BELIEF NETWORKS I

Lecture 23
(Chapter 15, 12)
Artificial Intelligence I
Autumn 2001
Henry Kautz

Independence

Two random variables $A,B$ are (absolutely) independent iff

$$P(A|B) = P(A)$$

or

$$P(A,B) = P(A|B)P(B) = P(A)P(B)$$

e.g., $A$ and $B$ are two coin tosses

If $n$ Boolean variables are independent, the full joint is

$$P(X_1, \ldots, X_n) = \prod P(X_i)$$

hence can be specified by just $n$ numbers

Absolute independence is a very strong requirement, seldom met

Conditional independence

Consider the dentist problem with three random variables:

Toothache, Cavity, Catch (steel probe catches in my tooth)

The full joint distribution has $2^3 - 1 = 7$ independent entries

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

1. $P(\text{Catch}|\text{Toothache}, \text{Cavity}) = P(\text{Catch}|\text{Cavity})$

i.e., Catch is conditionally independent of Toothache given Cavity

The same independence holds if I haven't got a cavity:

2. $P(\text{Catch}|\text{Toothache}, \neg\text{Cavity}) = P(\text{Catch}|\neg\text{Cavity})$
Conditional independence contd.

Product rule:

\[ P(\text{Toothache, Catch, Cavity}) = P(\text{Cavity})P(\text{Catch|Cavity})P(\text{Toothache|Catch, Cavity}) \]

Independence:

\[ P(\text{Toothache, Catch, Cavity}) = P(\text{Cavity})P(\text{Catch|Cavity})P(\text{Toothache|Cavity}) \]

Full joint distribution can now requires only 5 independent numbers (instead of 7)

Belief networks

A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions.

Syntax:
- a set of nodes, one per variable
- a directed, acyclic graph (link \(\approx\) “directly influences”)
- a conditional distribution for each node given its parents:
\[ P(X_i|\text{Parents}(X_i)) \]

In the simplest case, conditional distribution represented as a conditional probability table (CPT)

Example

I’m at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn’t call. Sometimes it’s set off by minor earthquakes. Is there a burglar?

Variables: Burglar, Earthquake, Alarm, JohnCalls, MaryCalls

Network topology reflects “causal” knowledge:

```
  Burglar  Earthquake
   \ \  /  \
  \   /   \
 Alarm  
   \ /   \\  \
  \       / \\
 JohnCalls  MaryCalls
```

Note: \( \leq k \) parents \( \Rightarrow O(d^k n) \) numbers vs. \( O(d^n) \)

Semantics

“Global” semantics defines the full joint distribution as the product of the local conditional distributions:

\[ P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i)) \]

e.g., \( P(J \land M \land A \land \neg B \land \neg E) \) is given by ??
Semantics

"Global" semantics defines the full joint distribution as the product of the local conditional distributions:

\[ P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i)) \]

e.g., \( P(J \land M \land A \land \neg B \land \neg E) \) is given by?

\[ = P(\neg B)P(\neg E)P(A|B \land \neg E)P(J|A)P(M|A) \]

"Local" semantics: each node is conditionally independent of its non-descendants given its parents

Theorem: Local semantics \( \iff \) global semantics

---

Markov blanket

Each node is conditionally independent of all others given its Markov blanket: parents + children + children's parents

---

Constructing belief networks

Need a method such that a series of locally testable assertions of conditional independence guarantees the required global semantics

1. Choose an ordering of variables \( X_1, \ldots, X_n \)
2. For \( i = 1 \) to \( n \)
   - add \( X_i \) to the network
   - select parents from \( X_1, \ldots, X_{i-1} \) such that
     \[ P(X_i|\text{Parents}(X_i)) = P(X_i|X_1, \ldots, X_{i-1}) \]

This choice of parents guarantees the global semantics:

\[ P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i)) \text{ (chain rule)} \]

\[ = \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i)) \text{ by construction} \]

---

Example

Suppose we choose the ordering \( M, J, A, B, E \)

\[ P(J|M) = P(J)? \]
\[ P(A | J, M) = P(A | J) \? \quad P(A | J, M) = P(A) \? \]
\[ P(B | A, J, M) = P(B | A) \? \quad P(B | A, J, M) = P(B) \? \]
\[ P(E | B, A, J, M) = P(E | A) \? \quad P(E | B, A, J, M) = P(E | A, B) \? \]
\[ \text{Earthquake} \]
\[ \text{Alarm} \]
\[ \text{Burglary} \]
Example: Car diagnosis

Initial evidence: engine won’t start
Testable variables (thin ovals), diagnosis variables (thick ovals)
Hidden variables (shaded) ensure sparse structure, reduce parameters

Compact conditional distributions

CPT grows exponentially with no. of parents
CPT becomes infinite with continuous valued parent or child
Solution: canonical distributions that are defined compactly

Deterministic nodes are the simplest case:

\[ X = f(Parents(X)) \] for some function \( f \)

E.g., Boolean functions

NorthAmerican \( \Leftrightarrow \) Canadian \( \lor \) US \( \lor \) Mexican

E.g., numerical relationships among continuous variables

\[ \frac{\partial \text{Level}}{\partial t} = \text{inflow} + \text{precipitation} \cdot \text{outflow} \cdot \text{evaporation} \]

Example: Car insurance

Predict claim costs (medical, liability, property)
given data on application form (other unshaded nodes)

Compact conditional distributions contd.

Noisy-\text{OR} distributions model multiple noninteracting causes
1) Parents \( U_1, \ldots, U_k \) include all causes (can add leak node)
2) Independent failure probability \( q_i \) for each cause alone

\[ P(X|U_1, \ldots, U_j, \neg U_{j+1}, \ldots, \neg U_k) = 1 - \prod_{i=1}^{k} q_i \]

<table>
<thead>
<tr>
<th>Cold</th>
<th>Flu</th>
<th>Malaria</th>
<th>( P(\text{Fever}) )</th>
<th>( P(\neg \text{Fever}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>0.98</td>
<td>0.02 = 0.2 \times 0.1</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>0.94</td>
<td>0.06 = 0.6 \times 0.1</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>0.88</td>
<td>0.12 = 0.6 \times 0.2</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.988</td>
<td>0.012 = 0.6 \times 0.2 \times 0.1</td>
</tr>
</tbody>
</table>

Number of parameters linear in number of parents
Naive Bayes

Very simple but surprisingly useful model:

All findings conditionally independent given cause

\[ P(e_1, \ldots e_n | h) = \prod_{i=1}^{n} P(e_i | h) \]

Therefore the cause \( h \) that maximizes \( P(h | e_1, \ldots e_n) \) is

just the one that maximizes \( P(h) \prod_{i=1}^{n} P(e_i | h) \) !

**Pathfinder**: first BN medical diagnosis system. Using naive Bayes outperformed doctors! Full BN version saved 1 life in 1000.

1) Better at incorporating prior probability of different diseases,
2) Uses all evidence — humans focus on only 7–9 pieces

**CPCS**: internal diseases — 448 nodes, 906 edges