Inference in Bayesian networks

Chapter 14.4–5
Outline

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

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

 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 $P(\text{outcome}|action, evidence)$

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

Conditional probability (CP) query is when we ask for the probability of a variable or set of variables given set of events.

Chapter 13 shows that CP can be computed by summing terms from full joint distribution. i.e. A query can be answered using a Bayesian network by computing sums of products of conditional probabilities from network.

\[ P(X|e) = \alpha P(X, e) = \alpha \sum_y P(X, e, y) \]

The complexity of this approach is \( O(n2^n) \)

Chapter 13 shows that CP can be computed by summing terms from full joint distribution.

Inference by enumeration is slightly intelligent way to sum out variables from the joint without actually constructing its explicit representation
Simple query on the burglary network:
\[ P(B|j, m) = P(B, j, m) / P(j, m) \]
\[ = \alpha P(B, j, m) \]
\[ = \alpha \sum_e \sum_a P(B, e, a, j, m) \]

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

Recursive depth-first enumeration: \( O(n) \) space, \( O(d^n) \) time
**Enumeration algorithm**

```plaintext
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 $\{X\} \cup E \cup Y$

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

for each value $x_i$ of $X$ do
    extend $e$ with value $x_i$ for $X$
    $Q(x_i) \gets$ Enumerate-All($\text{Vars}[bn], e$)

return Normalize($Q(X)$)
```

```plaintext
function Enumerate-All($vars, e$) returns a real number

if Empty?($vars$) then return 1.0

$Y \gets$ First($vars$)

if $Y$ has value $y$ in $e$
    then return $P(y | Pa(Y)) \times$ Enumerate-All($\text{Rest}(vars), e$)
else return $\sum_y P(y | Pa(Y)) \times$ Enumerate-All($\text{Rest}(vars), e_y$)
    where $e_y$ is $e$ extended with $Y = y$
```
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

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

\[ = \alpha P(B) \sum_e P(e) \sum_a P(a|B, e) P(j|a) f_M(a) \]

\[ = \alpha P(B) \sum_e P(e) \sum_a P(a|B, e) f_J(a) f_M(a) \]

(make factor) \[= \alpha P(B) \sum_e P(e) f_{\overline{A}JM}(b, e) \] (sum out \(A\); make factor)

\[ = \alpha P(B) f_{E\overline{A}JM}(b) \] (sum out \(E\); make factor)

\[ = \alpha f_B(b) \times f_{E\overline{A}JM}(b) \]
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

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

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., \( f_1(a, b) \times f_2(b, c) = f(a, b, c) \)
### Variable elimination: Basic operations

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>$f_1(A, B)$</th>
<th></th>
<th>$f_2(B, C)$</th>
<th></th>
<th>$f_3(A, B, C)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>.3</td>
<td>T</td>
<td>.2</td>
<td>T</td>
<td>.3 $\times$ .2 = .06</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>.7</td>
<td>T</td>
<td>.8</td>
<td>T</td>
<td>.3 $\times$ .8 = .24</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>.9</td>
<td>T</td>
<td>.6</td>
<td>T</td>
<td>.7 $\times$ .6 = .42</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>.1</td>
<td>F</td>
<td>.4</td>
<td>T</td>
<td>.7 $\times$ .4 = .28</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td></td>
<td>F</td>
<td></td>
<td>F</td>
<td>.9 $\times$ .2 = .18</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td></td>
<td>F</td>
<td>.1</td>
<td>F</td>
<td>.9 $\times$ .8 = .72</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td></td>
<td>F</td>
<td>.1</td>
<td>T</td>
<td>.1 $\times$ .6 = .06</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td></td>
<td>F</td>
<td>.1</td>
<td>F</td>
<td>.1 $\times$ .4 = .04</td>
</tr>
</tbody>
</table>

**Figure 14.10** Illustrating pointwise multiplication: $f_1(A, B) \times f_2(B, C) = f_3(A, B, C)$. 
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 $\leftarrow []$
  for each $var$ in Order($bn$, 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))
Consider the query $P(JohnCalls|\text{Burglary} = \text{true})$

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

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

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

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

so $\text{MaryCalls}$ is 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

1. $A \lor B \lor C$
2. $C \lor D \lor \neg A$
3. $B \lor C \lor \neg D$
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
function `Prior-Sample(bn)` returns an event sampled from `bn`

inputs: `bn`, a belief network specifying joint distribution `P(X_1, \ldots, X_n)`

`x` ← an event with `n` elements

for `i = 1` to `n` do

`x_i` ← a random sample from `P(X_i | parents(X_i))`

given the values of `Parents(X_i)` in `x`

return `x`
Example

- **Cloudy**
  - \( P(C) = 0.50 \)

- **Sprinkler**
  - \( \begin{array}{c|c}
  C & P(S|C) \\
  \hline
  T & 0.10 \\
  F & 0.50 \\
  \end{array} \)

- **Rain**
  - \( \begin{array}{c|c}
  C & P(R|C) \\
  \hline
  T & 0.80 \\
  F & 0.20 \\
  \end{array} \)

- **Wet Grass**
  - \( \begin{array}{c|c|c}
  S & R & P(W|S,R) \\
  \hline
  T & T & 0.99 \\
  T & F & 0.90 \\
  F & T & 0.90 \\
  F & F & 0.01 \\
  \end{array} \)
Example

\[
\begin{array}{c|c}
C & P(S|C) \\
T & .10 \\
F & .50 \\
\end{array}
\]

\[
\begin{array}{c|c}
C & P(R|C) \\
T & .80 \\
F & .20 \\
\end{array}
\]

\[
\begin{array}{c|c|c}
S & R & P(W|S,R) \\
T & T & .99 \\
T & F & .90 \\
F & T & .90 \\
F & F & .01 \\
\end{array}
\]
Example

**Cloudy**

- $P(C) = 0.50$

|   | $P(S|C)$ |
|---|----------|
| T | 0.10     |
| F | 0.50     |

**Sprinkler**

**Wet Grass**

**Rain**

|   | $P(R|C)$ |
|---|----------|
| T | 0.80     |
| F | 0.20     |

| S | R | $P(W|S,R)$ |
|---|---|-----------|
| T | T | 0.99      |
| T | F | 0.90      |
| F | T | 0.90      |
| F | F | 0.01      |
### Example

**Cloudy**

**Sprinkler**

**Wet Grass**

**Rain**

| C | P(S|C) |
|---|-------|
| T | .10   |
| F | .50   |

| C | P(R|C) |
|---|-------|
| T | .80   |
| F | .20   |

| S | R | P(W|S,R) |
|---|---|---------|
| T | T | .99     |
| T | F | .90     |
| F | T | .90     |
| F | F | .01     |
Example

| C | P(S|C) |
|---|-------|
| T | .10   |
| F | .50   |

| C | P(R|C) |
|---|-------|
| T | .80   |
| F | .20   |

| S | R | P(W|S,R) |
|---|---|---------|
| T | T | .99     |
| T | F | .90     |
| F | T | .90     |
| F | F | .01     |
Example

Cloudy

Sprinkler

Rain

Wet Grass

| C | P(S|C) |
|---|-------|
| T | .10   |
| F | .50   |

| C | P(R|C) |
|---|-------|
| T | .80   |
| F | .20   |

| S | R | P(W|S,R) |
|---|---|---------|
| T | T | .99     |
| T | F | .90     |
| F | T | .90     |
| F | F | .01     |
Example

\[
\begin{array}{c|c}
C & P(S|C) \\
T & .10 \\
F & .50 \\
\end{array}
\]

\[
\begin{array}{c|c}
C & P(R|C) \\
T & .80 \\
F & .20 \\
\end{array}
\]

\[
\begin{array}{c|c|c}
S & R & P(W|S,R) \\
T & T & .99 \\
T & F & .90 \\
F & T & .90 \\
F & F & .01 \\
\end{array}
\]
Probability that \textsc{PriorSample} generates a particular event

\[ S_{PS}(x_1 \ldots x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i)) = P(x_1 \ldots x_n) \]

i.e., the true prior probability

E.g., \[ S_{PS}(t, f, t, t) = 0.5 \times 0.9 \times 0.8 \times 0.9 = 0.324 = P(t, f, t, t) \]

Let \( N_{PS}(x_1 \ldots x_n) \) be the number of samples generated for event \( x_1, \ldots, x_n \)

Then we have

\[ \lim_{N \to \infty} \hat{P}(x_1, \ldots, x_n) = \lim_{N \to \infty} N_{PS}(x_1, \ldots, x_n)/N \]

\[ = S_{PS}(x_1, \ldots, x_n) \]

\[ = P(x_1 \ldots x_n) \]

That is, estimates derived from \textsc{PriorSample} are \textbf{consistent}

Shorthand: \( \hat{P}(x_1, \ldots, x_n) \approx P(x_1 \ldots x_n) \)
Rejection sampling

$\hat{P}(X|e)$ estimated from samples agreeing with $e$

```
function REJECTION-SAMPLING($X, e, bn, N$) returns an estimate of $P(X|e)$

    local variables: $N$, a vector of counts over $X$, initially zero

    for $j = 1$ to $N$ do
        $x \leftarrow$ PRIOR-SAMPLE($bn$)
        if $x$ is consistent with $e$ then
            $N[x] \leftarrow N[x] + 1$ where $x$ is the value of $X$ in $x$
    return NORMALIZE($N[X]$)
```

E.g., estimate $P(Rain|Sprinkler = true)$ using 100 samples

  27 samples have $Sprinkler = true$
  Of these, 8 have $Rain = true$ and 19 have $Rain = false$.

$\hat{P}(Rain|Sprinkler = true) = \text{NORMALIZE}(\langle 8, 19 \rangle) = \langle 0.296, 0.704 \rangle$

Similar to a basic real-world empirical estimation procedure
Analysis of rejection sampling

\[ \hat{P}(X|e) = \alpha N_{PS}(X, e) \quad \text{(algorithm defn.)} \]
\[ = \frac{N_{PS}(X, e)}{N_{PS}(e)} \quad \text{(normalized by } N_{PS}(e)\text{)} \]
\[ \approx \frac{P(X, e)}{P(e)} \quad \text{(property of PRIORSAMPLE)} \]
\[ = \frac{P(X|e)}{P(e)} \quad \text{(defn. of conditional probability)} \]

Hence rejection sampling returns consistent posterior estimates

Problem: hopelessly expensive if \( P(e) \) is small

\( P(e) \) drops off exponentially with number of evidence variables!
Likelihood weighting

Idea: fix evidence variables, sample only nonevidence variables, and weight each sample by the likelihood it accords the evidence

```
function LIKELIHOOD-WEIGHTING(X, e, bn, N) returns an estimate of P(X|e)
    local variables: W, a vector of weighted counts over X, initially zero
    for j = 1 to N do
        x, w ← WEIGHTED-SAMPLE(bn)
        W[x] ← W[x] + w where x is the value of X in x
    return Normalize(W[X])
```

```
function WEIGHTED-SAMPLE(bn, e) returns an event and a weight
    x ← an event with n elements; w ← 1
    for i = 1 to n do
        if X_i has a value x_i in e
            then w ← w × P(X_i = x_i | parents(X_i))
            else x_i ← a random sample from P(X_i | parents(X_i))
    return x, w
```
\[ w = 1.0 \]
Likelihood weighting example

\[ w = 1.0 \]
Likelihood weighting example

\[
\begin{array}{c|c}
C & P(S|C) \\
\hline
T & .10 \\
F & .50 \\
\end{array}
\]

\[
\begin{array}{c|c}
C & P(R|C) \\
\hline
T & .80 \\
F & .20 \\
\end{array}
\]

\[
\begin{array}{c|c|c}
S & R & P(W|S,R) \\
\hline
T & T & .99 \\
T & F & .90 \\
F & T & .90 \\
F & F & .01 \\
\end{array}
\]

\[w = 1.0\]
Likelihood weighting example

\[ w = 1.0 \times 0.1 \]
Likelihood weighting example

\[ w = 1.0 \times 0.1 \]
$w = 1.0 \times 0.1$
Likelihood weighting example

\[ w = 1.0 \times 0.1 \times 0.99 = 0.099 \]
Likelihood weighting analysis

Sampling probability for `WeightedSample` is
\[ S_{WS}(z, e) = \prod_{i=1}^{l} P(z_i | \text{parents}(Z_i)) \]
Note: pays attention to evidence in `ancestors` only
\[ \Rightarrow \text{ somewhere “in between” prior and posterior distribution} \]

Weight for a given sample \( z, e \) is
\[ w(z, e) = \prod_{i=1}^{m} P(e_i | \text{parents}(E_i)) \]

Weighted sampling probability is
\[ S_{WS}(z, e)w(z, e) \]
\[ = \prod_{i=1}^{l} P(z_i | \text{parents}(Z_i)) \prod_{i=1}^{m} P(e_i | \text{parents}(E_i)) \]
\[ = P(z, e) \text{ (by standard global semantics of network)} \]

Hence likelihood weighting returns consistent estimates but performance still degrades with many evidence variables because a few samples have nearly all the total weight