Introduction to Belief Networks II$^1$

Hennie de Villiers

Computer Science Division, Stellenbosch University

$^1$These slides are adapted from those accompanying the book *Bayesian Reasoning and Machine Learning*. The book and demos can be downloaded from [www.cs.ucl.ac.uk/staff/D.Barber/brml](http://www.cs.ucl.ac.uk/staff/D.Barber/brml). We acknowledge David Barber for providing the original slides.
Graphical Models

- Representation of knowledge ← We are here
- Inference
- Learning
Today

- Uncertain evidence
- Unreliable evidence
- Independence Relationships in Belief Nets
- Causality Traps
Belief Networks (Bayesian Networks)

- Directed acyclic graphs
- Each node: Conditional probability given its parents

Sally’s burglar Alarm is sounding. Has she been Burgled, or was the alarm triggered by an Earthquake? She turns the car Radio on for news of earthquakes.

Joint distribution: Product of the conditional probabilities:

\[ p(A, B, C, D, E) = p(A)p(B)p(C|A, B)p(D|C)p(E|B, C) \]
Example – Part I
Sally’s burglar Alarm is sounding. Has she been Burgled, or was the alarm triggered by an Earthquake? She turns the car Radio on for news of earthquakes.

Choosing an ordering
Without loss of generality, we can write

\[
p(A, R, E, B) = p(A|R, E, B)p(R, E, B)
\]
\[
= p(A|R, E, B)p(R|E, B)p(E, B)
\]
\[
= p(A|R, E, B)p(R|E, B)p(E|B)p(B)
\]

Assumptions:
- The alarm is not directly influenced by any report on the radio,
  \[p(A|R, E, B) = p(A|E, B)\]
- The radio broadcast is not directly influenced by the burglar variable,
  \[p(R|E, B) = p(R|E)\]
- Burglaries don’t directly ‘cause’ earthquakes, \[p(E|B) = p(E)\]

Therefore

\[
p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B)
\]
Example – Part II: Specifying the Tables

\[ p(A|B, E) \]

<table>
<thead>
<tr>
<th>Alarm = 1</th>
<th>Burglar</th>
<th>Earthquake</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9999</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.99</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0.99</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0.0001</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ p(R|E) \]

<table>
<thead>
<tr>
<th>Radio = 1</th>
<th>Earthquake</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The remaining tables are \( p(B = 1) = 0.01 \) and \( p(E = 1) = 0.000001 \). The tables and graphical structure fully specify the distribution.
Example Part III: Inference

Initial Evidence: The alarm is sounding

\[
p(B = 1|A = 1) = \frac{\sum_{E,R} p(B = 1, E, A = 1, R)}{\sum_{B,E,R} p(B, E, A = 1, R)}
\]

\[
= \frac{\sum_{E,R} p(A = 1|B = 1, E)p(B = 1)p(E)p(R|E)}{\sum_{B,E,R} p(A = 1|B, E)p(B)p(E)p(R|E)} \approx 0.99
\]

Additional Evidence: The radio broadcasts an earthquake warning:

A similar calculation gives \( p(B = 1|A = 1, R = 1) \approx 0.01 \).

- Initially: Alarm sounds → Sally thinks that she’s been burgled.
- Now: Earthquake report → Probability of burglary drops dramatically.
- Earthquake ‘explains away’ burglary.
Uncertain Evidence

$y$ has the states $\text{dom}(y) = \{\text{red, blue, green}\}$

**hard evidence**

- Certain variable in particular state.
- Probability mass in one of vector components, $(0, 0, 1)$.

**soft evidence**

- Variable in more than one state
- Strength of our belief given by probabilities.
- For example, if $y$ has the states $\text{dom}(y) = \{\text{red, blue, green}\}$ the vector $(0.6, 0.1, 0.3)$ could represent the probabilities of the respective states.
Uncertain Evidence

Can be achieved using Bayes’ rule.

Writing the soft evidence as $\tilde{y}$, we have

$$p(x|\tilde{y}) = \sum_y p(x|y)p(y|\tilde{y})$$

where $p(y = i|\tilde{y})$ represents the probability that $y$ is in state $i$ under the soft-evidence.
Jeffrey’s rule

For variables $x$, $y$, and $p_1(x, y)$, how do we form a joint distribution given soft-evidence $\tilde{y}$?

Form the conditional

We first define

$$p_1(x|y) = \frac{p_1(x, y)}{\sum_x p_1(x, y)}$$

Define the joint

The soft evidence $p(y|\tilde{y})$ then defines a new joint distribution

$$p_2(x, y|\tilde{y}) = p_1(x|y)p(y|\tilde{y})$$

One can therefore view soft evidence as defining a new joint distribution. We use a dashed circle to represent a variable in an uncertain state.
Revisiting the earthquake scenario, we think we hear the burglar alarm sounding, but are not sure, specifically $p(A = \text{tr}) = 0.7$. For this binary variable case we represent this soft-evidence for the states ($\text{tr, fa}$) as $\tilde{A} = (0.7, 0.3)$. What is the probability of a burglary under this soft-evidence?

$$p(B = \text{tr} | \tilde{A}) = \sum_A p(B = \text{tr} | A) p(A | \tilde{A})$$

$$= p(B = \text{tr} | A = \text{tr}) \times 0.7 + p(B = \text{tr} | A = \text{fa}) \times 0.3 \approx 0.6930$$

This value is lower than 0.99, the probability of being burgled when we are sure we heard the alarm. The probabilities $p(B = \text{tr} | A = \text{tr})$ and $p(B = \text{tr} | A = \text{fa})$ are calculated using Bayes’ rule from the original distribution, as before.
Uncertain Evidence

A different viewpoint
Include “dummy node” \( \tilde{A} \) which CONTROLS \( A \).

Implement logic:
- if \( \tilde{A} = 0 \) \( \implies \) \( A = 0 \).
- if \( \tilde{A} = 1 \) \( \implies \) \( A = 1 \).
- \( p(\tilde{A}) \) encode prior beliefs about \( \tilde{A} \) and hence \( A \).

Conditional distribution:

| \( p(A|\tilde{A}) \) | \( A \) | \( \tilde{A} \) |
|----------------------|------|------|
| 1                    | 0    | 0    |
| 0                    | 0    | 1    |
| 0                    | 1    | 0    |
| 1                    | 1    | 1    |
Uncertain Evidence

Conditional distribution:

\[
\begin{array}{c|cc}
\ p(A|\tilde{A}) & A & \tilde{A} \\
1 & 0 & 0 \\
0 & 0 & 1 \\
0 & 1 & 0 \\
1 & 1 & 1 \\
\end{array}
\]

Prior: \( p_{\tilde{A}}(\tilde{A} = 1) = 0.7 \)

New joint with assumption

\[
p(A|\tilde{A}, B, E) \propto p(A|\tilde{A})p(A|B, E)
\]

\[
p(A, \tilde{A}, R, E, B) \propto p(A|E, B)p(R|E)p(E)p(B)p(A|\tilde{A})p_{\tilde{A}}(\tilde{A})
\]

Inference as usual (keeping in mind normalizing constant)
Uncertain Evidence

\[
p(A, \tilde{A}, R, E, B) = p(A|E, B)p(R|E)p(E)p(B)p(A|\tilde{A})p_{\tilde{A}}(\tilde{A})
\]

\(\tilde{A}\) determines \(A\) uniquely.
Doesn’t this “disconnect” \(B\)?
No, \(\tilde{A}\) is not observed! So,

\[
p(A, R, E, B) \propto \sum_{\tilde{A}} p(A, \tilde{A}, R, E, B)
\]

\[
= p(A|E, B)p(R|E)p(E)p(B) \sum_{\tilde{A}} p(A|\tilde{A})p_{\tilde{A}}(\tilde{A})
\]

\[
= p(A|E, B)p(R|E)p(E)p(B)p_{\tilde{A}}(A)
\]

Just introduced a prior over \(A\), keep in mind normalization...
Unreliable evidence (likelihood evidence)

Under potentially confusing reports, you decide to replace the influence of the radio variable with your own model. You decide that you want the radio evidence to influence the inference 80% towards being an earthquake and 20% to not being an earthquake.

\[ p(R|E) \rightarrow p(R|E) = \begin{cases} 0.8 & E = \text{tr} \\ 0.2 & E = \text{fa} \end{cases} \]

This then gives a distribution, with \( R \) in an arbitrary fixed state,

\[ p(B, E, A, R) = p(A|B, E)p(B)p(E)p(R|E) \]

This can then be used to form inference.
Examples of Belief Networks in Machine Learning

Prediction (discriminative)
\[ p(\text{class}|\text{input}) \]

Prediction (generative)
\[ p(\text{class}|\text{input}) \propto p(\text{input}|\text{class})p(\text{class}) \]

Time-series

Unsupervised learning
\[ p(\text{data}) = \sum_{\text{latent}} p(\text{data}|\text{latent})p(\text{latent}). \]

And many more
Independence ⊥ in Belief Networks – Part I

- Graphical models encode *independence* relationships between variables.
- Unconditional independence \( A \perp B|\emptyset, P(A, B) = P(A)P(B) \)
- Conditional independence \( A \perp B|C, P(A, B|C) = P(A|C)P(B|C) \)
- A graphical model encodes a set of such independence relationships.
- Can belief networks encode all possible sets of independence relationships? ...
Independence $\perp \!
\!
\!\perp$ in Belief Networks – Part I

All belief networks with three nodes and two links:

(a) $A \perp \perp B \mid C$

(b) $A \perp \!
\!
\!\perp B \mid C$

(c) $A \perp \perp B \mid C$

(d) $A \not\perp \!
\!
\!\perp B \mid C$

- In (a), (b) and (c), $A, B$ are conditionally independent given $C$.

(a) $p(A, B \mid C) = \frac{p(A, B, C)}{p(C)} = \frac{p(A \mid C)p(B \mid C)p(C)}{p(C)} = p(A \mid C)p(B \mid C)$

(b) $p(A, B \mid C) = \frac{p(A)p(C \mid A)p(B \mid C)}{p(C)} = \frac{p(A, C)p(B \mid C)}{p(C)} = p(A \mid C)p(B \mid C)$

(c) $p(A, B \mid C) = \frac{p(A \mid C)p(C \mid B)p(B)}{p(C)} = \frac{p(A \mid C)p(B, C)}{p(C)} = p(A \mid C)p(B \mid C)$

- In (d) the variables $A, B$ are conditionally dependent given $C$, $p(A, B \mid C) \propto p(C \mid A, B)p(A)p(B)$. 
In (a), (b) and (c), the variables $A, B$ are marginally dependent.

In (d) the variables $A, B$ are marginally independent.

$$p(A, B) = \sum_C p(A, B, C) = \sum_C p(A)p(B)p(C|A, B) = p(A)p(B)$$
Independence \( \perp \) in Belief Networks – Part III

\[ A \perp B \]

\begin{align*}
\text{(a)} & \quad A \perp B | D \\
\text{(b)} & \quad A \perp B
\end{align*}

- Careful! Descendants can matter! On right, \( p(D|C) \) can be marginalized away with \( \sum_D \), not so with left.
Collider

A collider contains two or more incoming arrows along a chosen path.

Summary of two previous slides:

If $C$ has more than one incoming link, then $A \perp \perp B$ and $A \nparallel B \mid C$. In this case $C$ is called **collider**.

If $C$ has at most one incoming link, then $A \perp \perp B \mid C$ and $A \nparallel B$. In this case $C$ is called **non-collider**.
General Rule for Independence in Belief Networks

Given three sets of nodes $\mathcal{X}, \mathcal{Y}, \mathcal{C}$, if all paths from any element of $\mathcal{X}$ to any element of $\mathcal{Y}$ are blocked by $\mathcal{C}$, then $\mathcal{X}$ and $\mathcal{Y}$ are conditionally independent given $\mathcal{C}$.

A path $\mathcal{P}$ is blocked by $\mathcal{C}$ if at least one of the following conditions is satisfied:

1. there is a collider in the path $\mathcal{P}$ such that neither the collider nor any of its descendants is in the conditioning set $\mathcal{C}$.

2. there is a non-collider in the path $\mathcal{P}$ that is in the conditioning set $\mathcal{C}$.

d-connected/separated
We use the phrase ‘d-connected’ if there is a path from $\mathcal{X}$ to $\mathcal{Y}$ in the ‘connection’ graph – otherwise the variable sets are ‘d-separated’. Note that d-separation implies that $\mathcal{X} \perp \perp \mathcal{Y} | \mathcal{Z}$, but d-connection does not necessarily imply conditional dependence.
The ‘connection’-graph

All paths in the connection graph need to be blocked to obtain $\perp$:

- $A \perp D \mid B, C$
- $A \perp D$

Non-collider in the conditioning set blocks a path. Collider outside the conditioning set blocks a path.

$B \perp C \mid A$
$B \perp C \mid A$

All paths need to be blocked to obtain $\perp$. 
Markov Equivalence

**skeleton**
Formed from a graph by removing the arrows

**immorality**
An immorality in a DAG is a configuration of three nodes, $A, B, C$ such that $C$ is a child of both $A$ and $B$, with $A$ and $B$ not directly connected. *rolls eyes*

**Markov equivalence**
Two graphs represent the same set of independence assumptions if and only if they have the same skeleton and the same set of immoralities.

![Diagram](attachment:image.png)
Limitations of expressibility

\[ p(t_1, t_2, y_1, y_2, h) = p(t_1)p(t_2)p(y_1|t_1, h)p(y_2|t_2, h) \]

\[ t_1 \perp \perp t_2, y_2, \quad t_2 \perp \perp t_1, y_1 \]

\[ p(t_1, t_2, y_1, y_2) = p(t_1)p(t_2) \sum_h p(y_1|t_1, h)p(y_2|t_2, h) \]

Still holds that:

\[ t_1 \perp \perp t_2, y_2, \quad t_2 \perp \perp t_1, y_1 \]

No Belief network on \( t_1, t_2, y_1, y_2 \) can’t represent all the conditional independence statements contained in \( p(t_1, t_2, y_1, y_2) \). Sometimes we can extend the representation by adding for example a bidirectional link, but this is no longer a Belief Network.
### Causality

<table>
<thead>
<tr>
<th></th>
<th>Recovered</th>
<th>Not Recovered</th>
<th>Rec. Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Males</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Given Drug</td>
<td>18</td>
<td>12</td>
<td>60%</td>
</tr>
<tr>
<td>Not Given Drug</td>
<td>7</td>
<td>3</td>
<td>70%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Recovered</th>
<th>Not Recovered</th>
<th>Rec. Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Females</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Given Drug</td>
<td>2</td>
<td>8</td>
<td>20%</td>
</tr>
<tr>
<td>Not Given Drug</td>
<td>9</td>
<td>21</td>
<td>30%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Recovered</th>
<th>Not Recovered</th>
<th>Rec. Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Combined</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Given Drug</td>
<td>20</td>
<td>20</td>
<td>50%</td>
</tr>
<tr>
<td>Not Given Drug</td>
<td>16</td>
<td>24</td>
<td>40%</td>
</tr>
</tbody>
</table>

### Simpson’s paradox

For the males, it’s best not to give the drug. For the females, it’s also best not to give the drug. However, for the combined data, it’s best to give the drug!
Resolving the paradox

We can write the distribution as

\[ p(S, D, R) = p(R|S, D)p(D|S)p(S) \]

Our observational calculation computed \( p(R|S, D) \) and \( p(R|D) \) using the above distribution.

Sampling from the distribution

The above formula suggests that we would first choose a patient’s sex (the term \( p(S) \)) then decide whether or not to give the drug (the term \( p(D|S) \)).
Resolving the paradox

interventional calculation

We must use a distribution that is consistent with an interventional experiment. In this case, the term \( p(D|S) \) should play no role. That is, we need to consider a modified distribution (conditioned on the drug)

\[
\tilde{p}(S, R|D) = p(R|S, D)p(S)
\]

\[
p(R||D) = \sum_S \tilde{p}(S, R|D) = \sum_S p(R|S, D)p(S)
\]

This gives the non-paradoxical result:

\[
p(\text{recovery}|\text{drug}) = 0.6 \times 0.5 + 0.2 \times 0.5 = 0.4
\]
\[
p(\text{recovery}|\text{no drug}) = 0.7 \times 0.5 + 0.3 \times 0.5 = 0.5
\]

The moral of the story is that you have to make the distribution match the experimental conditions, otherwise apparent paradoxes may arise.
For next time, read Chapter 4...