Bayesian Network

Outline

- Syntax
- Semantics
- Exact inference by enumeration
- Exact inference by variable elimination

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")

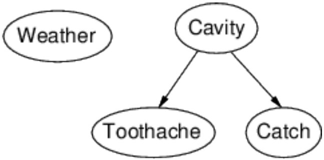
a conditional distribution for each node given its parents:

$\mathbf{P}(X_i|Parents(X_i))$

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*

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:

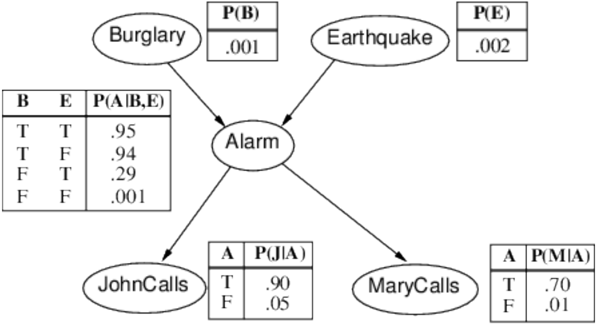
– A burglar can set the alarm off

– An earthquake can set the alarm off

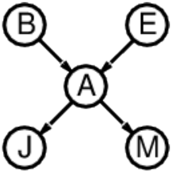
– The alarm can cause Mary to call

– The alarm can cause John to call

Example contd.

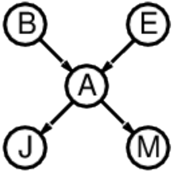


Compactness



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. $2^5 - 1 = 31$)

Global semantics

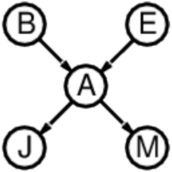


Global semantics defines the full joint distribution as the product of the local conditional distributions:

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

e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

Global semantics



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

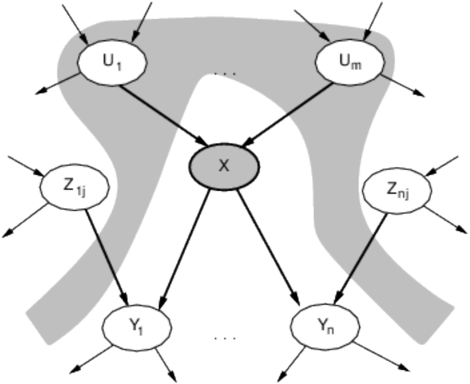
$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

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

Local semantics

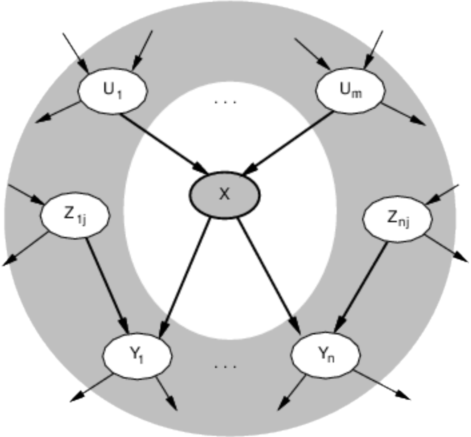
Local semantics: each node is conditionally independent of its nondescendants given its parents



Theorem: Local semantics ⇔ global semantics

Markov blanket

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



Answering queries on conditional dependencies

◊ Q: are X and Y conditionally independent given evidence variables $\{Z\}$?

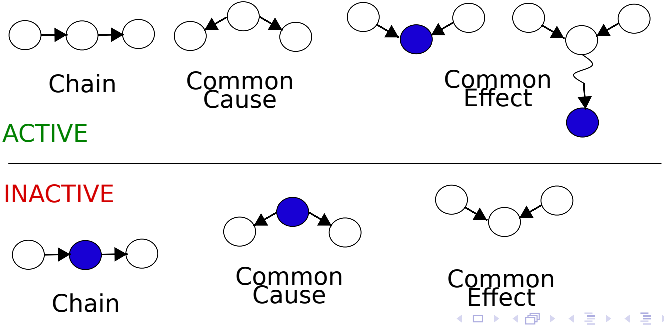
◊ Can write this as: $X \perp\!\!\!\perp Y \mid \{Z\}$

◊ We can analyze the undirected graph defined by the BN:

- $X \perp\!\!\!\perp Y \mid \{Z\}$ is true if X and Y are separated by $\{Z\}$
- consider all (undirected) paths between X and Y
- If no active paths ⇒ independence
- active path if each triple is active
- triplet: specific configurations of three variables.

Active and Inactive triples

- A triple is active if:
  - Causal Chain: $A \rightarrow B \rightarrow C$ where B is not observed (both directions)
  - Common Cause: $A \leftarrow B \rightarrow C$ where B is not observed
  - Common Effect: $A \rightarrow B \leftarrow C$ where B (or one of its descendents) is observed



Reachability (or D-separation)

- Given a query $X \perp\!\!\!\perp Y \mid \{Z\}$
- Highlight all evidence variables ($\{Z\}$)
- For all undirected paths between X and Y
  - if a path is active $\rightarrow X \perp\!\!\!\perp Y \mid \{Z\}$ is not guaranteed
- If no undirected path is active $\rightarrow X \perp\!\!\!\perp Y \mid \{Z\}$ is guaranteed

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

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

This choice of parents guarantees the global semantics:

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

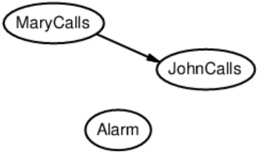
Example

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

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

Example

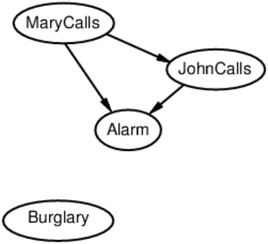
Suppose we choose the ordering M, J, A, B, E



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

Example

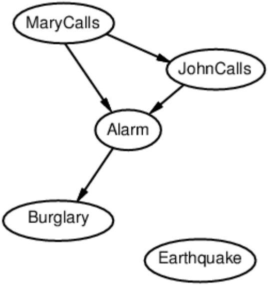
Suppose we choose the ordering M, J, A, B, E



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

Example

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



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

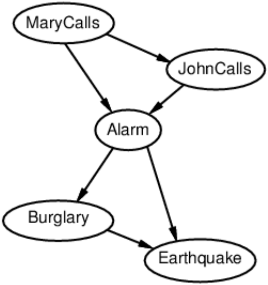
$P(A|J,M) = P(A|J)$? $P(A|J,M) = P(A)$? No

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

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

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

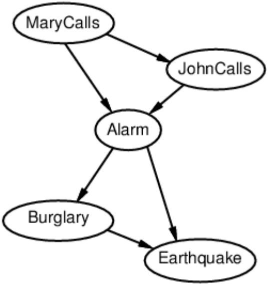
$P(B|A,J,M) = P(B|A)$? Yes

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

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

Example

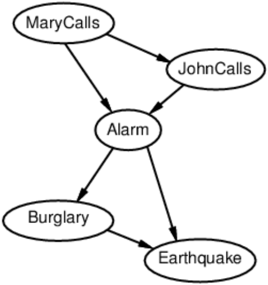
Suppose we choose the ordering M, J, A, B, E



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

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

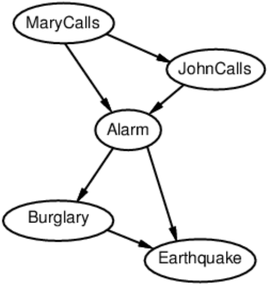
$P(B|A,J,M) = P(B|A)$? Yes

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

$P(E|B,A,J,M) = P(E|A)$? No

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

Example contd.



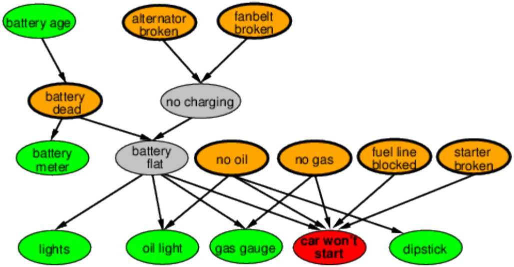
◊ Deciding conditional independence is hard in noncausal directions (Causal models and conditional independence seem hardwired for humans!)

◊ Assessing conditional probabilities is hard in noncausal directions

◊ Network is less compact: $1 + 2 + 4 + 2 + 4 = 13$ numbers needed

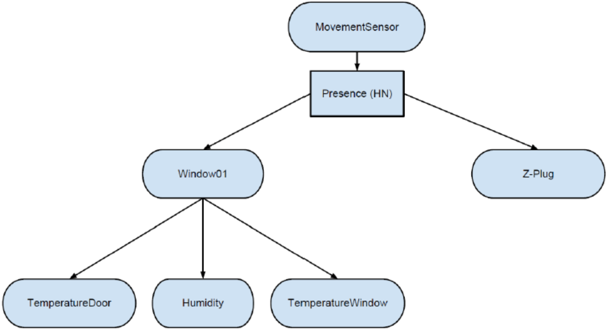
Example: Car diagnosis

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



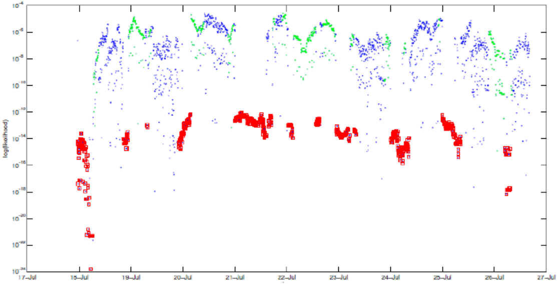
Example: energy usage

Course Project by Ambrosini and Scapin Conditional dependence for sensors in a facility room (coffee room)



Example: energy usage

Model learning using BNT
Joint Distribution of data based on sensor readings (log-likelihood)
Red Squares = fake readings artificially inserted



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:
$X = f(Parents(X))$ for some function f
E.g., Boolean functions
$NorthAmerican \Leftrightarrow Canadian \vee US \vee Mexican$
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	$0.012 = 0.6 \times 0.2 \times 0.1$

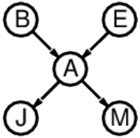
Number of parameters **linear** in number of parents

Inference tasks

Simple queries: compute posterior marginal $\mathbf{P}(X_i|\mathbf{E}=\mathbf{e})$
e.g., $P(NoGas|Gauge=empty, Lights=on, Starts=false)$
Conjunctive queries:
$\mathbf{P}(X_i, X_j|\mathbf{E}=\mathbf{e}) = \mathbf{P}(X_i|\mathbf{E}=\mathbf{e})\mathbf{P}(X_j|X_i, \mathbf{E}=\mathbf{e})$
Optimal decisions: decision networks include utility information; probabilistic inference required for
$P(outcome|action, evidence)$
Value of information: which evidence to seek next?
Sensitivity analysis: which probability values are most critical?
Explanation: why do I need a new starter motor?

Inference by enumeration

Slightly intelligent way to sum out variables from the joint without actually constructing its explicit representation



Simple query on the burglary network:

$\mathbf{P}(B|j,m)$
$= \mathbf{P}(B,j,m)/P(j,m)$
$= \alpha\mathbf{P}(B,j,m)$
$= \alpha \sum_e \sum_a \mathbf{P}(B,e,a,j,m)$

Rewrite full joint entries using product of CPT entries:

$\mathbf{P}(B|j,m)$
$= \alpha \sum_e \sum_a \mathbf{P}(B)P(e)\mathbf{P}(a|B,e)P(j|a)P(m|a)$
$= \alpha\mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e)P(j|a)P(m|a)$

Recursive depth-first enumeration: $O(n)$ space, $O(d^n)$ time

Enumeration algorithm

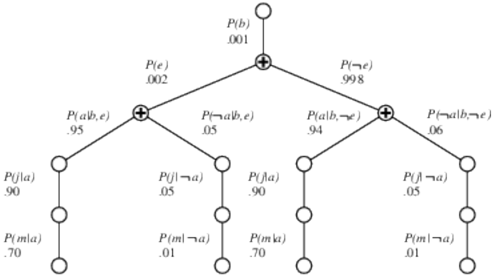
function ENUMERATION-ASK(X, **e**, bn) **returns** a distribution over X
  inputs: X, the query variable
    e, observed values for variables **E**
    bn, a Bayesian network with variables $\{X\} \cup \mathbf{E} \cup \mathbf{Y}$

  $\mathbf{Q}(X) \leftarrow$ a distribution over X, initially empty
  for each value x_i of X do
    extend **e** with value x_i for X
    $\mathbf{Q}(x_i) \leftarrow$ ENUMERATE-ALL(VARS[bn], **e**)
  return NORMALIZE($\mathbf{Q}(X)$)

---

function ENUMERATE-ALL($vars$, **e**) **returns** a real number
  if EMPTY?($vars$) then return 1.0
  $Y \leftarrow$ FIRST($vars$)
  if Y has value y in **e**
    then return $P(y \mid Pa(Y)) \times$ ENUMERATE-ALL(REST($vars$), **e**)
    else return $\sum_y P(y \mid Pa(Y)) \times$ ENUMERATE-ALL(REST($vars$), $\mathbf{e}_y$)
      where $\mathbf{e}_y$ is **e** extended with $Y = y$

Evaluation tree



Enumeration is inefficient: repeated computation

e.g., computes $P(j|a)P(m|a)$ for each value of e

Inference by variable elimination

Variable elimination: carry out summations right-to-left, storing intermediate results (factors) to avoid recomputation

$$\begin{aligned}
&\mathbf{P}(B|j,m) \\
&= \alpha \underbrace{\mathbf{P}(B)}_{B} \sum_e \underbrace{P(e)}_{E} \sum_a \underbrace{\mathbf{P}(a|B,e)}_{A} \underbrace{P(j|a)}_{J} \underbrace{P(m|a)}_{M} \\
&= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e) P(j|a) f_M(a) \\
&= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e) f_J(a) f_M(a) \\
&= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a f_A(a,b,e) f_J(a) f_M(a) \\
&= \alpha \mathbf{P}(B) \sum_e P(e) f_{\bar{A}JM}(b,e) \text{ (sum out } A\text{)} \\
&= \alpha \mathbf{P}(B) f_{\bar{E}\bar{A}JM}(b) \text{ (sum out } E\text{)} \\
&= \alpha f_B(b) \times f_{\bar{E}\bar{A}JM}(b)
\end{aligned}$$

Variable elimination: Basic operations

Summing out a variable from a product of factors:

move any constant factors outside the summation

add up submatrices in pointwise product of remaining factors

$\sum_x f_1 \times \cdots \times f_k = f_1 \times \cdots \times f_i \; \sum_x \; f_{i+1} \times \cdots \times f_k =$

$f_1 \times \cdots \times f_i \times f_{\bar{X}}$

assuming $f_1, \ldots, f_i$ do not depend on X

Pointwise product of factors f_1 and f_2:

$f_1(x_1, \ldots, x_j, y_1, \ldots, y_k) \times f_2(y_1, \ldots, y_k, z_1, \ldots, z_l)$

$= f(x_1, \ldots, x_j, y_1, \ldots, y_k, z_1, \ldots, z_l)$

E.g., $f_1(a, b) \times f_2(b, c) = f(a, b, c)$

Complexity of exact inference

Singly connected networks (or polytrees):

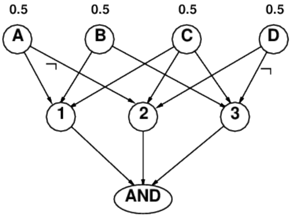
– any two nodes are connected by at most one (undirected) path

– time and space cost of variable elimination are $O(d^k n)$

Multiply connected networks:

– can reduce 3SAT to exact inference $\Longrightarrow$ NP-hard

1. A ∨ B ∨ C
2. C ∨ D ∨ ¬A
3. B ∨ C ∨ ¬D



Approximate Inference by stochastic simulation

Basic idea:

1) Draw N samples from a sampling distribution
2) Compute an approximate posterior probability $\hat{P}$
3) Show this converges to the true probability P

Outline:

– Sampling from an empty network
– Rejection sampling: reject samples disagreeing with evidence
– Likelihood weighting: use evidence to weight samples
– Markov chain Monte Carlo (MCMC): sample from a stochastic process whose stationary distribution is the true posterior

Summary

◊ Bayes nets provide a natural representation for (causally induced)
conditional independence
◊ Topology + CPTs = compact representation of joint distribution
◊ Exact Inference can exploit this compact representation
◊ In general exact inference is hard