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

\[
P(X_1 \ldots X_n) = \prod_{i} P(X_i \mid Pa(X_i))
\]

\( Pa(X) = \) 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 (i.e., 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 Allergies both cause Sinus problems
- Sinus problems cause Headaches and runny Nose
• Suppose we are interested in joint assignment \( <F=f, A=a, S=s, H=h, N=n> \)

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

\[
P(f) P(a) P(s|f,a) P(h|s) P(n|h,s)\]

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

• How do we calculate \( P(N=n) \)?

\[
P(N=n) = \sum_{f,a,h,s} P(F=f, A=a, H=h, S=s, N=1) P(f) P(a) P(s|f,a) P(h|s) P(n|h,s)
\]

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 \( P(F,A,S,H,N) \)?

1. Randomly draw a value for \( F = f \)
2. Draw \( r \in [0,1] \) uniformly
3. If \( r < \theta_{F=1} \) then output \( f = 1 \)
4. Draw \( f, a, s, h, i, n \)

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 \( P(F=1|H=1, N=0) \)

\( \rightarrow \) weak but general method for estimating any 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

eg., chain

\[ P(c=1) \]

Complexity 

8 terms 4-way Mult

Variable Elimination

Prob. of marginals: not so easy

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

eg., chain
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 (i.e., 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 \( \{X_2, X_3\} \)
  - 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 (i.e., 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

\[ \text{X1} \rightarrow \text{X2} \rightarrow \text{X3} \]

\[ \text{X1} \rightarrow \text{X4} \]

\[ \text{X3} \rightarrow \text{X4} \]
Easy Network 1: Head to Tail

prove A cond indep of B given C?

\[ p(a|b|c) = p(a|c) p(b|c) \]

\[
p(a|b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)} 
= \frac{p(a)p(c|a)}{p(c)} \cdot \frac{p(b|c)}{p(c)} 
= p(a|c) p(b|c)
\]

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?

\[ p(a|b|c) = p(a|c) p(b|c) \]

let's use \( p(a,b) \) as shorthand for \( p(A=a, B=b) \)
prove A cond indep of B given C?  ie., $p(a,b|c) = p(a|c) p(b|c)$

but true that $p(a,b) = p(a)p(b)$

$\begin{align*}
p(a,b) &= \sum_{c \in C} p(a,b,c) \\
      &= p(a)p(b) \sum_{c \in C} p(c|a,b) \\
      &= p(a)p(b) \sum_{c \in C} p(c|a,b) \\
      &= p(a)p(b) \sum_{c \in C} p(c|a,b) \\
      &= p(a)p(b)
\end{align*}$

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

Summary:
- $p(a,b)=p(a)p(b)$
- $p(a,b|c) \neq p(a|c)p(b|c)$

Explaining away.

e.g.,
- A=earthquake
- B=breakIn
- C=motionAlarm
X and Y are conditionally independent given Z, if and only if 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 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

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

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 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 {}?

X4 indep of X1 given X3?
X4 indep of X1 given {X3, X2}?
X4 indep of X1 given {}?
Markov Blanket

The Markov blanket of a node $x_i$, comprises the set of parents, children and co-parents of the node. It has the property that the conditional distribution of $x_i$, 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 $\theta$ for each CPD

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

   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…
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’