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

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

<table>
<thead>
<tr>
<th>B</th>
<th>P(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>false</td>
<td>0.999</td>
</tr>
<tr>
<td>true</td>
<td>0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E</th>
<th>P(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>false</td>
<td>0.998</td>
</tr>
<tr>
<td>true</td>
<td>0.002</td>
</tr>
</tbody>
</table>

| Alarm | 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).

   - 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

<table>
<thead>
<tr>
<th>B</th>
<th>P(B)</th>
<th>E</th>
<th>P(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>false</td>
<td>0.999</td>
<td>false</td>
<td>0.998</td>
</tr>
<tr>
<td>true</td>
<td>0.001</td>
<td>true</td>
<td>0.002</td>
</tr>
</tbody>
</table>

| 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 $P(X_i | \text{Parents}(X_i))$ 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

| 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  |

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

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

$P(A=\text{true}|B=\text{false},E=\text{false}) + P(A=\text{false}|B=\text{false},E=\text{false})=1$

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

<table>
<thead>
<tr>
<th>Coin</th>
<th>P(Coin)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tails</td>
<td>0.5</td>
</tr>
<tr>
<td>heads</td>
<td>0.5</td>
</tr>
</tbody>
</table>

| 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?
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 abbreviations:
  – \( P(x_1, \ldots, x_n) \) for \( P(X_1 = x_1 \land \ldots \land X_n = x_n) \)
  – \( \text{parents}(X_i) \) for the values of the parents of \( X_i \)
• From the Bayes net, we can calculate:

\[
P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | \text{parents}(X_i))
\]
The Full Joint Distribution

\[ P(x_1, \ldots, x_n) \]

\[ = P(x_n | x_{n-1}, \ldots, x_1)P(x_{n-1}, \ldots, x_1) \quad \text{(Chain Rule)} \]

\[ = P(x_n | x_{n-1}, \ldots, x_1)P(x_{n-1} | x_{n-2}, \ldots, x_1)P(x_{n-2}, \ldots, x_1) \quad \text{(Chain Rule)} \]

\[ = P(x_n | x_{n-1}, \ldots, x_1)P(x_{n-1} | x_{n-2}, \ldots, x_1) \cdots P(x_2 | x_1)P(x_1) \]

\[ = \prod_{i=1}^{n} P(x_i | x_{i-1}, \ldots, x_1) \quad \text{(Chain Rule)} \]

\[ = \prod_{i=1}^{n} P(x_i | \text{parents}(x_i)) \]

We’ll look at this step more closely

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

1. \text{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

\[
P(\text{JohnCalls}, \text{MaryCalls}, \text{Alarm}, \text{Burglary}, \text{Earthquake})
\]

\[
= P(\text{JohnCalls} \mid \text{Alarm}) \ P(\text{MaryCalls} \mid \text{Alarm}) \ P(\text{Alarm} \mid \text{Burglary, Earthquake}) \ P(\text{Burglary}) \ P(\text{Earthquake})
\]

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 \((Z_{ij}, Z_{nj})\), given its parents \((U_1, U_m)\).
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 \((Z_{i1j}, Z_{nj})\) — that is, given its Markov blanket.

- 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 $I(X, Y \mid E)$ 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 $I(X, Y \mid E)$.
• 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”.

![Diagram of Case 1]

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”.

![Diagram of Case 2]

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
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.

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
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)
D-separation Examples

I(A, F | E)?

This path from A to F is blocked by E (Case 2)

This path from A to F is blocked by G, which is not an evidence node (Case 3)

I(A, F | E)? Yes
D-separation Examples

I(C, D | F)?

But this path from C to D is not blocked. This is because F (which is a descendant of E) is in the evidence set (Case 3).

This path from C to D is blocked by G (not in evidence set) (Case 3) and by F (Case 2).
D-separation Examples

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

This path from A to G is blocked by B (Case 2)

This path from A to G is blocked by F (Case 2)
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