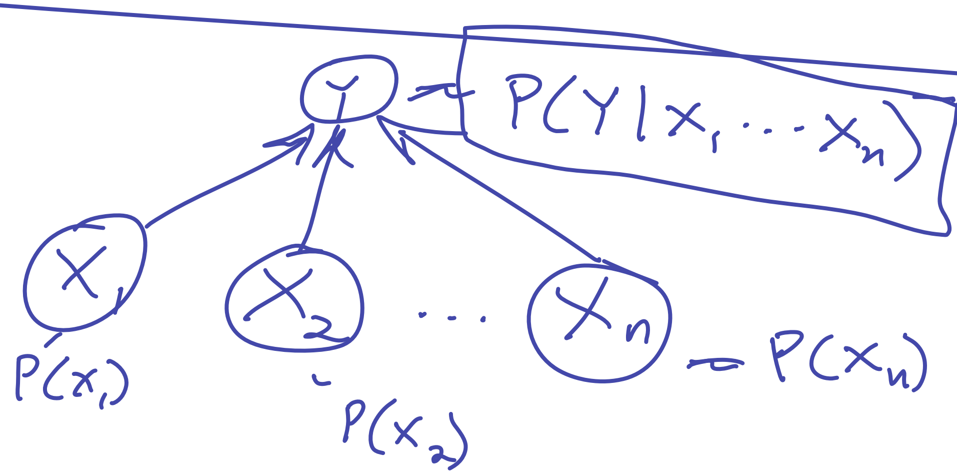
What is the Bayes Network for Naïve Bayes?

$P(Y \mid X_1, \dots X_n)$

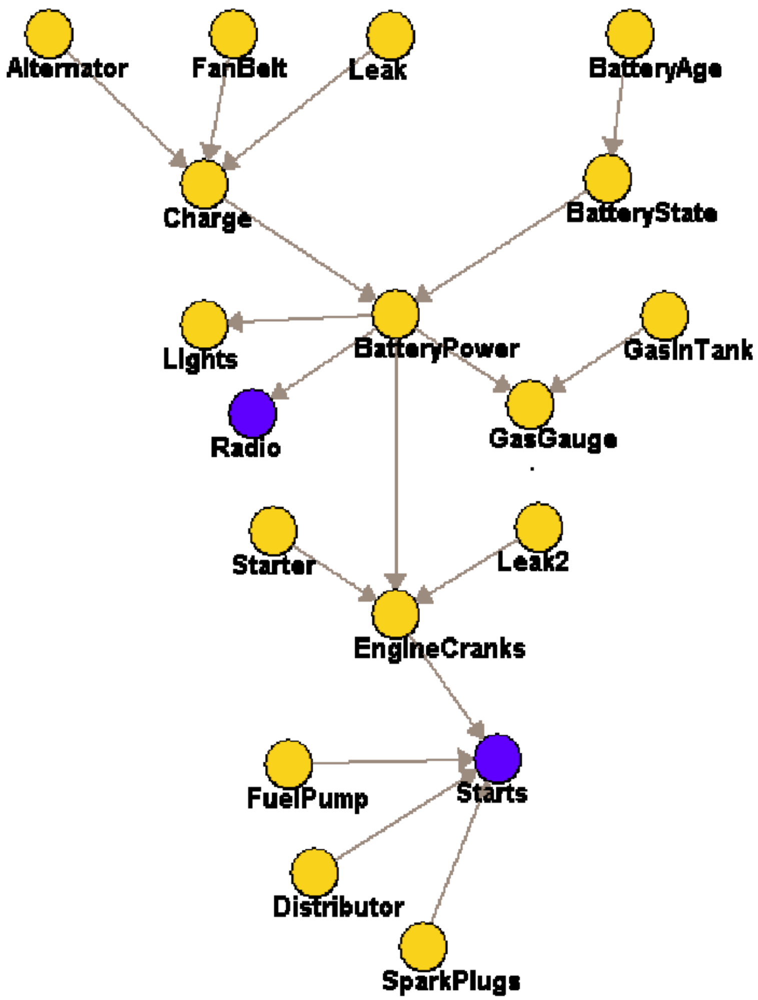


$P(X_1 \dots X_n Y) = P(Y) P(X_1 \mid Y) \dots P(X_n \mid Y)$

$\boxed{P(Y \mid X_1 \dots X_n)}$

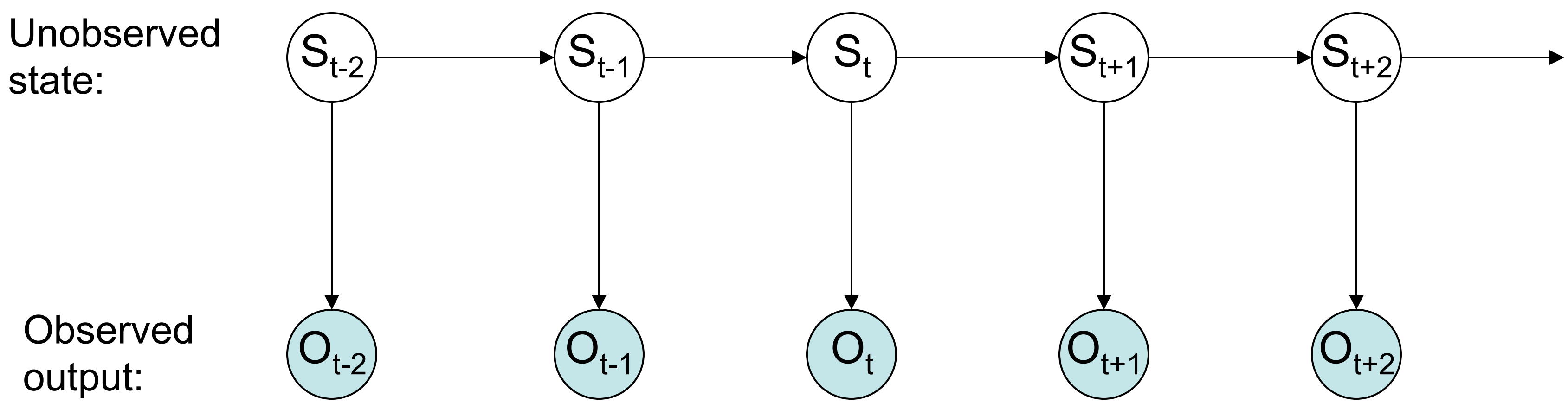


What do we do if variables are mix of discrete and real valued?



Bayes Network for a Hidden Markov Model

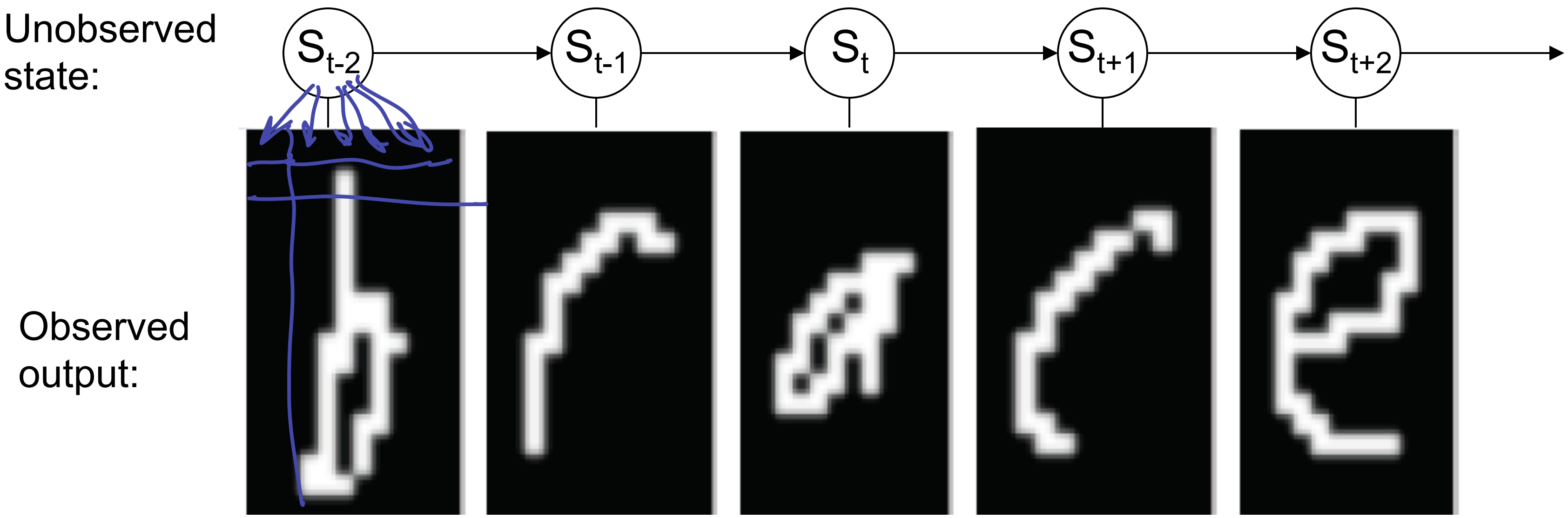
Assume the future is conditionally independent of the past, given the present



$$P(S_{t-2}, O_{t-2}, S_{t-1}, \ldots, O_{t+2}) =$$

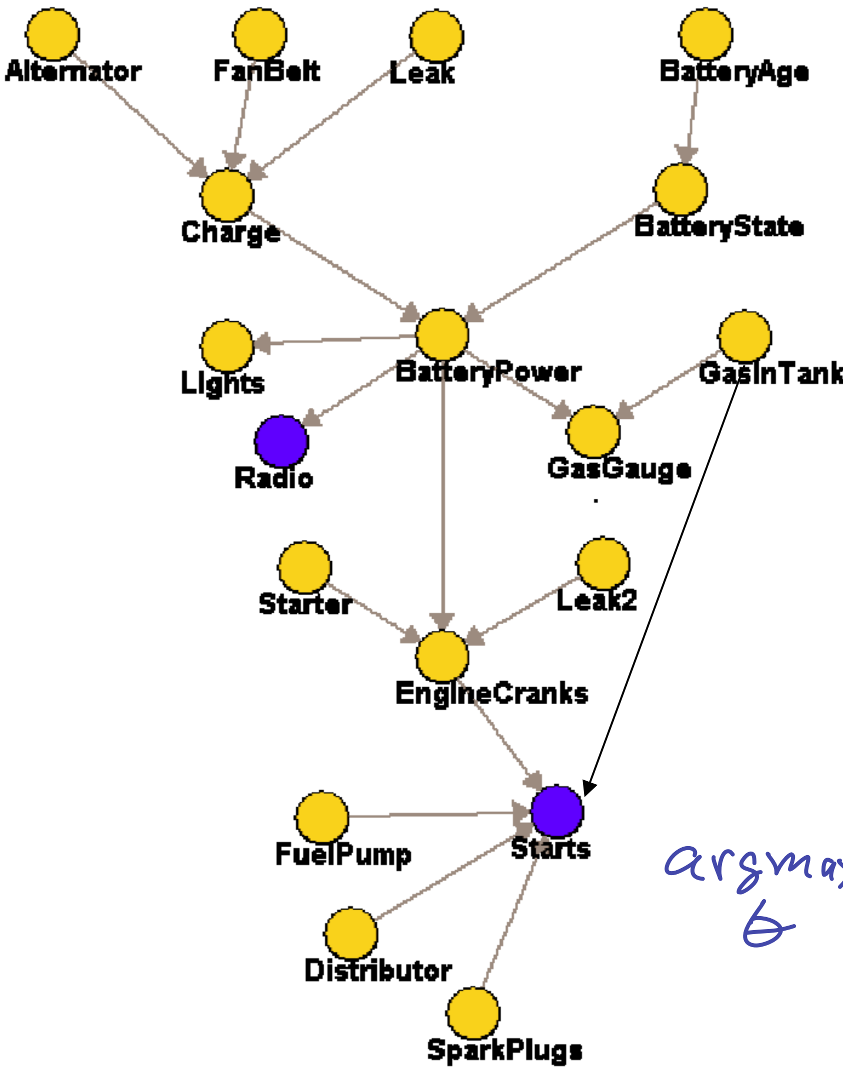
Bayes Network for a Hidden Markov Model

Assume each character t+1 is conditionally independent of the distant past, given character t



$P(S_{t-2}, O_{t-2}, S_{t-1}, \ldots, O_{t+2}) =$

How Can We Train a Bayes Net



1. when graph is given, and each training example gives value of every RV?

likelihood

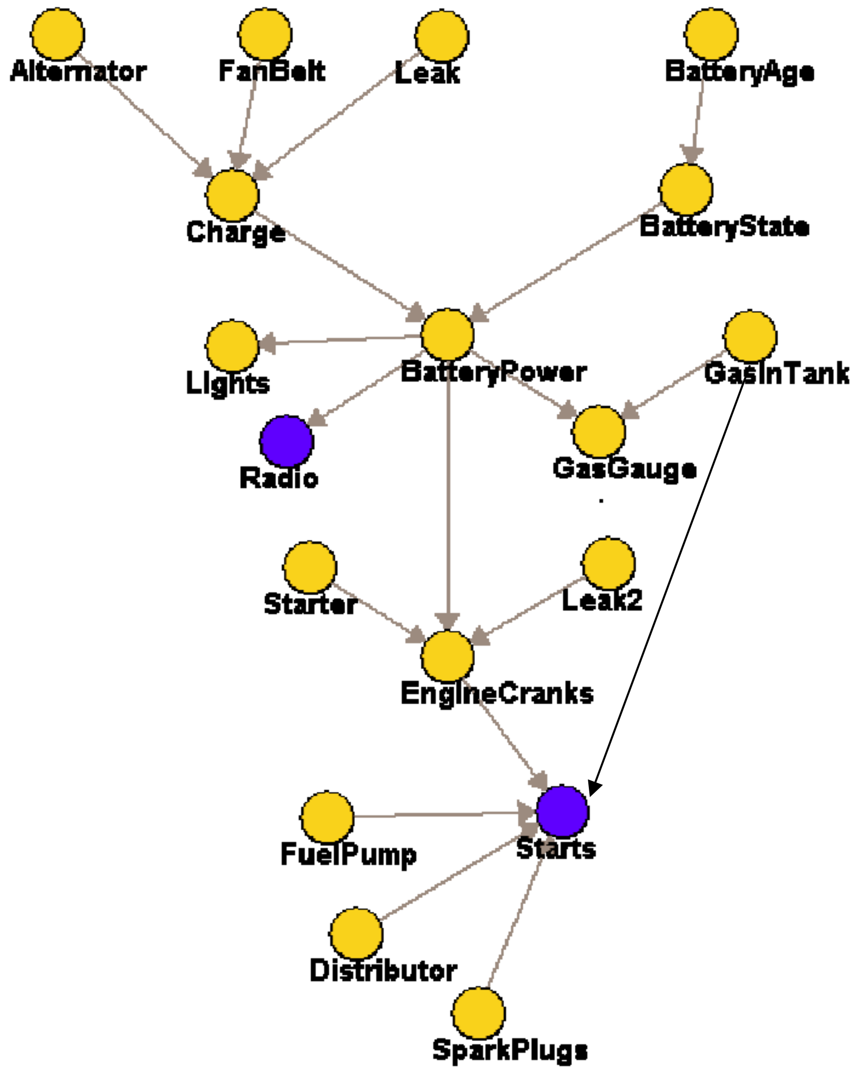
vector of parameters

$$\prod_{l \text{ examples}} \underbrace{P(\text{all the var values}}_{\prod_l \prod_{\text{nodes}} P(v \mid Pa(v); \theta)} \mid \theta)$$

~~nodes~~ Variables

$$\arg\max_{\theta} P(BS \mid BA_s; \theta)$$

How Can We Train a Bayes Net



1. when graph is given, and each training example gives value of every RV?

Easy: use data to obtain MLE or MAP estimates of θ for each CPD

$P(X_i \mid Pa(X_i); \theta)$

e.g. like training the CPD's of a naïve Bayes classifier

2. when graph unknown or some RV's unobserved?

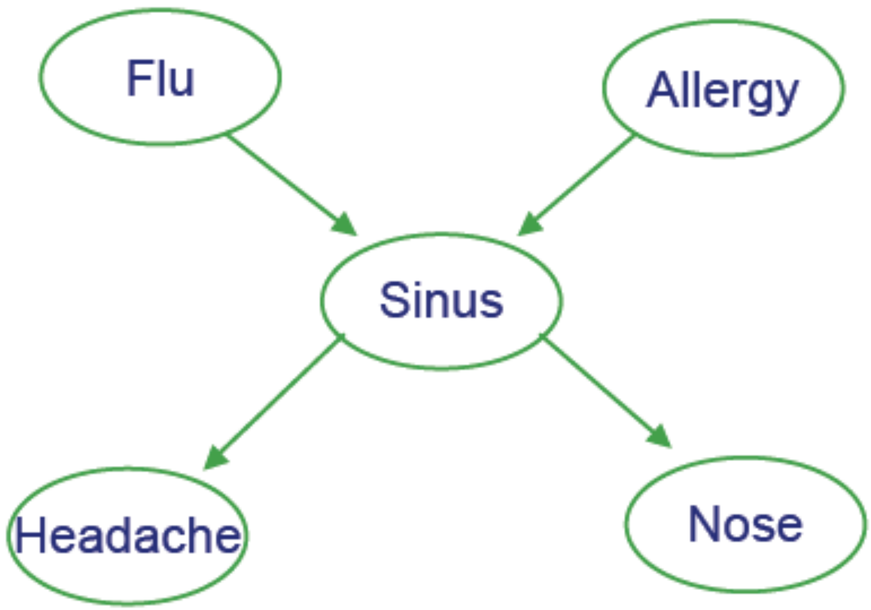
this is more difficult… later…

Inference in Bayes Nets

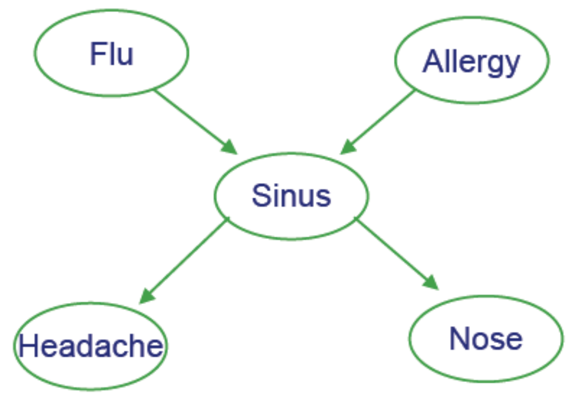
- In general, intractable (NP-complete)
- For certain cases, tractable
  - Assigning probability to fully observed set of variables
  - Or if just one variable unobserved
  - Or for singly connected graphs (ie., no undirected loops)
    - Belief propagation
- For multiply connected graphs
    - Junction tree
- 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

Example

- Bird flu and Allegies both cause Sinus problems
- Sinus problems cause Headaches and runny Nose



Prob. of joint assignment: easy



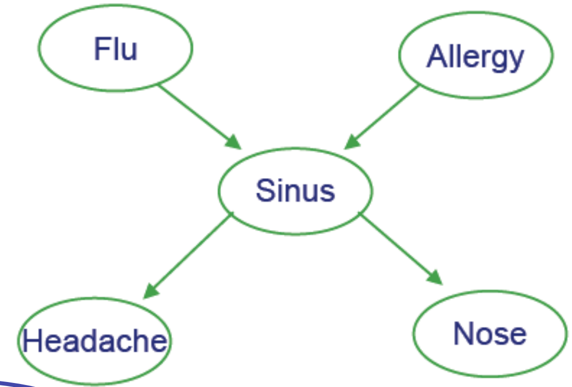
- Suppose we are interested in joint assignment <F=f,A=a,S=s,H=h,N=n>

1 0 1 1 0

What is P(f,a,s,h,n)? $= P(f)\, P(a)\, P(s|f,a)\, P(h|s)\, P(n|s)$

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

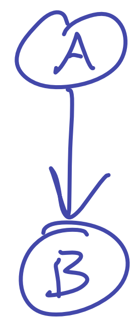
Prob. of marginals: not so easy



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

$$\sum_{f\in F}\ \sum_{a\in A}\ \sum_{h\in H}\ \sum_{s\in S} P(f\ a\ h\ s\ n)$$

values that A can take

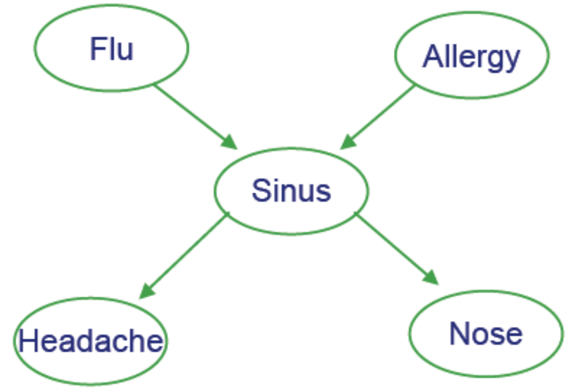


$P(B|A) =$

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

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)?

$$P(F)\, P(A)\, P(S|FA)\, P(H|S)\, P(N|S)$$

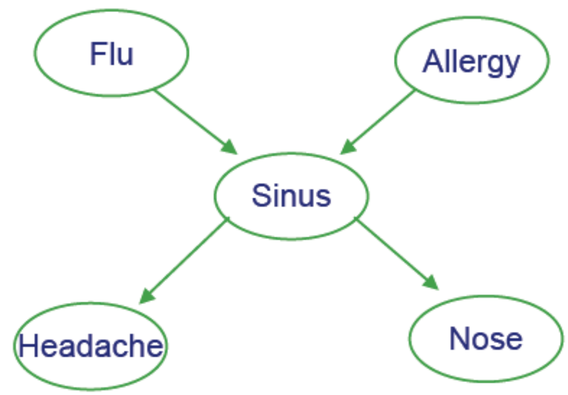
f a

$P(1) = \theta$

$P(0) = 1-\theta$

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

Generating a sample from joint distribution: easy



Note we can calculate marginals

like P(N=n) by generating many samples

from joint distribution, and then counting the fraction for which N=n

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

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

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

Prob. of marginals: not so easy

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

$P(C=1)$

eg., chain $A \to B \to C \to D \to E$ $\quad (C = 1)$

Brute force

$$P(C=1) = \sum_{a\in A}\sum_{b\in B}\sum_{d\in D}\sum_{e\in E} P(a\,b\,1\,d\,e)$$

Smarter $\sum_{A}\sum_{B}\sum_{D}\sum_{E} P(a)P(b|a)P(c=1|b)P(d|c=1)P(e|d)$

16×5 mults

$$\sum_{A}\sum_{B}\sum_{D} P \cdots \left[\sum_{E} P(e|d)\right]$$

8×5 mults

Variable Eliminates