Uncertain Knowledge

Bùi Tiến Lên

01/09/2019

KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Contents

1. Quantifying Uncertainty

2. Bayesian Networks

Quantifying Uncertainty

Uncertainty

- Let action A_t = "leave for airport t minutes before flight". Will A_t get me there on time?
- Problems:
  1. partial observability (road state, other drivers' plans, etc.)
  2. noisy sensors (KCBS traffic reports)
  3. uncertainty in action outcomes (flat tire, etc.)
  4. immense complexity of modelling and predicting traffic
- Hence a purely logical approach either
  1. risks falsehood: "A_{25} will get me there on time" or
  2. leads to conclusions that are too weak for decision making: "A_{25} will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact etc etc."

  (A_{1440} might reasonably be said to get me there on time but I'd have to stay overnight in the airport ...)

Methods for handling uncertainty

- **Default** or **nonmonotonic logic**
  - Assume my car does not have a flat tire
  - Assume A_{25} works unless contradicted by evidence
  - **Issues**: What assumptions are reasonable? How to handle contradiction?
- **Rules with fudge factors**
  - $A_{25} \mapsto_{0.3} AtAirportOnTime$
  - $Sprinkler \mapsto_{0.99} WetGrass$
  - $WetGrass \mapsto_{0.7} Rain$
  - **Issues**: Problems with combination, e.g., *Sprinkler* causes *Rain*?
- **Probability**
  - Given the available evidence, A_{25} will get me there on time with probability 0.04
- **Fuzzy logic** handles *degree of truth* NOT uncertainty
  - e.g., *WetGrass* is true to degree 0.2

Probability

- Probabilistic assertions *summarize* effects of
  - **laziness**: failure to enumerate exceptions, qualifications, etc.
  - **ignorance**: lack of relevant facts, initial conditions, etc.
- **Subjective** or **Bayesian** probability
  - Probabilities relate propositions to one's own state of knowledge

    $$P(A_{25}|\text{no reported accidents}) = 0.06$$

    These are *not* claims of a "probabilistic tendency" in the current situation (but might be learned from past experience of similar situations)
  - Probabilities of propositions change with new evidence

    $$P(A_{25}|\text{no reported accidents}, 5 \text{ a.m.}) = 0.15$$

Uncertainty and rational decisions

- Suppose I believe the following:

$$
\begin{aligned}
P(A_{25} \text{ gets me there on time} | \ldots) &= 0.04 \\
P(A_{90} \text{ gets me there on time} | \ldots) &= 0.70 \\
P(A_{120} \text{ gets me there on time} | \ldots) &= 0.95 \\
P(A_{1440} \text{ gets me there on time} | \ldots) &= 0.9999
\end{aligned}
$$

  Which action to choose?
- Depends on my **preferences** for missing flight vs. airport cuisine, etc.
- **Utility theory** is used to represent and infer preferences

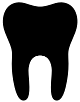
  Decision theory = utility theory + probability theory

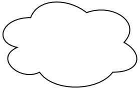
Uncertainty and rational decisions (cont.)

```
function DT-AGENT(percept) returns an action
persistent: belief_state, probabilistic beliefs about
                        the current state of the world
            action, the agent's action
  update belief_state based on action and percept
  calculate outcome probabilities for actions,
    given action descriptions and current belief_state
  select action with highest expected utility
    given probabilities of outcomes and utility information
return action
```

Examples

1. Diagnosing a dental patient's toothache with four random variables: *Toothache*, *Cavity*, *Catch*, *Weather*
2. Alarm problem has three variables: *Earthquake*, *Burglary*, *Alarm*





Syntax for propositions

- **Propositional** or **Boolean** random variables
  - e.g., *Cavity* (do I have a cavity?)
  - $\mathit{Cavity} = \mathit{true}$ is a proposition (also written *cavity*)
- **Discrete** random variables (**finite** or **infinite**)
  - e.g., *Weather* is one of $\langle \mathit{sunny}, \mathit{rain}, \mathit{cloudy}, \mathit{snow} \rangle$
  - $\mathit{Weather} = \mathit{rain}$ is a proposition
  - Values must be exhaustive and mutually exclusive
- **Continuous** random variables (**bounded** or **unbounded**)
  - e.g., $\mathit{Temp} = 21.6$; also allow, e.g., $\mathit{Temp} < 22.0$.
- Arbitrary Boolean combinations of basic propositions

Degrees of belief or Probability

Concept 1

- A **degree of belief** or **probability** in [0,1] is assigned to each world ω and denote it by $P(\omega)$
- The belief in, or probability of, a sentence α can then be defined as

$$P(\alpha) = \sum_{\omega:\omega\models\alpha} P(\omega)$$

Degrees of belief or Probability (cont.)

Table 1: A state of belief

world/model	**Earthquake**	**Burglary**	**Alarm**	**P**
ω_1	true	true	true	.0190
ω_2	true	true	false	.0010
ω_3	true	false	true	.0560
ω_4	true	false	false	.0240
ω_5	false	true	true	.1620
ω_6	false	true	false	.0180
ω_7	false	false	true	.0072
ω_8	false	false	false	.7128

Properties of beliefs

1. $0 \leq P(\alpha) \leq 1$ for any sentence α
2. $P(\alpha) = 0$ when α is inconsistent
3. $P(\alpha) = 1$ when α is valid
4. $P(\alpha) + P(\neg\alpha) = 1$ for any sentence α

Entropy

Concept 2

Entropy is an uncertainty quantification about a random variable or several random variables X

$$Entropy(X) = -\sum_{x} P(x) \times \log_2 P(x)$$

where $0 \log_2 0 = 0$ by convention

	Earthquake	**Burglary**	**Alarm**
true	.1	.2	.2442
false	.9	.8	.7558
$Entropy(\cdot)$	.469	.722	.802

Entropy (cont.)

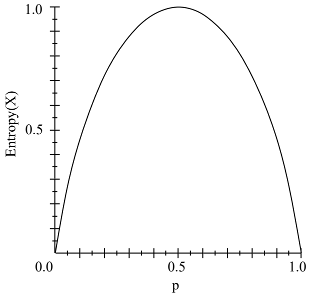


Figure 1: The entropy for a binary variable X with $P(X) = p$

Syntax for probability

- **Probabilities** of propositions

$$P(Cavity = true) = 0.1 \text{ and } P(Weather = sunny) = 0.72$$

- **Probability distribution** gives values for all possible assignments:

$$\mathbf{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$$

- **Joint probability distribution** for a set of r.v.s gives the probability of every atomic event on those r.v.s

$$\mathbf{P}(Weather, Cavity) = \begin{array}{lcccc} Weather = & sunny & rain & cloudy & snow \\ Cavity = true & 0.144 & 0.02 & 0.016 & 0.02 \\ Cavity = false & 0.576 & 0.08 & 0.064 & 0.08 \end{array}$$

Bayes' Rule

Concept 3

- **Prior** or **unconditional probabilities** are corresponded to belief prior to arrival of any (new) evidence
- **Posterior probability** or conditional probability (Bayes' rule)

$$P(a \mid b) = \frac{P(a \wedge b)}{P(b)} \text{ if } P(b) \neq 0$$

- **Product rule** gives an alternative formulation:

$$P(a \wedge b) = P(a \mid b)P(b) = P(b \mid a)P(a)$$

Bayes' Rule (cont.)

- **Chain rule** is derived by successive application of product rule:

$$
\begin{aligned}
\mathbf{P}(X_1, \ldots, X_n) &= \mathbf{P}(X_1, \ldots, X_{n-1})\mathbf{P}(X_n \mid X_1, \ldots, X_{n-1}) \\
&= \mathbf{P}(X_1, \ldots, X_{n-2})\mathbf{P}(X_{n-1} \mid X_1, \ldots, X_{n-2})\mathbf{P}(X_n \mid X_1, \ldots, X_{n-1}) \\
&= \cdots \\
&= \prod_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})
\end{aligned}
$$

Inference by enumeration

Enumeration

For **any proposition** α, sum the **atomic events** where it is true

$$P(\alpha) = \sum_{\omega:\omega \models \alpha} P(\omega)$$

- Start with the sample joint distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

Inference by enumeration (cont.)

	$toothache$		$\neg\ toothache$	
	$catch$	$\neg\ catch$	$catch$	$\neg\ catch$
$cavity$	**.108**	**.012**	**.072**	**.008**
$\neg\ cavity$	**.016**	**.064**	**.144**	**.576**

$$\begin{aligned} P(toothache) &= 0.108 + 0.012 + 0.016 + 0.064 \\ &= 0.2 \end{aligned}$$

Inference by enumeration (cont.)

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

$$\begin{aligned} P(\mathit{cavity} \vee \mathit{toothache}) &= 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 \\ &= 0.28 \end{aligned}$$

Inference by enumeration (cont.)

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

$$
\begin{aligned}
P(\neg cavity \mid toothache) &= \frac{P(\neg cavity \land toothache)}{P(toothache)} \\
&= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} \\
&= 0.4
\end{aligned}
$$

Inference by enumeration (cont.)

Problem

Let $\boldsymbol{X}$ be all the variables. We want the posterior joint distribution of the **query variables** $\boldsymbol{Y}$ given specific values $\boldsymbol{e}$ for the **evidence variables** $\boldsymbol{E}$

Solution

- General idea: compute distribution on query variable by fixing **evidence variables** and summing over **hidden variables**
- Let the **hidden variables** be $\boldsymbol{H} = \boldsymbol{X} - \boldsymbol{Y} - \boldsymbol{E}$ and denominator can be viewed as a **normalization constant** α

$$\boldsymbol{P}(\boldsymbol{Y} \mid \boldsymbol{E} = \boldsymbol{e}) = \alpha\, \boldsymbol{P}(\boldsymbol{Y}, \boldsymbol{E} = \boldsymbol{e}) = \alpha \sum_{\boldsymbol{h}} \boldsymbol{P}(\boldsymbol{Y}, \boldsymbol{E} = \boldsymbol{e}, \boldsymbol{H} = \boldsymbol{h})$$

■

Inference by enumeration (cont.)

- Worst-case time complexity $O(d^n)$ where d is the largest arity
- Space complexity $O(d^n)$ to store the joint distribution

Inference by enumeration (cont.)

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

$$\begin{aligned}
& \mathbf{P}(Cavity|toothache) \\
= \ & \alpha\, \mathbf{P}(Cavity, toothache) \\
= \ & \alpha\, [\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)] \\
= \ & \alpha\, [\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] \\
= \ & \alpha\, \langle 0.12, 0.08\rangle \\
= \ & \langle 0.6, 0.4\rangle
\end{aligned}$$

Independence

Concept 4

A and B are **independent** $(A \perp B)$ iff

$$\boldsymbol{P}(A \mid B) = \boldsymbol{P}(A)$$

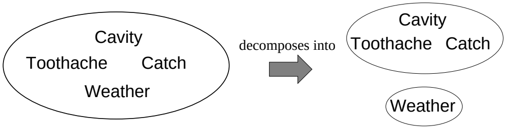
or

$$\boldsymbol{P}(B \mid A) = \boldsymbol{P}(B)$$

or

$$\boldsymbol{P}(A, B) = \boldsymbol{P}(A) P(B)$$

Independence (cont.)



$$\mathbf{P}(Toothache, Catch, Cavity, Weather) = \mathbf{P}(Toothache, Catch, Cavity)\mathbf{P}(Weather)$$

- 32 entries reduced to 12; for n independent biased coins, $2^n \to n$
- Absolute independence powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional independence

Concept 5

A and B are **independent** given C ($A \perp B \mid C$) iff

$$\boldsymbol{P}(A \mid B, C) = \boldsymbol{P}(A \mid C)$$

or

$$\boldsymbol{P}(B \mid A, C) = \boldsymbol{P}(B \mid C)$$

or

$$\boldsymbol{P}(A, B \mid C) = \boldsymbol{P}(A \mid C)\boldsymbol{P}(B \mid C)$$

Conditional independence (cont.)

- $\mathbf{P}(\mathit{Toothache}, \mathit{Cavity}, \mathit{Catch})$ 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:
$$P(\mathit{catch} \mid \mathit{toothache}, \mathit{cavity}) = P(\mathit{catch} \mid \mathit{cavity})$$
- The same independence holds if I haven't got a cavity:
$$P(\mathit{catch} \mid \mathit{toothache}, \neg \mathit{cavity}) = P(\mathit{catch} \mid \neg \mathit{cavity})$$
- *Catch* is **conditionally independent** of *Toothache* given *Cavity*:
$$\mathbf{P}(\mathit{Catch} \mid \mathit{Toothache}, \mathit{Cavity}) = \mathbf{P}(\mathit{Catch} \mid \mathit{Cavity})$$

Conditional independence (cont.)

- Equivalent statements:

$$\mathbf{P}(Toothache \mid Catch, Cavity) = \mathbf{P}(Toothache \mid Cavity)$$

$$\mathbf{P}(Toothache, Catch \mid Cavity) = \mathbf{P}(Toothache \mid Cavity)\mathbf{P}(Catch \mid Cavity)$$

- Write out full joint distribution using chain rule:

$$\begin{aligned}
& \mathbf{P}(Toothache, Catch, Cavity) \\
= \; & \mathbf{P}(Toothache \mid Catch, Cavity)\mathbf{P}(Catch, Cavity) \\
= \; & \mathbf{P}(Toothache \mid Catch, Cavity)\mathbf{P}(Catch \mid Cavity)\mathbf{P}(Cavity) \\
= \; & \mathbf{P}(Toothache \mid Cavity)\mathbf{P}(Catch \mid Cavity)\mathbf{P}(Cavity)
\end{aligned}$$

I.e., $2 + 2 + 1 = 5$ independent numbers

Conditional independence (cont.)

- **In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n.**
- *Conditional independence is our most basic and robust* form of knowledge about uncertain environments.

Bayes' rule and conditional independence

- Bayes' rule

$$\mathbf{P}(Y \mid X) = \frac{\mathbf{P}(X \mid Y)\mathbf{P}(Y)}{\mathbf{P}(X)} = \alpha\, \mathbf{P}(X \mid Y)\mathbf{P}(Y)$$

Useful for assessing **diagnostic** probability from **causal** probability

$$\mathbf{P}(Cause \mid Effect) = \frac{\mathbf{P}(Effect \mid Cause)\mathbf{P}(Cause)}{\mathbf{P}(Effect)}$$

$$\begin{aligned} & \mathbf{P}(Cavity \mid toothache \wedge catch) \\ = \quad & \alpha\, \mathbf{P}(toothache \wedge catch \mid Cavity)\mathbf{P}(Cavity) \\ = \quad & \alpha\, \mathbf{P}(toothache \mid Cavity)\mathbf{P}(catch \mid Cavity)\mathbf{P}(Cavity) \end{aligned}$$

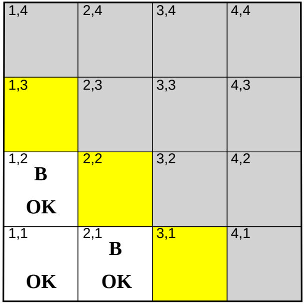
Bayes' rule and conditional independence (cont.)

- This is an example of a **naive Bayes** model

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i \mid Cause)$$

- Total number of parameters is *linear* in n

The Wumpus World Revisitted



- $P_{ij} = \mathit{true}$ iff $[i,j]$ contains a pit
- $B_{ij} = \mathit{true}$ iff $[i,j]$ is breezy
- Include only $B_{1,1}, B_{1,2}, B_{2,1}$ in the probability model

Specifying the probability model

- The full joint distribution is $\boldsymbol{P}(P_{1,1}, \ldots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1})$
- Apply product rule: $\boldsymbol{P}(B_{1,1}, B_{1,2}, B_{2,1} \mid P_{1,1}, \ldots, P_{4,4})\boldsymbol{P}(P_{1,1}, \ldots, P_{4,4})$ (Do it this way to get $P(Effect \mid Cause)$.)
- First term: 1 if pits are adjacent to breezes, 0 otherwise
- Second term: pits are placed randomly, probability 0.2 per square:

$$\boldsymbol{P}(P_{1,1}, \ldots, P_{4,4}) = \prod_{i,j=1,1}^{4,4} \boldsymbol{P}(P_{i,j}) = 0.2^n \times 0.8^{16-n}$$

for n pits.

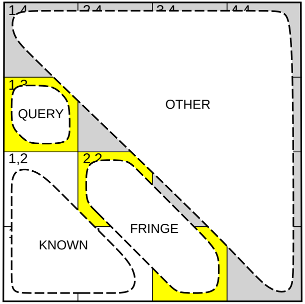
Observations and query

- We know the following facts:
  $b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$
  $known = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$
  Define $Unknown = P_{ij}$s other than $P_{1,3}$ and $Known$
- Query is $\mathbf{P}(P_{1,3} \mid known, b)$
- For inference by enumeration, we have

$$\mathbf{P}(P_{1,3} \mid known, b) = \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b)$$

Grows exponentially with number of squares!

Using conditional independence



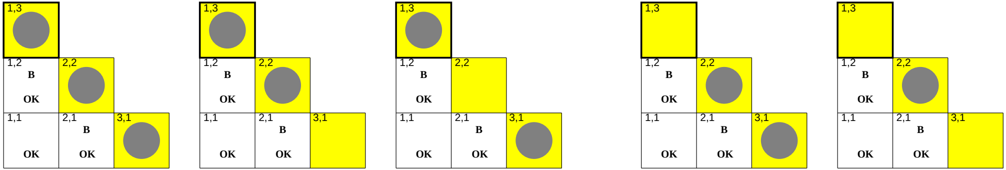
- Basic insight: observations are conditionally independent of other hidden squares given neighbouring hidden squares
- Define $Unknown = Fringe(\text{or } Frontier) \cup Other$
  $\boldsymbol{P}(b \mid P_{1,3}, Known, Unknown) = \boldsymbol{P}(b \mid P_{1,3}, Known, Fringe)$

Using conditional independence (cont.)

- Manipulate query into a form where we can use this!

$$
\begin{aligned}
& \mathbf{P}(P_{1,3} \mid known, b) \\
= \quad & \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b) \\
= \quad & \alpha \sum_{unknown} \mathbf{P}(b \mid P_{1,3}, known, unknown) \mathbf{P}(P_{1,3}, known, unknown) \\
= \quad & \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b \mid known, P_{1,3}, fringe, other) \mathbf{P}(P_{1,3}, known, fringe, other) \\
= \quad & \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b \mid known, P_{1,3}, fringe) \mathbf{P}(P_{1,3}, known, fringe, other) \\
= \quad & \alpha \sum_{fringe} \mathbf{P}(b \mid known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3}, known, fringe, other) \\
= \quad & \alpha \sum_{fringe} \mathbf{P}(b \mid known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3}) \mathbf{P}(known) \mathbf{P}(fringe) \mathbf{P}(other) \\
= \quad & \alpha \, \mathbf{P}(known) P(P_{1,3}) \sum_{fringe} \mathbf{P}(b \mid known, P_{1,3}, fringe) \mathbf{P}(fringe) \sum_{other} \mathbf{P}(other) \\
= \quad & \alpha' \, \mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b \mid known, P_{1,3}, fringe) \mathbf{P}(fringe)
\end{aligned}
$$

Using conditional independence (cont.)



0.2 x 0.2 = 0.04 0.2 x 0.8 = 0.16 0.8 x 0.2 = 0.16 0.2 x 0.2 = 0.04 0.2 x 0.8 = 0.16

$$
\begin{aligned}
P(P_{1,3} \mid known, b) &= \alpha' \langle 0.2(0.04 + 0.16 + 0.16), 0.8(0.04 + 0.16) \rangle \\
&\approx \langle 0.31, 0.69 \rangle \\
P(P_{2,2} \mid known, b) &\approx \langle 0.86, 0.14 \rangle
\end{aligned}
$$

Bayesian Networks

Bayesian networks

Concept 6

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

Syntax

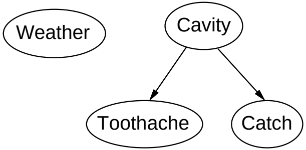
- A set of nodes, one per variable
- A directed, acyclic graph (link $\approx$ "directly influences")
- A conditional distribution for each node given its parents:

$$\boldsymbol{P}(X_i \mid parents(X_i))$$

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

Example (cont.)

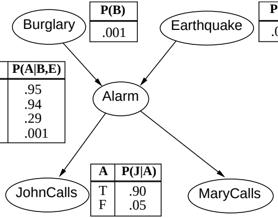
Problem

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



P(B)
.001

P(E)
.002

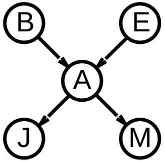
B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

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

A	P(M|A)
T	.70
F	.01

Compactness

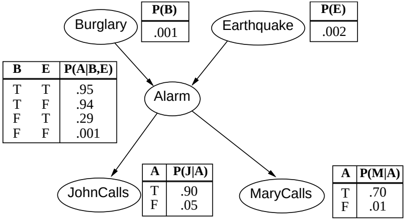
- A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values
- Each row requires one number p for $X_i = true$ (the number for $X_i = false$ is just $1 - p$)
- If each variable has no more than k parents, the complete network requires $O(n \times 2^k)$ numbers
  - 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. $2^5 - 1 = 31$)



Global semantics

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

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



- For example,

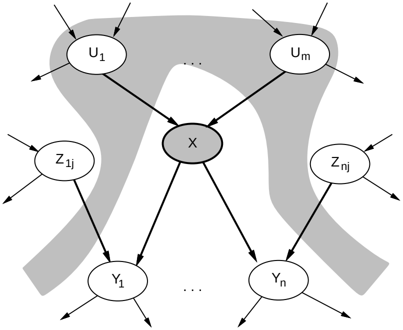
$$\begin{aligned} P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) &= P(j \mid a)P(m \mid a)P(a \mid \neg b, \neg e)P(\neg b)P(\neg e) \\ &= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ &\approx 0.00063 \end{aligned}$$

Local semantics

Local semantics: each node is conditionally independent of its nondescendants given its parents

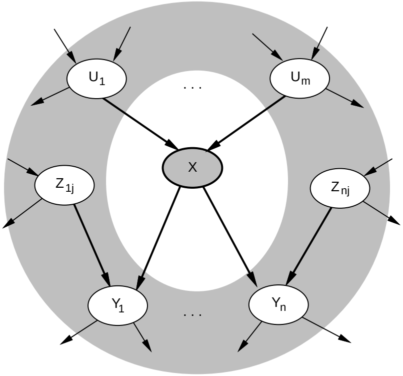
Theorem 1

Local semantics ≡ global semantics



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

$$\mathbf{P}(X_i \mid parents(X_i)) = \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})$$

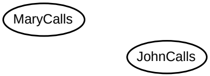
This choice of parents guarantees the global semantics:

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

Example: Burglary alarm

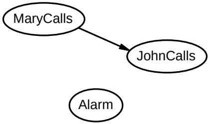
- Suppose we choose the ordering M, J, A, B, E

$$P(J \mid M) = P(J)?$$



Example: Burglary alarm (cont.)

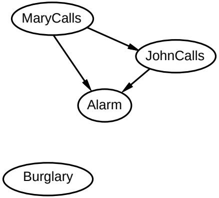
- Suppose we choose the ordering M, J, A, B, E



$P(J \mid M) = P(J)$? No
$P(A \mid J, M) = P(A \mid J)$? $P(A \mid J, M) = P(A)$?

Example: Burglary alarm (cont.)

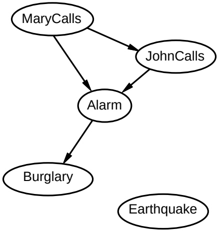
- Suppose we choose the ordering M, J, A, B, E



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

Example: Burglary alarm (cont.)

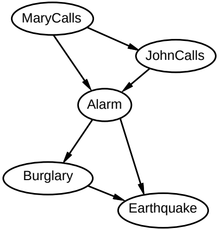
- Suppose we choose the ordering M, J, A, B, E



$P(J \mid M) = P(J)$? No
$P(A \mid J, M) = P(A \mid J)$? $P(A \mid J, M) = P(A)$? No
$P(B \mid A, J, M) = P(B \mid A)$? Yes
$P(B \mid A, J, M) = P(B)$? No
$P(E \mid B, A, J, M) = P(E \mid A)$?
$P(E \mid B, A, J, M) = P(E \mid A, B)$?

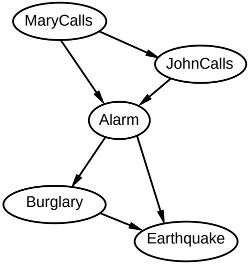
Example: Burglary alarm (cont.)

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



$P(J \mid M) = P(J)$? No
$P(A \mid J, M) = P(A \mid J)$? $P(A \mid J, M) = P(A)$? No
$P(B \mid A, J, M) = P(B \mid A)$? Yes
$P(B \mid A, J, M) = P(B)$? No
$P(E \mid B, A, J, M) = P(E \mid A)$? No
$P(E \mid B, A, J, M) = P(E \mid A, B)$? Yes

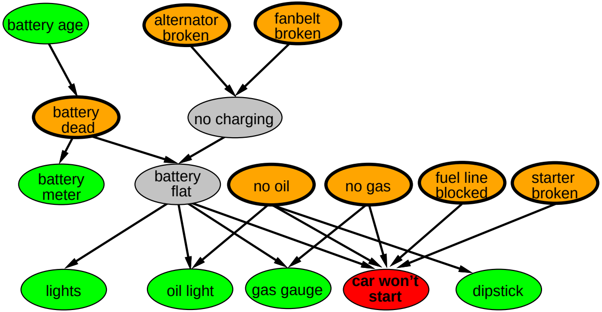
Example: Burglary alarm (cont.)



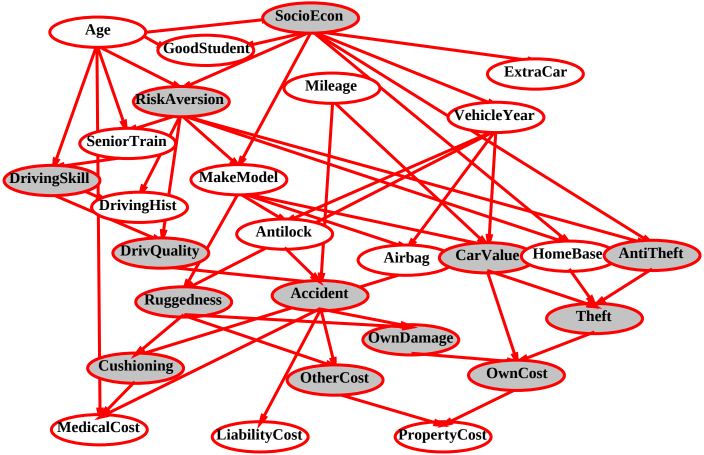
- Deciding conditional independence is hard in noncausal directions (Causal models and conditional independence seem hardwired for humans!)
- Assessing conditional probabilities is hard in noncausal directions
- Network is less compact: $1 + 2 + 4 + 2 + 4 = 13$ numbers needed

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

Problem

- 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:

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

  - Boolean functions

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

  - Numerical relationships among continuous variables

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

Compact conditional distributions (cont.)

- **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

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

Number of parameters *linear* in number of parents

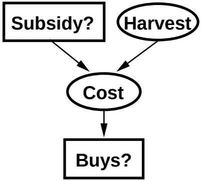
Compact conditional distributions (cont.)

$$
\begin{aligned}
q_{cold} &= P(\neg fever \mid cold, \neg flu, \neg malaria) = 0.6 \\
q_{flu} &= P(\neg fever \mid \neg cold, flu, \neg malaria) = 0.2 \\
q_{malaria} &= P(\neg fever \mid \neg cold, \neg flu, malaria) = 0.1
\end{aligned}
$$

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	$0.012 = 0.6 \times 0.2 \times 0.1$

Bayesian nets with continuous variables

Discrete (*Subsidy*? and *Buys*?); continuous (*Harvest* and *Cost*)



- 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** (LG) model, e.g.,:

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

- 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 (cont.)

- All-continuous network with LG distributions $\Longrightarrow$ full joint distribution is a multivariate Gaussian
- Discrete+continuous LG network is a **conditional Gaussian** network i.e., a multivariate Gaussian over all continuous variables for each combination of discrete variable values

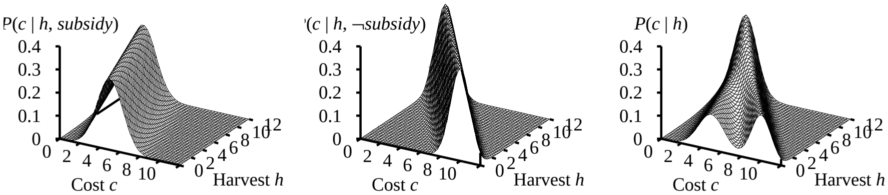
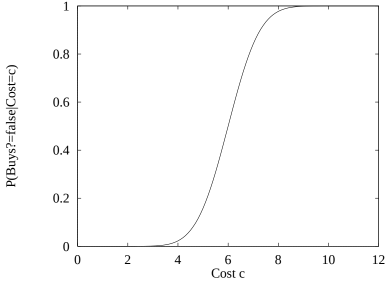


Figure 2: The graphs in (1) and (2) show the probability distribution over *Cost* as a function of *Harvest* size, with *Subsidy* true and false, respectively. Graph (3) shows the distribution $P(Cost \mid Harvest)$, obtained by summing over the two subsidy cases.

Discrete variable given continuous parents

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



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

$$\Phi(x) = \int_{-\infty}^{x} N(0,1)(x)dx$$

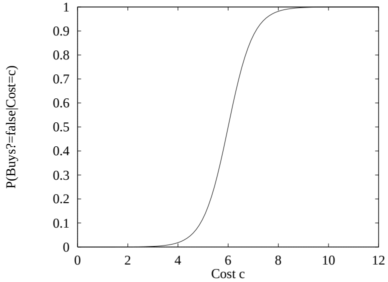
$$P(Buys? = true \mid Cost = c) = \Phi((-c+\mu)/\sigma)$$

Discrete variable given continuous parents

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

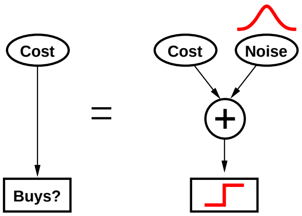
$$P(Buys? = true \mid Cost = c) = \frac{1}{1 + exp(-2\frac{-c+\mu}{\sigma})}$$

- Sigmoid has similar shape to probit but much longer tails:



Why the probit?

1. It's sort of the right shape
2. Can view as hard threshold whose location is subject to noise



References

Goodfellow, I., Bengio, Y., and Courville, A. (2016).
Deep learning.
MIT press.

Lê, B. and Tô, V. (2014).
Cở sở trí tuệ nhân tạo.
Nhà xuất bản Khoa học và Kỹ thuật.

Nguyen, T. (2018).
Artificial intelligence slides.
Technical report, HCMC University of Sciences.

Russell, S. and Norvig, P. (2016).
Artificial intelligence: a modern approach.
Pearson Education Limited.