**Bayesian 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 (\text{link} \Rightarrow \text{"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) giving the distribution over \( X_i \) for each combination of parent values.

**Example**

Topology of network encodes conditional independence assertions:

- Weather
- Cavity
- Toothache
- Catch

*Weather* is independent of the other variables

*Toothache* and *Catch* are conditionally independent given *Cavity*

---

**Outline**

- Syntax
- Semantics
- Parameterized distributions

---

**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:
- A burglar can set the alarm off
- An earthquake can set the alarm off
- The alarm can cause Mary to call
- The alarm can cause John to call

---

**Example contd.**

```
Burglary P(B)
T   .01
F   .99

Earthquake P(E)
T   .02
F   .98

Alarm P(A|B,E)
B   E   P(A|B,E)
T   T   .95
T   F   .05
F   T   .29
F   F   .01

JohnCalls P(J|A)
A   B   P(J|A)
T   T   .90
T   F   .05
F   T   .90
F   F   .05

MaryCalls P(M|A)
A   B   P(M|A)
T   T   .70
T   F   .01
F   T   .01
F   F   .30
```
**Compactness**

A CPT for Boolean $X_i$ with $k$ Boolean parents has $2^k$ rows for the combinations of parent values.

Each row requires one number $p$ for $X_i = \text{true}$ (the number for $X_i = \text{false}$ is just $1 - p$).

If each variable has no more than $k$ parents, the complete network requires $O(n \cdot 2^k)$ numbers.

I.e., grows linearly with $n$, vs. $O(2^n)$ for the full joint distribution.

For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$).

---

**Local semantics**

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

**Theorem:** Local semantics $\iff$ global semantics

---

**Global 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., $P(j \land m \land a \land \neg b \land \neg e) = P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e)$

---

**Markov blanket**

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

---

**Constructing Bayesian 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:

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

(chain rule)

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

(by construction)
Example

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

\[
\begin{align*}
P(J|M) &= P(J) \quad \text{No} \\
P(A|J, M) &= P(A|J) \quad P(A|J, M) = P(A) \quad \text{No} \\
P(B|A, J, M) &= P(B|A) \quad \text{Yes} \\
P(E|B, A, J, M) &= P(E|A) \quad \text{No} \\
P(E|B, A, J, M) &= P(E|A, B) \quad \text{Yes}
\end{align*}
\]

Example

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

\[
\begin{align*}
P(J|M) &= P(J) \\
P(A|J, M) &= P(A|J) \\
P(A|J, M) &= P(A) \\
P(B|A, J, M) &= P(B|A) \\
P(E|B, A, J, M) &= P(E|A) \\
P(E|B, A, J, M) &= P(E|A, B)
\end{align*}
\]

Example

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

\[
\begin{align*}
P(J|M) &= P(J) \quad \text{No} \\
P(A|J, M) &= P(A|J) \\
P(A|J, M) &= P(A) \\
P(B|A, J, M) &= P(B|A) \\
P(B|A, J, M) &= P(B)
\end{align*}
\]

Example contd.

Deciding conditional independence is hard in noncausal directions

(Causal models and conditional independence seem hardwired for humans)

Assessing conditional probabilities is hard in noncausal directions

Network is less compact: \( 1 + 2 + 4 + 2 + 4 = 13 \) numbers needed
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
\( X = f(\text{Parents}(X)) \) for some function \( f \)

E.g., Boolean functions

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

E.g., numerical relationships among continuous variables

\( \frac{\partial \text{Level}}{\partial t} = \text{inflow} + \text{precipitation} - \text{outflow} - \text{evaporation} \)

Continuous child variables

Need one conditional density function for child variable given continuous parents, for each possible assignment to discrete parents

Most common is the linear Gaussian model, e.g.:

\[
P(Cost = e | Harvest = h, Subsidy = \text{true}) = \mathcal{N}(a_h + b_t, \sigma_t) \\
= \frac{1}{\sigma_t \sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( \frac{e - (a_h + b_t)}{\sigma_t} \right)^2 \right)
\]

Mean \( Cost \) varies linearly with \( Harvest \), variance is fixed

Linear variation is unreasonable over the full range

but works OK if the likely range of \( Harvest \) is narrow
Continuous child variables

Discrete variable contd.

Sigmoid (or logit) distribution also used in neural networks:

\[ P(Buy^? = \text{true} \mid \text{Cost} = c) = \frac{1}{1 + \exp\left[-\frac{c - \mu}{\sigma}\right]} \]

Sigmoid has similar shape to probit but much longer tails:

Discrete variable w/ continuous parents

Summary

Bayes nets provide a natural representation for (causally induced)
conditional independence

Topology + CPTs = compact representation of joint distribution

Generally easy for (non)experts to construct

Canonical distributions (e.g., noisy-OR) = compact representation of CPTs

Continuous variables ⇒ parameterized distributions (e.g., linear Gaussian)

Why the probit?

1. It's sort of the right shape

2. Can view as hard threshold whose location is subject to noise

\[\text{Cost} \quad \text{Cost} + \text{Noise} \quad \text{Buys?}\]