Introduction to Belief Networks

Prof. Dr. H. H. Takada

Quantitative Research – Itaú Asset Management
Institute of Mathematics and Statistics – University of São Paulo

Belief Networks (Bayesian Networks)

Definition: Belief Network

A belief network is a distribution of the form

$$p\left(x_1, \ldots, x_D\right) = \prod_{i=1}^{D} p\left(x_i | \mathrm{pa}\left(x_i\right)\right),$$

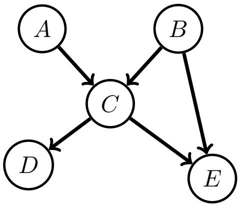
where $\mathrm{pa}\left(x_i\right)$ represent the parental variables of variable x_i. A belief network is a Directed Acyclic Graph (DAG) in which each node has associated the conditional probability of the node given its parents.

Belief Networks (Bayesian Networks)

Example

The joint distribution is obtained by taking the product of the conditional probabilities:

$$p(A,B,C,D,E) = p(A)p(B)p(C|A,B)p(D|C)p(E|B,C)$$



$p(E|B,C)$

Example – Part I

Sally's burglar **A**larm is sounding. Has she been **B**urgled, or was the alarm triggered by an **E**arthquake? She turns the car **R**adio on for news of earthquakes.

Choosing an ordering

Without loss of generality, we can write

$$\begin{aligned} p(A, R, E, B) &= p(A|R, E, B)p(R, E, B) \\ &= p(A|R, E, B)p(R|E, B)p(E, B) \\ &= p(A|R, E, B)p(R|E, B)p(E|B)p(B) \end{aligned}$$

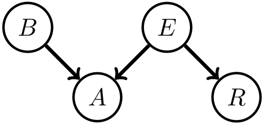
Assumptions:

- 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



$p(A|B,E)$

Alarm = 1	Burglar	Earthquake
0.9999	1	1
0.9900	1	0
0.9900	0	1
0.0001	0	0

$p(R|E)$

Radio = 1	Earthquake
1	1
0	0

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

$$
\begin{aligned}
p(B=1|A=1) &= \frac{\sum_{E,R} p(B=1,E,A=1,R)}{\sum_{B,E,R} p(B,E,A=1,R)} \\
&= \frac{\sum_{E,R} p(A=1|B=1,E)p(B=1)p(E)p(R|E)}{\sum_{B,E,R} p(A=1|B,E)p(B)p(E)p(R|E)} \approx 0.99
\end{aligned}
$$

Additional Evidence: The radio broadcasts an earthquake warning:

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

Initially, because the alarm sounds, Sally thinks that she's been burgled. However, this probability drops dramatically when she hears that there has been an earthquake.

The earthquake 'explains away' to an extent the fact that the alarm is ringing.

Example Part IV: Implementation

Matlab

```
>> setup
>> demoBurglar
```

Uncertain Evidence

In soft/uncertain evidence the variable is in more than one state, with the strength of our belief about each state being given by probabilities. For example, if y has the states $\mathrm{dom}(y) = \{\text{red}, \text{blue}, \text{green}\}$ the vector $(0.6, 0.1, 0.3)$ could represent the probabilities of the respective states.

---

hard evidence

We are certain that a variable is in a particular state. In this case, all the probability mass is in one of the vector components, $(0, 0, 1)$.

---

inference

Inference with soft-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 $p(y = \text{i}|\tilde{y})$ 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

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

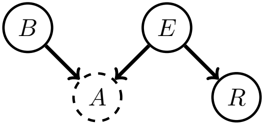
Define the joint

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.

Uncertain evidence example



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 (tr, fa) as $\tilde{A} = (0.7, 0.3)$. What is the probability of a burglary under this soft-evidence?

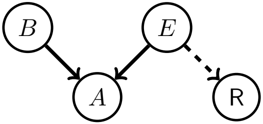
$$\begin{aligned} p(B = \text{tr}|\tilde{A}) &= \sum_A p(B = \text{tr}|A) p(A|\tilde{A}) \\ &= p(B = \text{tr}|A = \text{tr}) \times 0.7 + p(B = \text{tr}|A = \text{fa}) \times 0.3 \approx 0.6930 \end{aligned}$$

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.

Unreliable evidence (likelihood evidence or virtual 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.

$$p(R|E) \to p(\mathsf{R}|E) = \begin{cases} 0.8 & E = \mathsf{tr} \\ 0.2 & E = \mathsf{fa} \end{cases}$$



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

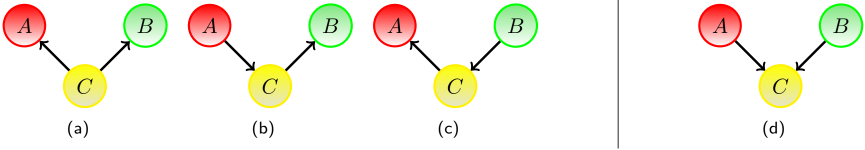
$$p(B, E, A, \mathsf{R}) = p(A|B, E)p(B)p(E)p(\mathsf{R}|E)$$

This can then be used to form inference.

Independence $\perp\!\!\!\perp$ in Belief Networks – Part I

All belief networks with three nodes and two links:

$A \perp\!\!\!\perp B \,|\, C$ $\qquad\qquad$ $A \not\!\perp\!\!\!\perp B \,|\, C$



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

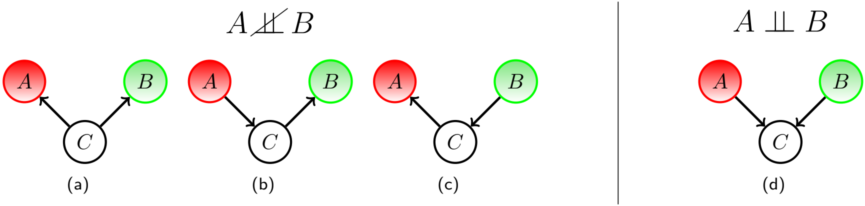
  (a) $p(A,B|C) = \frac{p(A,B,C)}{p(C)} = \frac{p(A|C)p(B|C)p(C)}{p(C)} = p(A|C)p(B|C)$

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

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

- In (d) the variables A, B are conditionally dependent given C, $p(A,B|C) \propto p(C|A,B)p(A)p(B)$.

Independence $\perp\!\!\!\perp$ in Belief Networks – Part II

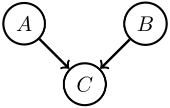


- In (a), (b) and (c), the variables A, B are marginally dependent.
- In (d) the variables A, B are marginally independent.

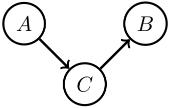
$$p(A, B) = \sum_C p(A, B, C) = \sum_C p(A)p(B)p(C|A, B) = p(A)p(B)$$

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 $A \not\perp\!\!\!\perp B \mid C$. In this case C is called collider.

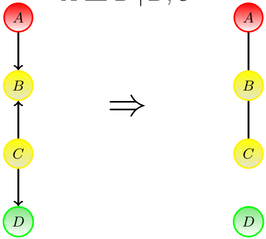


If C has at most one incoming link, then $A \perp\!\!\!\perp B \mid C$ and $A \not\perp\!\!\!\perp B$. In this case C is called non-collider.

The 'connection'-graph

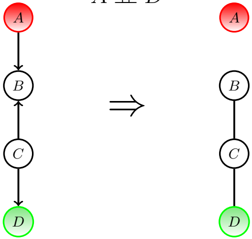
All paths in the connection graph need to be blocked to obtain $\perp\!\!\!\perp$:

$A \perp\!\!\!\perp D \,|\, B, C$



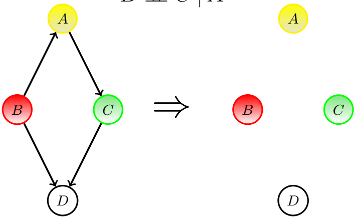
non-collider in the conditioning set blocks a path

$A \perp\!\!\!\perp D$

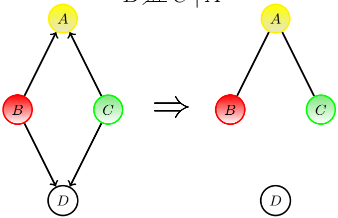


collider outside the conditioning set blocks a path

$B \perp\!\!\!\perp C \,|\, A$



$B \not\!\perp\!\!\!\perp C \,|\, A$

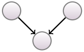


all paths need to be blocked to obtain $\perp\!\!\!\perp$

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.

Markov Equivalence

skeleton

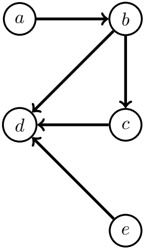
Formed from a graph by removing the arrows

immorality

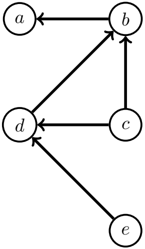
An immorality in a DAG is a configuration of three nodes, A,B,C such that C is a child of both A and B, with A and B not directly connected.

Markov equivalence

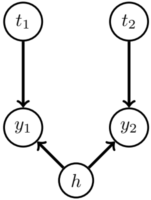
Two graphs represent the same set of independence assumptions if and only if they have the same skeleton and the same set of immoralities.



$=?$



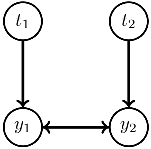
Limitations of expressibility



$$p(t_1, t_2, y_1, y_2, h) = p(t_1)p(t_2)p(y_1|t_1, h)p(y_2|t_2, h)p(h)$$

$$t_1 \perp\!\!\!\perp t_2, y_2, \qquad t_2 \perp\!\!\!\perp t_1, y_1$$

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



Still holds that:

$$t_1 \perp\!\!\!\perp t_2, y_2, \qquad t_2 \perp\!\!\!\perp t_1, y_1$$

No Belief network on t_1, t_2, y_1, y_2 can 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

Males	Recovered	Not Recovered	Rec. Rate
Given Drug	18	12	60%
Not Given Drug	7	3	70%

Females	Recovered	Not Recovered	Rec. Rate
Given Drug	2	8	20%
Not Given Drug	9	21	30%

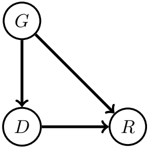
Combined	Recovered	Not Recovered	Rec. Rate
Given Drug	20	20	50%
Not Given Drug	16	24	40%

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



G: Gender; D: Drug; R: Recovered

---

observational calculation

$$p(G, D, R) = p(R|G, D)p(D|G)p(G)$$

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

Sampling from the distribution

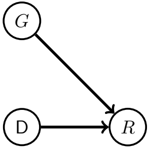
The above formula suggests that we would first chose a gender (the term $p(G)$) then decide whether or not to give the drug (the term $p(D|G)$).

Resolving the paradox

interventional calculation

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

$$\tilde{p}(G, R|D) = p(R|G, D)p(G)$$



$$p(R||D) = \sum_G \tilde{p}(G, R|D) = \sum_G p(R|G, D)p(G)$$

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.