Bayesian networks

AIMA2e Chapter 14
Outline

♦ Syntax
♦ Semantics
♦ Constructing Bayesian Networks
♦ Exact inference by enumeration
♦ Exact inference by variable elimination (SKIP)
♦ Approximate inference by stochastic simulation (SKIP)
♦ Approximate inference by Markov chain Monte Carlo (SKIP)
Bayesian networks

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

A Bayesian Belief Network is a directed acyclic graph where:
- a set of nodes, one per random variable
- a directed, acyclic graph (link \( \approx \) “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 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.

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

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

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

\[
P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}) \quad \text{(chain rule)}
\]

\[
= \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i)) \quad \text{(by construction)}
\]
Example

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

\[ P(J|M) = P(J)? \]
Example

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

\[ P(J|M) = P(J) \quad \text{No} \]
\[ P(A|J, M) = P(A|J) \quad P(A|J, M) = P(A) ? \]

MaryCalls

\[ \rightarrow \]

JohnCalls

Alarm
Example

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

\[ P(J|M) = P(J)? \quad \text{No} \]
\[ P(A|J, M) = P(A|J)? \quad P(A|J, M) = P(A)? \quad \text{No} \]
\[ P(B|A, J, M) = P(B|A)? \]
\[ P(B|A, J, M) = P(B)? \]
Example

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

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

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

\[
\begin{align*}
P(J|M) &= P(J) \quad \text{No} \\
P(A|J,M) &= P(A|J) \quad P(A|J,M) = P(A) \quad \text{No} \\
P(B|A,J,M) &= P(B|A) \quad \text{Yes} \\
P(B|A,J,M) &= P(B) \quad \text{No} \\
P(E|B,A,J,M) &= P(E|A) \quad \text{No} \\
P(E|B,A,J,M) &= P(E|A,B) \quad \text{Yes}
\end{align*}
\]
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
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 the small burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$).

For the large burglary net, the network is less compact: $1 + 2 + 4 + 2 + 4 = 13$ numbers needed.
Global semantics of Bayesian Networks

A Bayesian network 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., \( P(j \land m \land a \land \neg b \land \neg e) \)

= 
Global semantics of Bayesian Networks

A Bayesian network 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., \( P(j \land m \land a \land \neg b \land \neg e) \)

\[ = P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \]
Local semantics: each node is conditionally independent of its nondescendants given its parents

Theorem: \textbf{Local semantics} $\iff$ \textbf{global semantics}
Markov blanket

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

SOME OF THE TOPIC (more examples and how belief update is done) IS COVERED IN Bayesian Belief Net.doc!
Outline for Chapter 15

♦ Exact inference by enumeration
♦ Exact inference by variable elimination (SKIP)
♦ Approximate inference by stochastic simulation (SKIP)
♦ Approximate inference by Markov chain Monte Carlo (SKIP)
Inference tasks

Simple queries: compute posterior marginal $P( X_i | E = e )$
  e.g., $P( \text{NoGas} | \text{Gauge} = \text{empty}, \text{Lights} = \text{on}, \text{Starts} = \text{false} )$

Conjunctive queries: $P( X_i, X_j | E = e ) = P( X_i | E = e ) P( X_j | X_i, E = e )$

Optimal decisions: decision networks include utility information;
  probabilistic inference required for $P( \text{outcome} | \text{action}, \text{evidence} )$

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:
\[ P(B|j, m) = \frac{P(B, j, m)}{P(j, m)} = \frac{\alpha P(B, j, m)}{\alpha \sum_e \sum_a P(B, e, a, j, m)} \]

Rewrite full joint entries using product of CPT entries:
\[ P(B|j, m) = \frac{\alpha \sum_e \sum_a P(B) P(e) P(a|B, e) P(j|a) P(m|a)}{\alpha P(B) \sum_e P(e) \sum_a P(a|B, e) P(j|a) P(m|a)} \]

Recursive depth-first enumeration: \( O(n) \) space, \( O(d^m) \) 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 $\{X\} \cup E \cup Y$

$Q(X) \leftarrow$ a distribution over $X$, initially empty

for each value $x_i$ of $X$ do

- extend $e$ with value $x_i$ for $X$
- $Q(x_i) \leftarrow$ Enumerate-All($\text{VARS}[bn], e$)

**return** Normalize($Q(X)$)

**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($\text{REST}(vars), e$)

else return $\sum_y P(y \mid Pa(Y)) \times$ Enumerate-All($\text{REST}(vars), e_y$)

where $e_y$ 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

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

$$= \alpha P(B) \sum_{e} P(e) \sum_{a} P(a|B, e) P(j|a) f_M(a)$$

$$= \alpha P(B) \sum_{e} P(e) \sum_{a} f_A(a, b, e) f_J(a) f_M(a)$$

$$= \alpha P(B) \sum_{e} P(e) f_{\tilde{A}JM}(b, e) \text{(sum out } A)$$

$$= \alpha P(B) f_{\tilde{E}\tilde{A}JM}(b) \text{(sum out } E)$$

$$= \alpha f_B(b) \times f_{\tilde{E}\tilde{A}JM}(b)$$
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

\[
\sum_x f_1 \times \cdots \times f_k = f_1 \times \cdots \times f_i \sum_x f_{i+1} \times \cdots \times f_k = f_1 \times \cdots \times f_i \times f_X
\]

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., \(f_1(a, b) \times f_2(b, c) = f(a, b, c)\)
Variable elimination algorithm

function Elimination-Ask($X, e, bn$) returns a distribution over $X$

inputs: $X$, the query variable
         $e$, evidence specified as an event
         $bn$, a belief network specifying joint distribution $P(X_1, \ldots, X_n)$

$factors \leftarrow []; vars \leftarrow \text{Reverse(VARS[bn])}$

for each $var$ in $vars$ do
    $factors \leftarrow [\text{Make-Factor}(var, e)|factors]$
    if $var$ is a hidden variable then $factors \leftarrow \text{Sum-Out}(var, factors)$

return Normalize(Pointwise-Product($factors$))
Irrelevant variables

Consider the query $P(\text{JohnCalls}|\text{Burglary} = \text{true})$

$$P(J|b) = \alpha P(b) \sum_e P(e) \sum_a P(a|b, e) P(J|a) \sum_m P(m|a)$$

Sum over $m$ is identically 1; $M$ is irrelevant to the query

Thm 1: $Y$ is irrelevant unless $Y \in \text{Ancestors}({X} \cup E)$

Here, $X = \text{JohnCalls}$, $E = \{\text{Burglary}\}$, and $\text{Ancestors}({X} \cup E) = \{\text{Alarm, Earthquake}\}$

so $M$ is irrelevant

(Compare this to backward chaining from the query in Horn clause KBs)
Irrelevant variables contd.

Defn: **moral graph** of Bayes net: marry all parents and drop arrows

Defn: \( A \) is m-separated from \( B \) by \( C \) iff separated by \( C \) in the moral graph

Thm 2: \( Y \) is irrelevant if m-separated from \( X \) by \( E \)

For \( P(\text{JohnCalls}|\text{Alarm}=true) \), both \( \text{Burglary} \) and \( \text{Earthquake} \) are irrelevant
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 $\Rightarrow$ NP-hard
- equivalent to counting 3SAT models $\Rightarrow$ #P-complete
Inference by stochastic simulation

Basic idea:
1) Draw $N$ samples from a sampling distribution $S$
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

Exact inference by variable elimination:
- polytime on polytrees, NP-hard on general graphs
- space = time, very sensitive to topology