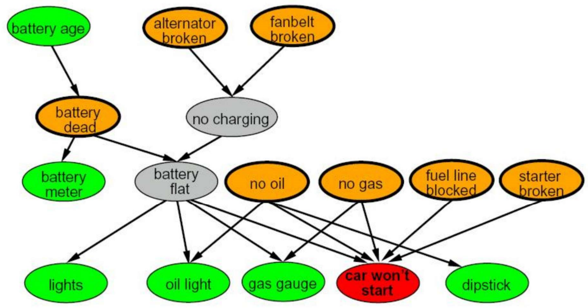
KI2 - 4

Bayesian Belief Networks

AIMA, Chapter 14

Kunstmatige Intelligentie / RuG

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

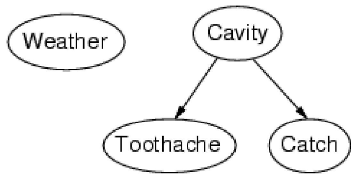


Bayesian Networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions
- The constituents of a Bayesian network:
  - a set of nodes, one per variable
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents
    P (X_i | 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:

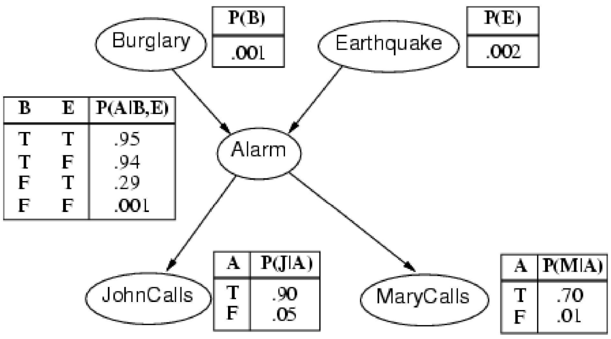


- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity (e.g.: p(T|Cav) = p(T|Cav, Catch)*
  =p(T|Cav, not Catch)

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: *Burglary*, *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.



P(B)
.001

P(E)
.002

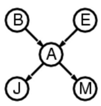
B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

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

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

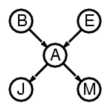
Compactness

- 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 $O(n \cdot 2^k)$ 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)

Computing the Full Joint Distribution



- The full joint distribution is equal to the product of the local conditional distributions:

$$\mathbf{P}(X_1, \ldots, X_n) = \pi_i\ \mathbf{P}(X_i \mid Parents(X_i))$$

e.g. $\mathbf{P}(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$
$= \mathbf{P}(j \mid a)\ \mathbf{P}(m \mid a)\ \mathbf{P}(a \mid \neg b, \neg e)\ \mathbf{P}(\neg b)\ \mathbf{P}(\neg e)$

Constructing Bayesian Networks

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

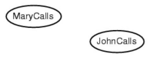
- This choice of parents guarantees:

$\mathbf{P}(X_1, \ldots, X_n) = \pi_i\ \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})$ (chain rule)

$= \pi_i\, \mathbf{P}(X_i \mid Parents(X_i))$ (by construction)

Example

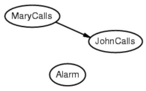
- Suppose we choose the ordering *M, J, A, B, E*



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

Example

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

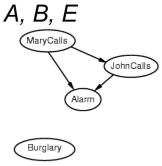


P(J | M) = P(J)? No, John can decide for himself

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

Example

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



P(J | M) = P(J)? No

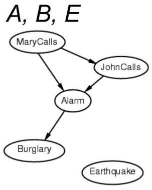
P(A | J, M) = P(A | J)? P(A | J, M) = P(A)? No

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)? No

P(A | J, M) = P(A | J)? P(A | J, M) = P(A)? No

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

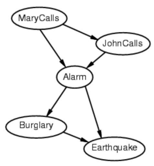
P(B | A, J, M) = P(B)? No, the burglar decides himself

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

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

Example

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



P(J | M) = P(J)? No

P(A | J, M) = P(A | J)? P(A | J, M) = P(A)? No

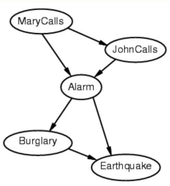
P(B | A, J, M) = P(B | A)? Yes

P(B | A, J, M) = P(B)? No, the burglar decides himself

P(E | B, A ,J, M) = P(E | A)? No, earthquakes don't wait for an alarm to go off

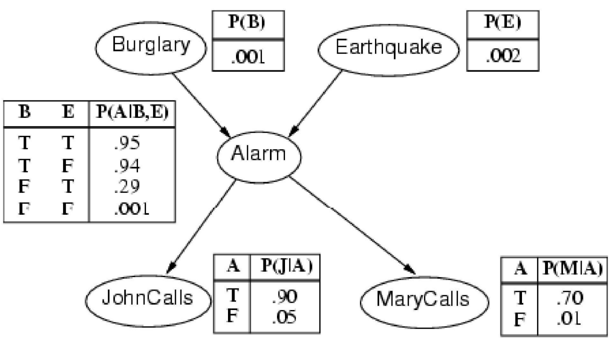
P(E | B, A, J, M) = P(E | A, B)? Yes ➔ *unclear picture*

Example contd.



- Deciding conditional independence is difficult in **noncausal** directions.
- Causal models and conditional independence seem hardwired in human reasoning!
- Also: the bad network was less compact: 1 + 2 + 4 + 2 + 4 = 13 numbers needed.

Causal ordering is natural, easy



P(B)
.001

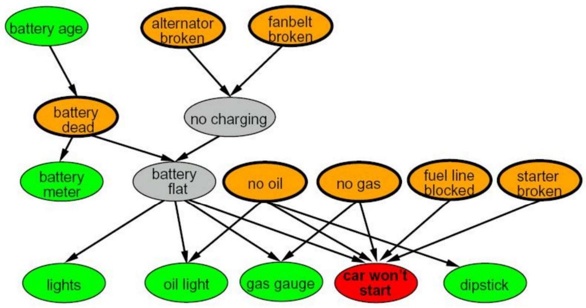
P(E)
.002

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

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



Summary

- Bayesian networks provide a natural representation for (causally induced) conditional independence.
- Topology + CPTs = compact representation of joint distribution.
- Generally easy for domain experts to construct.
- Belief networks have found increasing use in complex diagnosis problems (medical, cars, PC operating systems).

All is well in Belief Network Land?

- Problems:
  - Network construction: often gofai human construction labor (knowledge based, 'Cyc' etc.)
  - Estimation of probabilities
  - Product of probabilities:
    - Not p but p' = p + ε therefore

$\prod_i (p_i + \varepsilon)$ propagation of errors! especially in a long chain

Summary for Bayesian Methods

- Bayesian methods:
  - Learning = estimation of probability distributions of samples from different classes
  - Classification = use these estimates to determine which class is more likely for a new instance
- Naive Bayes:
  - Assumes that attributes are independent.
- Bayesian Belief Networks:
  - Assumes that subsets of attributes are independent.
- Bayesian methods allow combining prior knowledge about the world with evidence from the data stream.