Machine Learning 10-701

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

February 15, 2011

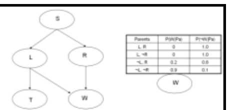
Today:

- Graphical models
- Inference
- Conditional independence and D-separation
- Learning from fully labeled data

Readings:

Required:
- Bishop chapter 8, through 8.2

Bayesian Networks Definition



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

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

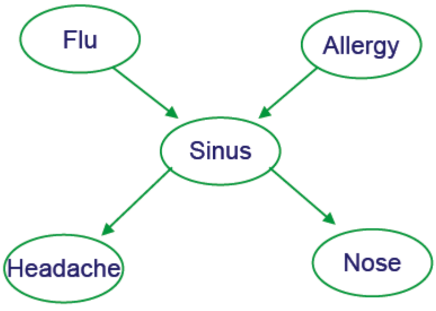
Pa(X) = immediate parents of X in the graph

Inference in Bayes Nets

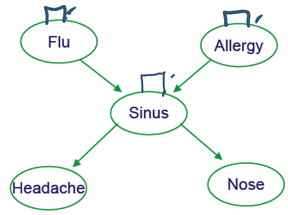
- In general, intractable (NP-complete)
- For certain cases, tractable
  - Assigning probability to fully observed set of variables
  - Or if just one variable unobserved
  - Or for singly connected graphs (ie., no undirected loops)
    - Belief propagation
- For multiply connected graphs
    - Junction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

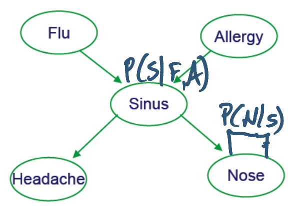
var value

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) ?

$$= \sum_s P(N=n|S=s)\,P(S=s)$$

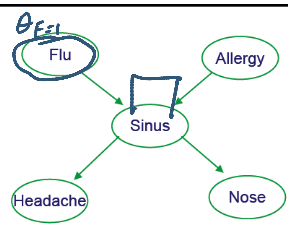
$$P(N=n) = \sum_{f,a,h,s} \underbrace{P(F=f, A=a, H=h, S=s, N=n)}_{P(f)\,P(a)\,P(s|f,a)\,P(h|s)\,P(n|s)}$$

K boolean vars

$$\text{Cost} = \underbrace{2^{K-1}}_{\text{terms in sum}} \cdot K \text{ mult}$$

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

Generating a sample from joint distribution: easy



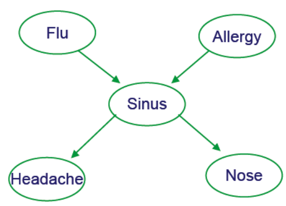
How can we generate random samples drawn according to P(F,A,S,H,N)?

randomly draw a value for $F=f$
draw $r \in [0,1]$ uniformly
if $r < \theta_{F=1}$ then output $f=1$
else $f=0$

draw f, a, $s|f,a$, $h|s$, $n|s$

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

Generating a sample from joint distribution: easy



Note we can estimate marginals like P(N=n) by generating many samples from joint distribution, by summing the probability mass 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…

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

Boolean vars. What is $P(c=1)$

eg., chain

$$\begin{tikzcd} A \arrow[r] & B \arrow[r] & C \arrow[r] & D \arrow[r] & E \end{tikzcd}$$

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

$$P(a)P(b|a)P(c=1|b)P(d|c=1)P(e|d)$$

Complexity Sum 16 terms, Mult 5 each

$$\sum_a\sum_b\sum_d P(a)P(b|a)P(c=1|b)P(d|c=1)\boxed{\sum_{e\in E}P(e|d)}$$

8 terms 4 way Mult

[Variable Elimination E]

Prob. of marginals: not so easy

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

eg., chain

$$\begin{tikzcd} A \arrow[r] & B \arrow[r] & C \arrow[r] & D \arrow[r] & E \end{tikzcd}$$

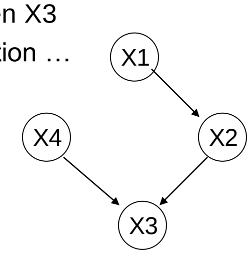
Inference in Bayes Nets

- 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
- 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 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 …

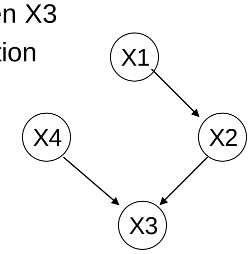


Inference in Bayes Nets

- 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
- 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

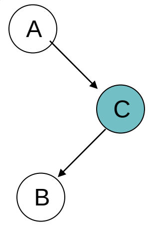
- We said:
  - Each node is conditionally independent of its non-descendents, given its immediate parents.
- Does this rule give us all of the conditional independence relations implied by the Bayes network?
  - No!
  - E.g., X1 and X4 are conditionally indep given {X2, X3}
  - But X1 and X4 not conditionally 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)



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

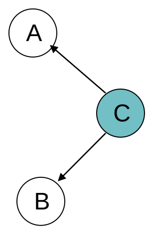
$$\boxed{\frac{P(a,c)}{P(c)}} \rightarrow P(a|c)$$

$$p(ab|c) = P(a|c)\,P(b|c)$$

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

Easy Network 2: Tail to Tail

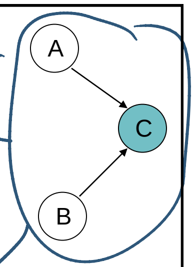
prove A cond indep of B given C?   ie., p(a,b|c) = p(a|c) p(b|c)



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

Easy Network 3: Head to Head

Collider



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

False

but true that

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

$$p(a\,b) = \sum_{c \in C} p(a\,b\,c)$$

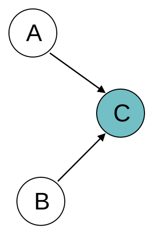
$$= \sum_{c \in C} P(a)\,P(b)\,P(c|a\,b)$$

$$= P(a)\,P(b) \underbrace{\boxed{\sum_{c \in C} P(c|a\,b)}}_{1}$$

$$= P(a)\,P(b)$$

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)

$\boxed{\text{Explaining away.}}$

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

Collider

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

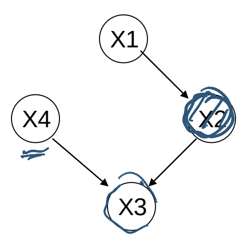
1. arrows on the path meet either head-to-tail or tail-to-tail at the node and this node is in Z

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

X1 indep of X3 given X2? Y I

X3 indep of X1 given X2? Y I

X4 indep of X1 given X2? Y



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

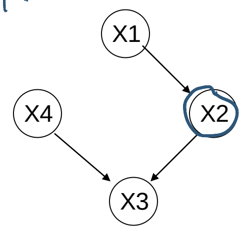
1.arrows on the path meet either head-to-tail or tail-to-tail at the node and this node is in Z

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}? $\leftarrow P(x_1 x_4 | x_2 x_3) = P(x_1 | x_2 x_3) \cdot P(x_4 | x_2 x_3)$

X4 indep of X1 given {}?



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**

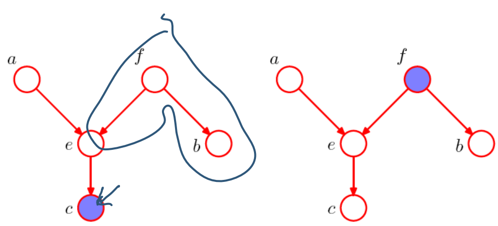
A path from variable A to variable B is **blocked** if it includes a node such that either

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?

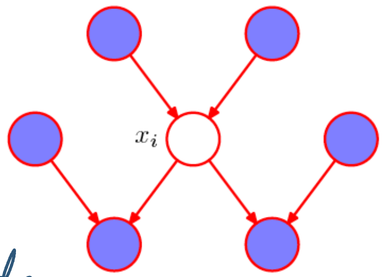
a indep of b given f ?



(a) (b)

Markov Blanket

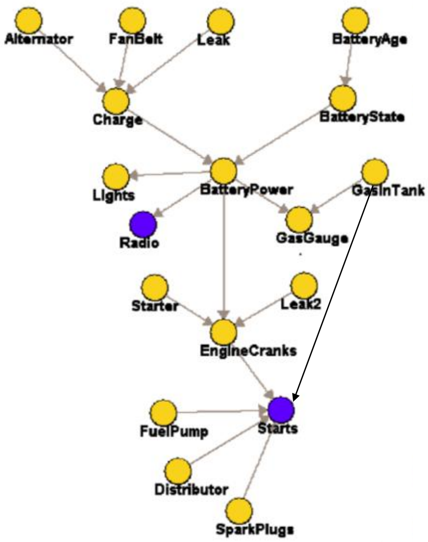
The Markov blanket of a node $\mathbf{x}_i$ comprises the set of parents, children and co-parents of the node. It has the property that the conditional distribution of $\mathbf{x}_i$, conditioned on all the remaining variables in the graph, is dependent only on the variables in the Markov blanket.



co-parent = other side of X_i's colliders

from [Bishop, 8.2]

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(Xi | Pa(Xi); θ)

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…

Learning in Bayes Nets

- Four categories of learning problems
  - Graph structure may be known/unknown
  - Variable values may be observed/unobserved
- Easy case: learn parameters for known graph structure, using fully observed data
- Gruesome case: learn graph and parameters, from partly unobserved data
- More on these in next lectures

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'