Bayesian networks: Chapter 14.1–3

July 25, 2021

Outline

- Syntax
- Semantics
- Parameterized distributions

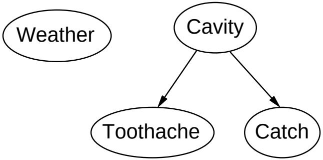
Bayesian networks

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 (link $\approx$ "directly influences")
  - a conditional distribution for each node given its parents $P(X_i | Parents(X_i))$.
    - In the simplest case, the 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:



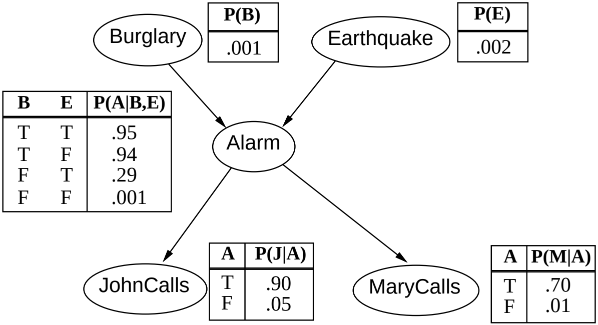
- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

Example: Burglary - Earthquake

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?

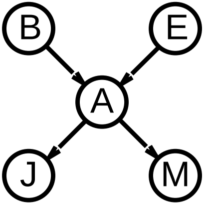
- Variables: *Burglar*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
- The network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example contd.



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 $X_i = true$ (the number for $X_i = false$ 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., it grows linearly with n, vs. $O(2^n)$ for the full joint distribution)
- For the B-E net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)

Global semantics

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

P(X1=x1,..., Xn = xn)

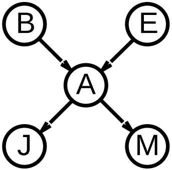
P(Xi=xi|parents(Xi))

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

P(J called, Mary calls, Alarm goes off, Not burglary, Not eartquake)

e.g.,

$$P(j \land m \land a \land \neg b \land \neg e) = \ldots.$$

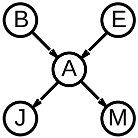


Global semantics

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

e.g.,

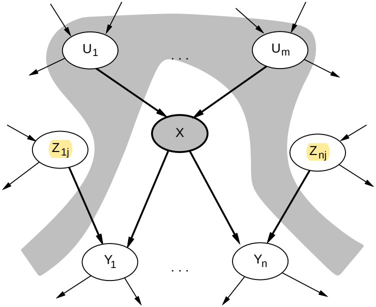
$$\begin{aligned} P(j \land m \land a \land \neg b \land \neg e) &= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \\ &= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ &\approx 0.00063 \end{aligned}$$



Local semantics

- **Local** semantics: each node is conditionally independent of its nondescendents given its parents

$P(X \mid Zi, U1,\ldots,UM) = P(X \mid U1,\ldots,UM)$



Theorem: The *Local Semantics* $\equiv$ *Global Semantics*

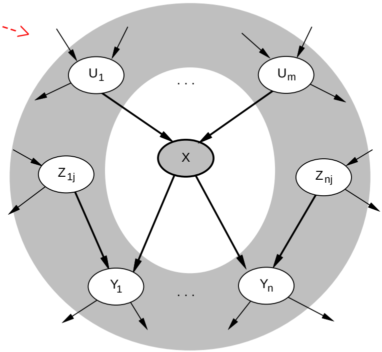
Markov blanket

Markov property states that "the present given the past depends only on the immediate past"

$P(X_t \mid X_{t-1}, X_{t-2}, \ldots, X_1) = P(X_t \mid X_{t-1})$

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

T1



$P(X \mid T1, U1,\ldots,Um,Y1,\ldots,Yn,Z1j\ldots,Znj) = P(X \mid U1,\ldots,Um,Y1,\ldots,Yn,Z1j\ldots,Znj)$

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
   2.1 add X_i to the network
   2.2 select parents from $\{X_1, \ldots, X_{i-1}\}$ such that

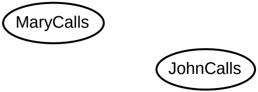
$$P(X_i | Parents(X_i)) = P(X_i | X_1, \ldots, X_{i-1})$$

This choice of parents guarantees the global semantics:

$$\begin{aligned} P(X_1, \ldots, X_n) &= \textstyle\prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1}) \text{(chain rule)} \\ &= \textstyle\prod_{i=1}^{n} P(X_i | Parents(X_i)) \text{(by construction)} \end{aligned}$$

Example

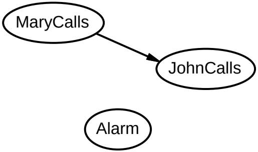
Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$?

Example

Suppose we choose the ordering M, J, A, B, E

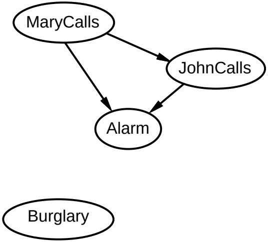


$$P(J|M) = P(J)? \text{ No}$$

$$P(A|J,M) = P(A|J)? \quad P(A|J,M) = P(A)?$$

Example

Suppose we choose the ordering M, J, A, B, E



$$P(J|M) = P(J) \text{ No}$$

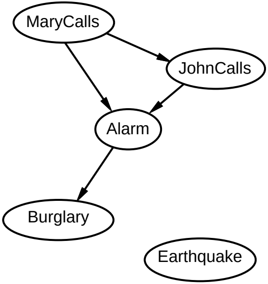
$$P(A|J,M) = P(A|J)? \;\; P(A|J,M) = P(A)? \text{ No}$$

$$P(B|A,J,M) = P(B|A)?$$

$$P(B|A,J,M) = P(B)?$$

Example

Suppose we choose the ordering M, J, A, B, E



$$P(J|M) = P(J) \text{ No}$$

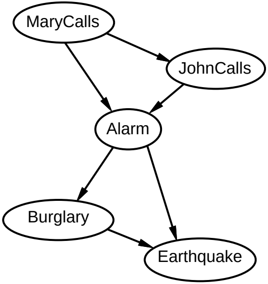
$$P(A|J,M) = P(A|J)? \;\; P(A|J,M) = P(A)? \text{ No}$$

$$P(B|A,J,M) = P(B|A)? \text{ Yes}$$

$$P(B|A,J,M) = P(B)? \text{ No}$$

Example

Suppose we choose the ordering M, J, A, B, E



$$P(J|M) = P(J)\text{No}$$

$$P(A|J, M) = P(A|J)? \;\; P(A|J, M) = P(A)?\text{No}$$

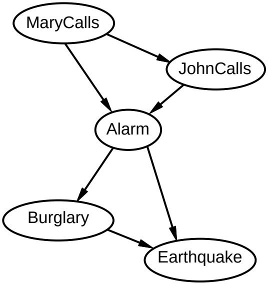
$$P(B|A, J, M) = P(B|A)?\text{YES}$$

$$P(B|A, J, M) = P(B)?\text{No}$$

$$P(E|B, A, J, M) = P(E|A)?\text{No}$$

$$P(E|B, A, J, M) = P(E|A, B)\text{YES}$$

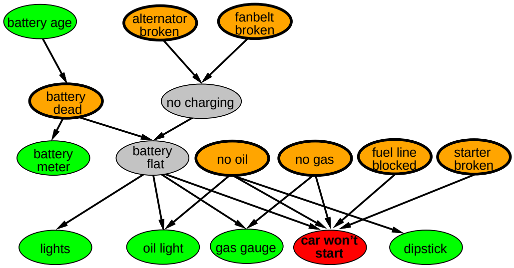
Example contd.



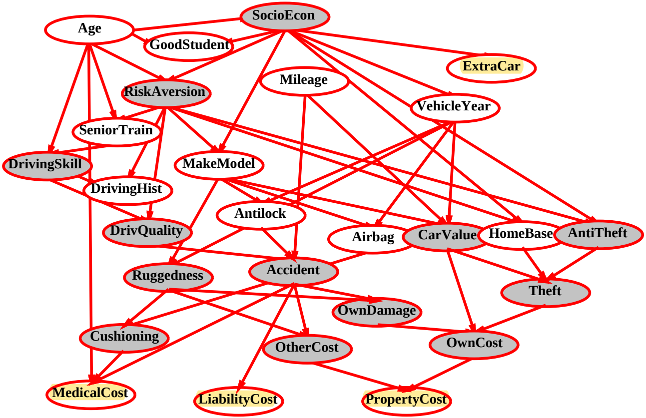
- Deciding conditional independence is hard in non-causal directions
- (Causal models and conditional independence seem hardwired for humans!???)
- Assessing conditional probabilities is hard in non-causal directions
- Network is less compact: need $1 + 2 + 4 + 2 + 4 = 13$ numbers

Example: Car diagnosis

- Initial evidence: car won't start
- Testable variables (green), "broken, so fix it" variables (orange)
- Hidden variables (gray) ensure sparse structure,reduce parameters



Example: Car insurance



Compact conditional distributions

- CPT grows exponentially with number of parents
- CPT becomes infinite with continuous-valued parent or child
- Solution: **canonical** distributions that are defined compactly
- **Deterministic** nodes are the simplest case:

$$X = f(Parents(X)), \text{ for some function } f$$

E.g., Boolean functions

$$NorthAmerican \equiv Canadian \vee US \vee Mexican$$

E.g., numerical relationships among continuous variables

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

Compact conditional distributions contd.

- **Noisy-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

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

$Cold$	Flu	$Malaria$	$P(Fever)$	$P(\neg Fever)$
F	F	F	**0.0**	1.0
F	F	T	0.9	**0.1**
F	T	F	0.8	**0.2**
F	T	T	0.98	$0.02 = 0.2 \times 0.1$
T	F	F	0.4	**0.6**
T	F	T	0.94	$0.06 = 0.6 \times 0.1$
T	T	F	0.88	$0.12 = 0.6 \times 0.2$
T	T	T	0.988	$0.012 = 0.6 \times 0.2 \times 0.1$

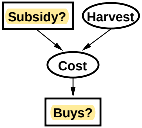
- Number of parameters **linear** in number of parents

Hybrid (discrete+continuous) networks

- Discrete (*Subsidy*? and *Buys*?);
- Continuous (*Harvest* and *Cost*)

P(Cost = c | Subsidy, Harvest)

CPT for Cost
P(Cost | h1, true)
P(Cost | h1, false)
P(Cost | h2, true)
P(Cost |h2, false)
P(Cost| h3, true)
P(Cost | h3 ,false)



Harvest = h1, h2, h3
Subsidy - true/false
Cost=c1, c2, c3

- **Option 1:** discretization—possibly large errors, large CPTs
- **Option 2:** finitely parameterized canonical families
  1. Continuous variable, discrete+continuous parents (e.g., *Cost*)
  2. Discrete variable, continuous parents (e.g., *Buys*?)

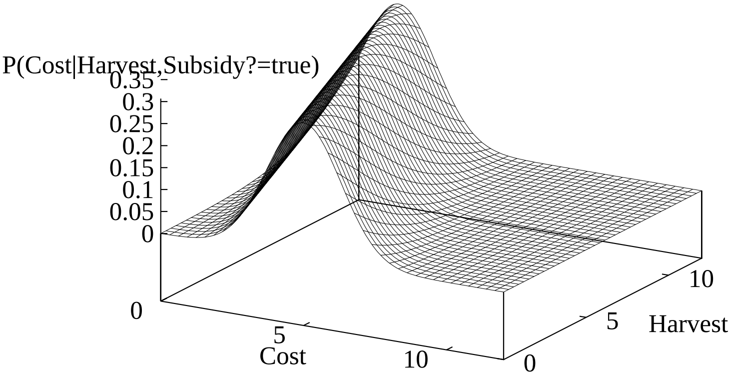
Continuous child variables

- 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.,:

$$
\begin{aligned}
& P(Cost = c | Harvest = h, Subsidy? = true) = \\
& = N(a_t h + b_t, \sigma_t)(c) = \\
& = \frac{1}{\sigma_t \sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{c-(a_t h + b_t)}{\sigma_t}\right)^2\right)
\end{aligned}
$$

- 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

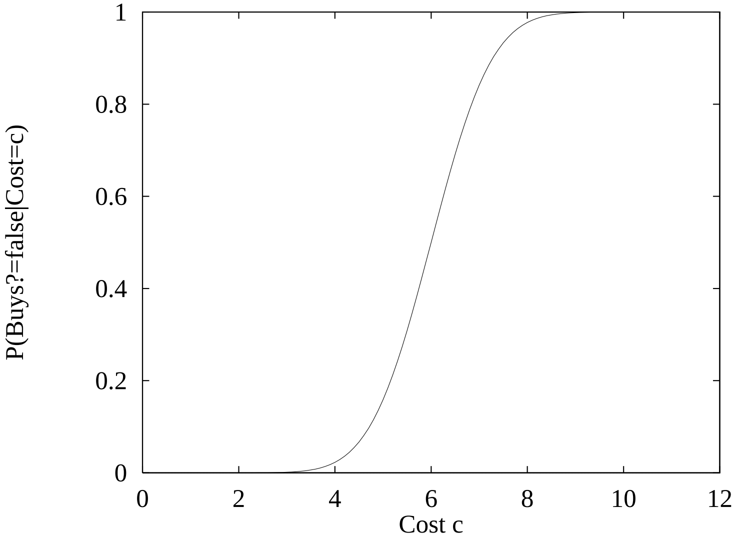
Continuous child variables



- All-continuous network with LG distributions $\implies$ full joint distribution is a multivariate Gaussian

Discrete variable w/ continuous parents

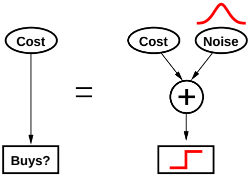
- Probability of *Buys*? given *Cost* should be a *soft* threshold:



- **Probit** distribution uses integral of Gaussian:

Why the probit?

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

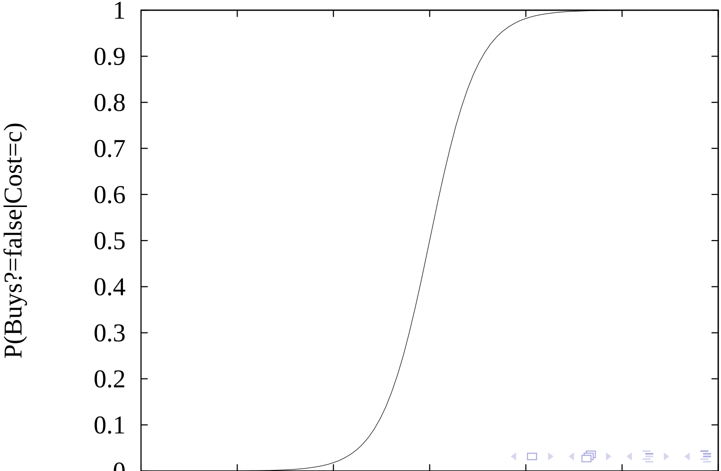


Discrete variable contd.

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

$$P(Buys? = true|Cost = c) = \frac{1}{1 + exp(-2\frac{-c+\mu}{\sigma})}$$

Sigmoid has similar shape to probit but much longer tails:



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 $\rightarrow$ parameterized distributions (e.g., linear Gaussian)