CS 331: Artificial Intelligence
Bayesian Networks

Thanks to Andrew Moore for some course material

Why This Matters

- Bayesian networks have been one of the most important contributions to the field of AI in the last 10-20 years
- Provide a way to represent knowledge in an uncertain domain and a way to reason about this knowledge
- Many applications: medicine, factories, help desks, spam filtering, etc.

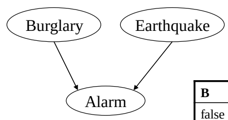
Outline

1. Brief Introduction to Bayesian networks
2. Semantics of Bayesian networks
   - Bayesian networks as a full joint probability distribution
   - Bayesian networks as encoding conditional independence relationships

A Bayesian Network

B	P(B)
false	0.999
true	0.001

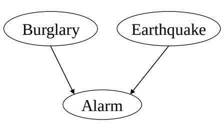
E	P(E)
false	0.998
true	0.002



A Bayesian network is made up of two parts:
1. A directed acyclic graph
2. A set of parameters

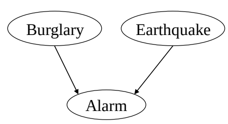
B	E	A	P(A|B,E)
false	false	false	0.999
false	false	true	0.001
false	true	false	0.71
false	true	true	0.29
true	false	false	0.06
true	false	true	0.94
true	true	false	0.05
true	true	true	0.95

A Directed Acyclic Graph



1. A directed acyclic graph:
- The nodes are random variables (which can be discrete or continuous)
- Arrows connect pairs of nodes (X is a parent of Y if there is an arrow from node X to node Y).

A Directed Acyclic Graph

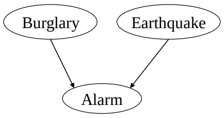


- Intuitively, an arrow from node X to node Y means X has a direct influence on Y (often X has a causal effect on Y)
- Easy for a domain expert to determine these relationships
- The absence/presence of arrows will be made more precise later on

A Set of Parameters

B	P(B)
false	0.999
true	0.001

E	P(E)
false	0.998
true	0.002



B	E	A	P(A|B,E)
false	false	false	0.999
false	false	true	0.001
false	true	false	0.71
false	true	true	0.29
true	false	false	0.06
true	false	true	0.94
true	true	false	0.05
true	true	true	0.95

Each node X_i has a conditional probability distribution $P(X_i \mid Parents(X_i))$ that quantifies the effect of the parents on the node

The parameters are the probabilities in these conditional probability distributions

Because we have discrete random variables, we have conditional probability tables (CPTs)

A Set of Parameters

Conditional Probability Distribution for Alarm

Stores the probability distribution for Alarm given the values of Burglary and Earthquake

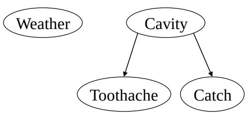
B	E	A	P(A|B,E)
false	false	false	0.999
false	false	true	0.001
false	true	false	0.71
false	true	true	0.29
true	false	false	0.06
true	false	true	0.94
true	true	false	0.05
true	true	true	0.95

For a given combination of values of the parents (B and E in this example), the entries for P(A=true|B,E) and P(A=false|B,E) must add up to 1 e.g. P(A=true|B=false,E=false) + P(A=false|B=false,E=false)=1

If you have a Boolean variable with k Boolean parents, how big is the conditional probability table?

How many entries are independently specifiable?

Bayesian Network Example



Things of note:

- Weather is independent of the other variables
- Toothache and Catch are conditionally independent given Cavity (this is represented by the fact that there is no link between Toothache and Catch and by the fact that they have Cavity as a parent)

Bayesian Network Example

Coin	P(Coin)
tails	0.5
heads	0.5

Coin	Card	P(Card | Coin)
tails	black	0.6
tails	red	0.4
heads	black	0.3
heads	red	0.7

Card	Candy	P(Candy | Card)
black	1	0.5
black	2	0.2
black	3	0.3
red	1	0.1
red	2	0.3
red	3	0.6

What does the DAG for this Bayes net look like?

Semantics of Bayesian Networks

Bayes Nets Formalized

A Bayes net (also called a belief network) is an augmented directed acyclic graph, represented by the pair V , E where:

- V is a set of vertices.
- E is a set of directed edges joining vertices. No loops of any length are allowed.

Each vertex in V contains the following information:

- The name of a random variable
- A probability distribution table indicating how the probability of this variable's values depends on all possible combinations of parental values.

Semantics of Bayesian Networks

Two ways to view Bayes nets:

1. A representation of a joint probability distribution
2. An encoding of a collection of conditional independence statements

A Representation of the Full Joint Distribution

- We will use the following abbrevations:
  - $P(x_1, \ldots, x_n)$ for $P(X_1 = x_1 \wedge \ldots \wedge X_n = x_n)$
  - *parents*(X_i) for the values of the parents of X_i
- From the Bayes net, we can calculate:

$$P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i \mid parents(X_i))$$

The Full Joint Distribution

$$P(x_1,\ldots,x_n)$$
$$= P(x_n \mid x_{n-1},\ldots,x_1)P(x_{n-1},\ldots,x_1) \quad \text{(Chain Rule)}$$
$$= P(x_n \mid x_{n-1},\ldots,x_1)P(x_{n-1} \mid x_{n-2},\ldots,x_1)P(x_{n-2},\ldots,x_1) \quad \text{(Chain Rule)}$$
$$= P(x_n \mid x_{n-1},\ldots,x_1)P(x_{n-1} \mid x_{n-2},\ldots,x_1)\ldots P(x_2 \mid x_1)P(x_1) \quad \text{(Chain Rule)}$$
$$= \prod_{i=1}^{n} P(x_i \mid x_{i-1},\ldots,x_1)$$
$$= \prod_{i=1}^{n} P(x_i \mid parents(x_i))$$

We'll look at this step more closely

The Full Joint Distribution

$$\prod_{i=1}^{n} P(x_i \mid x_{i-1},\ldots,x_1) = \prod_{i=1}^{n} P(x_i \mid parents(x_i))$$

To be able to do this, we need two things:

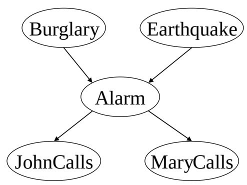
1. *Parents*(X_i) $\subseteq \{X_{i-1}, \ldots, X_1\}$

   This is easy – we just label the nodes according to the partial order in the graph

2. We need X_i to be conditionally independent of its predecessors given its parents

   This can be done when constructing the network. Choose parents that directly influence X_i.

Example



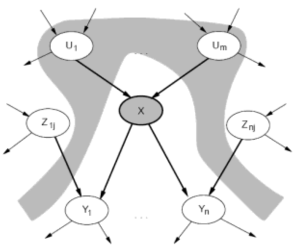
P(*JohnCalls, MaryCalls, Alarm, Burglary, Earthquake*)

= ***P***(*JohnCalls* | *Alarm*) ***P***(*MaryCalls* | *Alarm*) ***P***(*Alarm* | *Burglary, Earthquake*) ***P***(*Burglary*) ***P***(*Earthquake*)

Conditional Independence

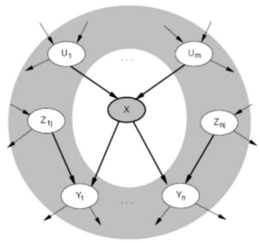
We can look at the actual graph structure and determine conditional independence relationships.

1. A node (X) is conditionally independent of its non-descendants (Z_{1j}, Z_{nj}), given its parents (U_1, U_m).



Conditional Independence

2. Equivalently, a node (X) is conditionally independent of all other nodes in the network, given its parents (U_1, U_m), children (Y_1, Y_n), and children's parents (Z_{1j},Z_{nj}) – that is, given its Markov blanket



Conditional Independence

- Previously, we conditioned on either the parent values or the values of the nodes in the Markov blanket
- There is a much more general topological criterion called d-separation
- d-separation determines whether a set of nodes X is independent of another set Y given a third set E
- You should use d-separation for determining conditional independence

D-separation

- We will use the notation I(X, Y | E) to mean that X and Y are conditionally independent given E
- Theorem [Verma and Pearl 1988]:
  If a set of evidence variables E d-separates X and Y in the Bayesian Network's graph, then I(X, Y | E)
- d-separation can be determined in linear time using a DFS-like algorithm

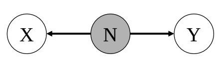
D-separation

- Let evidence nodes $E \subseteq V$ (where V are the vertices or nodes in the graph), and X and Y be distinct nodes in $V - E$.
- We say X and Y are d-separated by E in the Bayesian network if every undirected path between X and Y is blocked by E.
- What does it mean for a path to be blocked? There are 3 cases…

Case 1

There exists a node N on the path such that

- It is in the evidence set E (shaded grey)
- The arcs putting N in the path are "tail-to-tail".

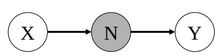


The path between X and Y is blocked by N

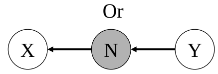
Case 2

There exists a node N on the path such that

- It is in the evidence set E
- The arcs putting N in the path are "tail-to-head".



Or

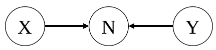


The path between X and Y is blocked by N

Case 3

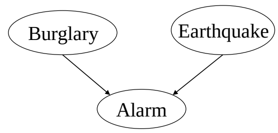
There exists a node N on the path such that

- It is NOT in the evidence set *E* (not shaded)
- Neither are any of its descendants
- The arcs putting *N* in the path are "head-to-head".



The path between X and Y is blocked by N (Note N is not in the evidence set)

Case 3 (Explaining Away)

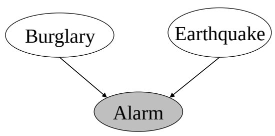


Your house has a twitchy burglar alarm that is also sometimes triggered by earthquakes

Given no evidence about Alarm, Burglary and Earthquake are independent i.e. learning about an earthquake when you know nothing about the status of your alarm doesn't give you any information about the burglary and vice versa

26

Case 3 (Explaining Away)



Suppose that while you are on vacation, your neighbor lets you know your alarm went off. If you knew that a medium-sized earthquake happened, then you're probably relieved that it's probably not a burglar

The earthquake "explains away" the hypothetical burglar

This means that Burglary and Earthquake are not independent given Alarm.

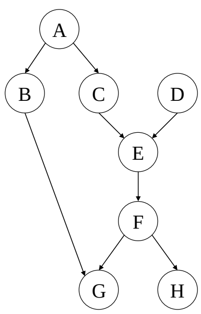
27

d-separation Recipe

- To determine if I(X, Y | E), ignore the directions of the arrows, find all paths between X and Y
- Now pay attention to the arrows. Determine if the paths are blocked according to the 3 cases
- If all the paths are blocked, X and Y are d-separated given E
- Which means they are conditionally independent given E

28

D-separation Examples



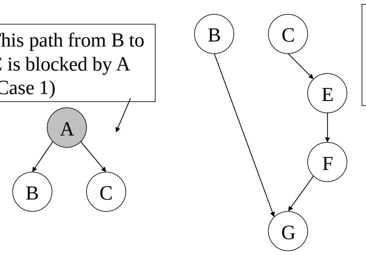
I(B, C | A)?

29

D-separation Examples

I(B, C | A)? Yes. Notice the two (undirected) paths between B and C

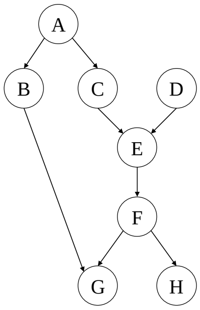
This path from B to C is blocked by A (Case 1)



This path from B to C is blocked by G, which is not in the evidence set (Case 3)

30

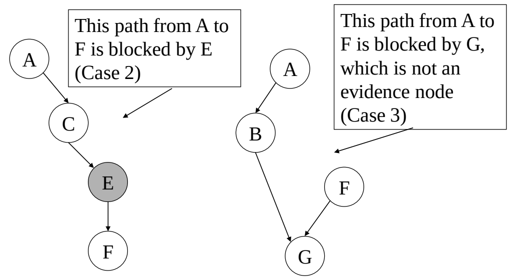
D-separation Examples



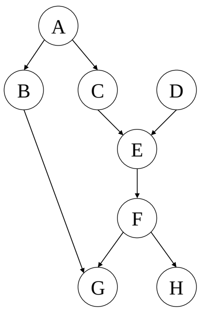
I(A , F | E)?

D-separation Examples

I(A , F | E)? Yes



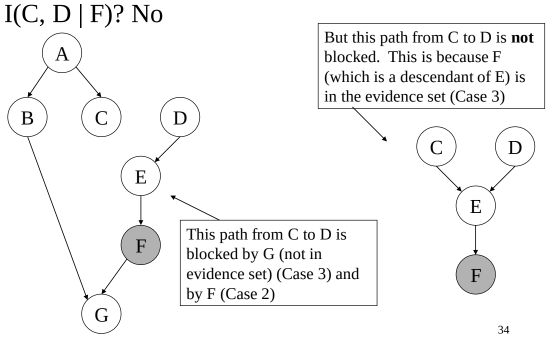
D-separation Examples



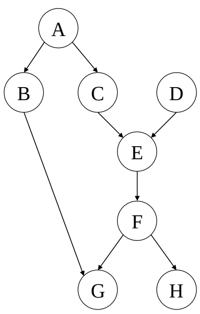
I(C, D | F)?

D-separation Examples

I(C, D | F)? No



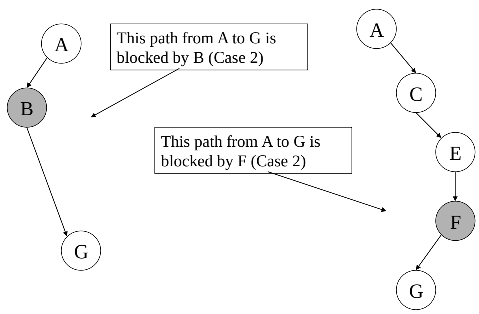
D-separation Examples



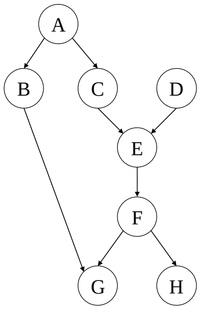
I(A, G | {B, F})?

D-separation Examples

I(A, G | {B, F})? Yes



Practice



I(B, D | {C, H})?

Conditional Independence

- Note: D-separation only finds random variables that are conditionally independent based on the topology of the network
- Some random variables that are not d-separated may still be conditionally independent because of the probabilities in their CPTs

What You Should Know

- How to compute the joint probability distribution from a Bayesian network
- How to determine conditional independence relationships using d-separation