Probabilistic Reasoning
Uncertainty and Bayesian networks

Peter Lucas
IRIS
Institute for Computing and Information Sciences
University of Nijmegen, The Netherlands

Uncertainty in Daily Life

- Empirical evidence:
  "If symptoms of fever, shortness of breath (dyspnoea), and coughing are present, and the patient has recently visited China, then the patient has probably SARS"

- Subjective belief:
  "The Balkenende II government is likely to resign soon"

- Temporal dimension:
  "There is less than 10% chance that the Dutch economy will recover in the next two years"

Uncertainty Representation and Manipulation

- Methods for dealing with uncertainty are not new:
  - 17th century: Fermat, Pascal, Huygens, Leibniz, Bernoulli
  - 18th century: Laplace, De Moivre, Bayes
  - 19th century: Gauss, Boole
  ⇒ you could have contributed too if you had been around

- Most important research question in early AI (1970–1987):
  - How to incorporate uncertainty reasoning in logical deduction?
  ⇒ you could have contributed too if you had been around

Early AI Methods of Uncertainty

- Rule-based uncertainty representation:
  \[ \text{fever} \land \text{dyspnoea} \Rightarrow \text{SARS}_{CF=0.4} \]

- Uncertainty calculus (certainty-factor (CF) model, subjective Bayesian method):
  - \( CF(\text{fever}, B) = 0.6; \ CF(\text{dyspnoea}, B) = 1 \) (\( B \) is background knowledge)
  - Combination functions:
    \[
    \begin{align*}
    \text{CF}(\text{SARS}, \{\text{fever, dyspnoea}\} \cup B) &= 0.4 \cdot \max\{0, \min\{\text{CF}(\text{fever}, B), \text{CF}(\text{dyspnoea}, B)\}\} \\
    &= 0.4 \cdot \max\{0, \min\{0.6, 1\}\} = 0.24
    \end{align*}
    \]
However ...

\[ \text{fever} \land \text{dyspnoea} \Rightarrow \text{SARS}_{CF=0.4} \]

- How likely is the occurrence of fever or dyspnoea given that the patient has SARS?
- How likely is the occurrence of fever or dyspnoea in the absence of SARS?
- How likely is the presence of SARS when just fever is present?
- How likely is no SARS when just fever is present?

### Bayesian Networks

\[ \text{Pr}(\text{CH, FL, RS, DY, FE, TEMP}) \]

- \[ \text{Pr}(\text{FE} = y \mid \text{FL} = y, \text{RS} = y) = 0.95 \]
- \[ \text{Pr}(\text{FE} = y \mid \text{FL} = n, \text{RS} = y) = 0.80 \]
- \[ \text{Pr}(\text{FE} = y \mid \text{FL} = y, \text{RS} = n) = 0.88 \]
- \[ \text{Pr}(\text{FE} = y \mid \text{FL} = n, \text{RS} = n) = 0.001 \]

- \[ \text{Pr}(\text{FL} = y) = 0.1 \]
- \[ \text{Pr}(\text{FL} = y) = 0.9 \]

- \[ \text{Pr}(\text{RS} = y \mid \text{CH} = y) = 0.3 \]
- \[ \text{Pr}(\text{RS} = y \mid \text{CH} = n) = 0.01 \]

- \[ \text{Pr}(\text{TEMP} \leq 37.5 \mid \text{FE} = y) = 0.1 \]
- \[ \text{Pr}(\text{TEMP} \leq 37.5 \mid \text{FE} = n) = 0.99 \]

- \[ \text{Pr}(\text{DY} = y \mid \text{RS} = y) = 0.9 \]
- \[ \text{Pr}(\text{DY} = y \mid \text{RS} = n) = 0.05 \]

- \[ \text{Pr}(\text{CH} = y) = 0.1 \]

### Reasoning: Evidence Propagation

- **Nothing known:**
  - Temperature $>37.5$ °C:
    - I just returned from China:
      - VisitToChina (CH) (yes/no)
      - SARS (RS) (yes/no)
      - fever (FE) (yes/no)
      - TEMP ($\leq 37.5$/$>37.5$)

  - Temperature $\leq 37.5$ °C:
    - VisitToChina (CH) (yes/no)
    - SARS (RS) (yes/no)
    - fever (FE) (yes/no)
    - TEMP ($\leq 37.5$/$>37.5$)

- **Temperature $>37.5$ °C:**
  - Flu (FL) (yes/no)
  - Fever (FE) (yes/no)
  - Dyspnoea (DY) (yes/no)

- **Temperature $\leq 37.5$ °C:**
  - Flu (FL) (yes/no)
  - Fever (FE) (yes/no)
  - Dyspnoea (DY) (yes/no)

- **I just returned from China:**
  - VisitToChina (CH) (yes/no)
  - SARS (RS) (yes/no)
  - Fever (FE) (yes/no)
  - TEMP ($\leq 37.5$/$>37.5$)
Bayesian Network Formally

A Bayesian network (BN) is a pair \( B = (G, \Pr) \), where:

- \( G = (V(G), A(G)) \) is an acyclic directed graph, with
  - \( V(G) = \{X_1, X_2, \ldots, X_n\} \), a set of vertices (nodes);
  - \( X \in V(G) \) corresponds to a random variable \( X \)
  - \( A(G) \subseteq V(G) \times V(G) \) a set of arcs reflecting (conditional) independences among variables

- \( \Pr : \wp(V(G)) \to [0, 1] \) is a joint probability distribution, such that
  \[
  \Pr(V(G)) = \prod_{i=1}^{n} \Pr(X_i | \pi_G(X_i))
  \]
  where \( \pi_G(X_i) \) denotes the set of immediate ancestors (parents) of vertex \( X_i \) in \( G \)

Factorisation

Conditional probability distribution:
\[
\Pr(X_1 | X_2, X_3) = \frac{\Pr(X_1, X_2, X_3)}{\Pr(X_2, X_3)}
\]
\[
\Rightarrow \Pr(X_1, X_2, X_3) = \Pr(X_1 | X_2, X_3) \Pr(X_2 | X_3) \Pr(X_3)
\]

Chain rule yields a factorisation:
\[
\Pr(\bigwedge_{i=1}^{n} X_i) = \prod_{i=1}^{n} \Pr(X_i | \bigwedge_{k=i+1}^{n} X_k)
\]

Independence Representation in Graphs

The set of variables \( X \) is conditionally independent of the set \( Z \) given the set \( Y \), notation \( X \perp \! \! \! \perp Z \mid Y \), iff
\[
\Pr(X \mid Y, Z) = \Pr(X \mid Y)
\]

Three flavours of graph-representation of (in)dependence:
- **Diverging**: \( Y \) blocks \( X \) and \( Z \): \( X \perp \! \! \! \perp Z \mid Y \)
- **Serial**: \( Y \) blocks \( X \) and \( Z \): \( X \perp \! \! \! \perp Z \mid Y \)
- **Converging**: \( Y \) connects \( X \) and \( Z \): \( X \not\perp \! \! \! \perp Z \mid Y \)

Use of Independence Information

General:
\[
\Pr(X_1, X_2, X_3) = \Pr(X_2 \mid X_1, X_3) \Pr(X_3 \mid X_1) \Pr(X_1)
\]

Assume that \( X_2 \perp X_3 \mid X_1 \), then:
\[
\Pr(X_2 \mid X_1, X_3) = \Pr(X_2 \mid X_1)
\]
and
\[
\Pr(X_3 \mid X_1, X_2) = \Pr(X_3 \mid X_1)
\]

Only \( 5 = 2 + 2 + 1 \) probabilities needed for \( \Pr(X_1, X_2, X_3) \)
(instead of 7)
Find the Independences

Examples:
- FLU \perp VisitToChina | ∅
- FLU \perp SARS | ∅
- FLU \perp SARS | FEVER, also FLU \perp SARS | TEMP
- SARS \perp TEMP | FEVER
- VisitToChina \perp DYSPNOEA | SARS

Probabilistic Reasoning

- Interested in marginal probability distributions:
  \[ \Pr(V_i | \mathcal{E}) = \Pr^\mathcal{E}(V_i) \]
  for (possibly empty) evidence \( \mathcal{E} \) (instantiated variables)
- Joint probability distribution \( \Pr(V) \):
  - marginalisation:
    \[ \Pr(W) = \sum_{V \setminus W} \Pr(V) \]
    \[ = \sum_{V \setminus W} \prod_{X \in V} \Pr(X | \pi(X)) \]
  - conditional probabilities and Bayes’ rule:
    \[ \Pr(Y, Z | X) = \frac{\Pr(X | Y, Z) \Pr(Y, Z)}{\Pr(X)} \]
- Many efficient Bayesian reasoning algorithms exist

Naive Probabilistic Reasoning: Evidence

\[ \begin{align*}
X_1 \quad & X_2 \\
\text{y/n} \quad & \text{y/n} \\
X_3 \\n\text{y/n} \\
X_4 \quad & \text{y/n} \\
\Pr(x_4 | x_3) = 0.4 \\
\Pr(x_4 | \neg x_3) = 0.1 \\
\Pr(x_3 | x_1, x_2) = 0.3 \\
\Pr(x_3 | \neg x_1, x_2) = 0.5 \\
\Pr(x_3 | x_1, \neg x_2) = 0.7 \\
\Pr(x_3 | \neg x_1, \neg x_2) = 0.9 \\
\Pr(x_1) = 0.6 \\
\Pr(x_2) = 0.2
\end{align*} \]

\[ \Pr^\mathcal{E}(x_2) = \Pr(x_2 | x_4) = \frac{\Pr(x_4 | x_2) \Pr(x_2)}{\Pr(x_4)} \text{ (Bayes’ rule)} \]

\[ \approx 0.14 \]

Judea Pearl’s Algorithm

- Object-oriented approach: vertices are objects, which have local information and carry out local computations
- Updating of probability distribution by message passing: arcs are communication channels
Data Fusion Lemma

Data fusion:
\[ Pr(\mathcal{E} | V_i) = Pr(V_i | \mathcal{E}) = \alpha \cdot \text{causal info for } V_i \cdot \text{diagnostic info for } V_i = \alpha \cdot \pi(V_i) \cdot \lambda(V_i) \]

where:
- \( \mathcal{E} = \mathcal{E}_V^+ \cup \mathcal{E}_V^- \): evidence
- \( \alpha \): normalisation constant

Problem Solving

Bayesian networks are declarative, i.e.:  
- mathematical basis  
- problem to be solved determined by (1) entered evidence \( \mathcal{E} \) (may include decisions); (2) given hypothesis \( H \):

\[ Pr(H | \mathcal{E}) \quad \text{(cf. KB} \land H \vdash \mathcal{E}) \]

Examples:
- Description of populations  
- Classification and diagnosis: \( D = \arg \max_H Pr(H | \mathcal{E}) \)  
- Temporal reasoning, prediction, what-if scenarios  
- Decision-making based on decision theory

\[ \text{MEU}(D | \mathcal{E}) = \max_{d \in D} \sum_{x \in X_d(V)} u(x) \Pr(x | d, \mathcal{E}) \]

Manual Construction of Bayesian Networks

Qualitative modelling:

- Colonisation by bacterium A  
- Colonisation by bacterium B  
- Colonisation by bacterium C  
- Body response to A  
- Body response to B  
- Body response to C  
- Infection  
- Fever  
- WBC  
- ESR

People become colonised by bacteria when entering a hospital, which may give rise to infection

Bayesian-network Modelling

Qualitative  
- causal modelling  

Quantitative  
- interaction modelling

\[ Pr(\text{Inf} | \text{BR}_A, \text{BR}_B, \text{BR}_C) \]

<table>
<thead>
<tr>
<th>( \text{BR}_A )</th>
<th>( \text{BR}_B )</th>
<th>( \text{BR}_C )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t )</td>
<td>( t )</td>
<td>( t )</td>
</tr>
<tr>
<td>( f )</td>
<td>( f )</td>
<td>( f )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \text{Inf} )</th>
<th>( \text{BR}_A )</th>
<th>( \text{BR}_B )</th>
<th>( \text{BR}_C )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t )</td>
<td>0.8</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>( f )</td>
<td>0.2</td>
<td>0.4</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Causal Independence

\[ \Pr(e \mid C_1, \ldots, C_n) = \sum_{I_1, \ldots, I_n} \Pr(e \mid I_1, \ldots, I_n) \prod_{k=1}^{n} \Pr(I_k \mid C_k) \]

Boolean functions: \( P(E \mid I_1, \ldots, I_n) \in \{0, 1\} \)

Interaction function \( f \), defined such that \( f(I_1, \ldots, I_n) = e \) if \( P(e \mid I_1, \ldots, I_n) = 1 \)

Example: noisy OR

- Interactions among causes: logical OR
- Meaning: presence of any one of the causes \( C_i \) with absolute certainty will cause the effect \( e \) (i.e. \( E = \text{true} \))

\[ \Pr(e \mid C_1, C_2) = \sum_{I_1 \lor I_2 = e} \Pr(e \mid I_1, I_2) \prod_{k=1,2} \Pr(I_k \mid C_k) \]

\[ = \Pr(i_1 \mid C_1) \Pr(i_2 \mid C_2) + \Pr(\neg i_1 \mid C_1) \Pr(i_2 \mid C_2) + \Pr(i_1 \mid C_1) \Pr(\neg i_2 \mid C_2) \]

Assessment of \( O(n) \) instead of \( O(2^n) \) probabilities

Example BN: non-Hodgkin Lymphoma

Bayesian Network Learning

Bayesian network \( B = (G, \Pr) \), with

- digraph \( G = (V(G), A(G)) \), and
- probability distribution \( \Pr \)
Learning Bayesian Networks

Problems:
- for many BNs too many probabilities have to be assessed
- complex BNs do not necessarily yield better classifiers
- complex BNs may yield better estimates of a probability distribution

Solution:
- use simple probabilistic models for classification:
  - naive (independent) form BN
  - Tree-Augmented Bayesian Network (TAN)
  - Forest-Augmented Bayesian Network (FAN)
- use background knowledge and clever heuristics

Naive (independent) form BN

- $C$ is a class variable
- The evidence variables $E_i$ in the evidence $\mathcal{E} \subseteq \{E_1, \ldots, E_m\}$ are conditionally independent given the class variable $C$

This yields:
$$P(C \mid \mathcal{E}) = \frac{P(\mathcal{E} \mid C)P(C)}{P(\mathcal{E})} = \frac{\Pi_{E \in \mathcal{E}} P(E \mid C)}{\sum_C P(\mathcal{E} \mid C)P(C)}$$
as $E_i \perp E_j \mid C$, for $i \neq j$

Classifier: $c_{\text{max}} = \arg \max_C P(C \mid \mathcal{E})$

Learning Structure from Data

Given the following dataset $D$:

<table>
<thead>
<tr>
<th>Student</th>
<th>Gender</th>
<th>IQ</th>
<th>High Mark for Maths</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>male</td>
<td>low</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>female</td>
<td>average</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td>male</td>
<td>high</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>female</td>
<td>high</td>
<td>yes</td>
</tr>
</tbody>
</table>

and the following Bayesian networks:

$G_1$: \( G \rightarrow I \rightarrow A \)

$G_2$: \( G \rightarrow I \rightarrow A \)

$G_3$: \( G \rightarrow I \rightarrow A \)

$G_4$: \( G \rightarrow I \rightarrow A \)

$G_5$: \( G \rightarrow I \rightarrow A \)

Which one is the best?

Quality Measure $Q$

$$Q(G, D) = \log \Pr(G) - |D| \cdot H(G, D) - \frac{1}{2}k \cdot \log |D|$$

where:
- $\Pr(G)$: prior probability of $G$
- $-H(G, D)$: negative value of match
- $-\frac{1}{2}k \cdot \log |D|$: penalty term
Research Issues

Qualitative modelling:
- To determine the structure of a network
- Assessment of $\Pr(V_i \mid \pi(V_i))$
- Enhancement of logical semantics

Learning
- Structure learning: determine the ‘best’ graph topology
- Parameter learning: determine the ‘best’ probability distribution (discrete or continuous)

⇒ you can contribute too …