Introduction to Belief Networks[1]

Dmitry Adamskiy, David Barber

University College London

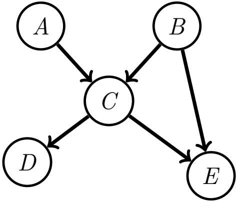
[1] These slides accompany the book *Bayesian Reasoning and Machine Learning*. The book and demos can be downloaded from `www.cs.ucl.ac.uk/staff/D.Barber/brml`. Feedback and corrections are also available on the site. Feel free to adapt these slides for your own purposes, but please include a link the above website.

Belief Networks (Bayesian Networks)

A belief network is a directed acyclic graph in which each node is associated with the conditional probability of the node given its parents.

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

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

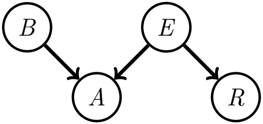
Assumptions:

- 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
0.99	1	0
0.99	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.

Uncertain Evidence

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

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

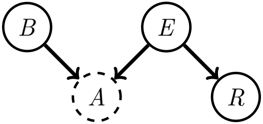
Define the joint

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?

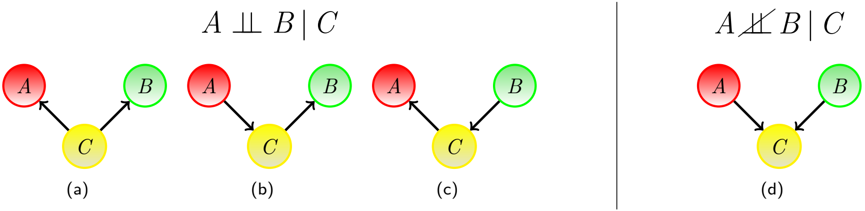
$$\begin{aligned} p(B = \text{tr}|\tilde{A}) &= \sum_A p(B = \text{tr}|A)p(A|\tilde{A}) \\ &= p(B = \text{tr}|A = \text{tr}) \times 0.7 + p(B = \text{tr}|A = \text{fa}) \times 0.3 \approx 0.6930 \end{aligned}$$

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.

Independence in Belief Networks

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

All belief networks with three nodes and two links:



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

$A \not\!\perp\!\!\!\perp B \mid C$

(a) (b) (c) (d)

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

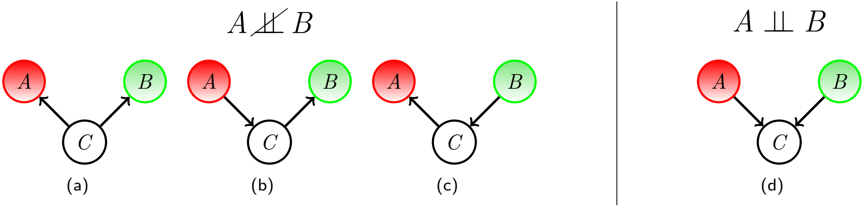
  (a) $p(A,B|C) = \frac{p(A,B,C)}{p(C)} = \frac{p(A|C)p(B|C)p(C)}{p(C)} = p(A|C)p(B|C)$

  (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



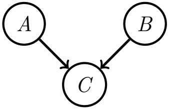
(a) (b) (c) (d)

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

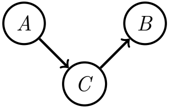
$$p(A, B) = \sum_C p(A, B, C) = \sum_C p(A)p(B)p(C|A, B) = p(A)p(B)$$

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.

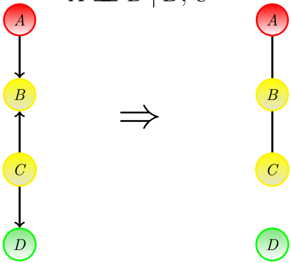


If C has at most one incoming link, then $A \perp\!\!\!\perp B \mid C$ and $A \not\!\perp\!\!\!\perp B$. In this case C is called non-collider.

The ‘connection’-graph (an informal and intuitive idea)

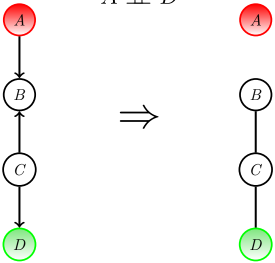
All paths in the connection graph need to be blocked to obtain $\perp\!\!\!\perp$:

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



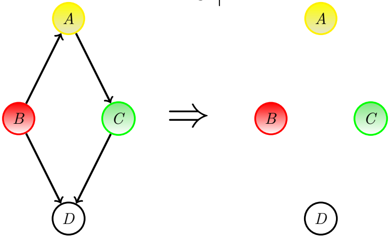
non-collider in the conditioning set blocks a path

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

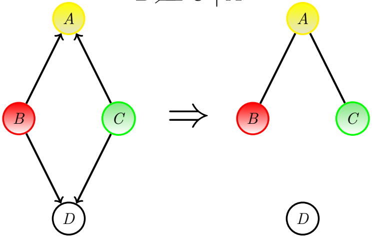


collider outside the conditioning set blocks a path

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



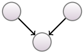
$B \not\!\perp\!\!\!\perp C \mid A$



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 $\mathcal{C}$.

Independence of $\mathcal{X}$ and $\mathcal{Y}$

- When the conditioning set is empty $\mathcal{C} = \emptyset$, then a path $\mathcal{P}$ from an element of $\mathcal{X}$ to an element of $\mathcal{Y}$ is blocked if there is a collider on the path.
- Hence $\mathcal{X}$ and $\mathcal{Y}$ are independent if every path from any element of $\mathcal{X}$ to any element of $\mathcal{Y}$ has a collider.

d-connected/separated?

d-connected

- We use the term that $\mathcal{X}$ and $\mathcal{Y}$ are ‘d-connected’ by $\mathcal{Z}$ if there is any path from $\mathcal{X}$ to $\mathcal{Y}$ that is not blocked by $\mathcal{Z}$. If $\mathcal{Z}$ is the empty set then we just say that $\mathcal{X}$ and $\mathcal{Y}$ are d-connected.
- If all of the paths are blocked then we say $\mathcal{X}$ and $\mathcal{Y}$ are ‘d-separated’ by $\mathcal{Z}$.

Separation and Independence

- Note first that d-separation and connection are properties of the graph (not of the distribution).
- d-separation implies that $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$, but d-connection does not necessarily imply conditional dependence.
- That is, for any distribution in which $\mathcal{X}$ and $\mathcal{Y}$ are ‘d-separated’ by $\mathcal{Z}$, then no matter what the settings of the conditional tables are, then conditional independence holds, namely $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$.

Graphical versus Distributional Independence

- Consider the distribution $p(a|b)p(b|c)p(c)$. Here it seems obvious 'from the graph' that a and c are dependent (indeed they are d-connected).
- However, there are (even non-trivial) settings of the distribution tables $p(a|b)$, $p(b|c)$ and $p(c)$ such that a and c are independent. This is far from obvious.
- In this case a and c are 'graphically dependent' (they are d-connected), this doesn't necessarily mean that a and c are dependent in all distributions that can be represented by the graph.
- On the other hand, consider $p(b|a, c)p(a)p(c)$. Here, a and c are d-separated and, no matter what the table settings are, a and c are always independent. In this case 'graphical independence' (d-separation) implies distributional independence.

Markov Equivalence

skeleton

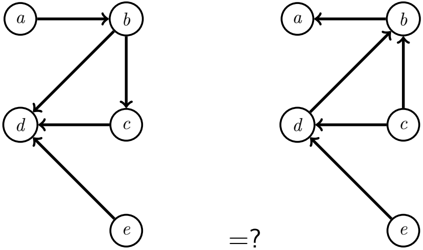
Formed from a graph by removing the arrows

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.



What can Belief Networks represent?

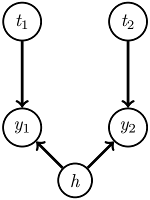
BNs can represent both everything and nothing

- Given any distribution $p(x_1, \ldots, x_N)$ we can always write this in the form

$$\begin{aligned}
p(x_1, \ldots, x_N) &= p(x_N|x_{1:N-1})p(x_{1:N-1}) \\
&= p(x_N|x_{1:N-1})p(x_{N-1}|x_{1:N-2})p(x_{1:N-2}) \\
&= \prod_{n=1}^{N} p(x_n|x_{1:n-1})
\end{aligned}$$

- Represented as a graph, this forms a 'cascade' in which every variable is connected to all preceding variables.
- The above is simply a 'numerical' statement. It does not mean that we can represent whatever independence statements are present (numerically) in p by a BN.
- Fundamentally, the actual numerical distribution p contains much more information than a graph can represent.

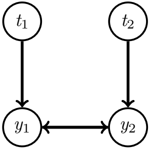
Limitations of expressibility



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

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

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



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.

Summary

- BNs are conventient ways to graphical represent conditional independence statements.
- Essentially they are statements about the way the joint distribution factorises.
- They do not say everything about the numerical content of the distirbution – they only express certain factorisation assumptions.
- There is a general rule that can ascertain whether x and y are independent (conditioned on z) and this relates to examining paths on the graph.
- There is a distinction between ‘graphical’ and ‘numerical’ independence.
- If the graph implies x and y are independence then this will numerically hold in all distributions consistent with the BN.
- However, if the graph imples that x and y are ‘graphically dependence’ (d-connected) then this does not imply that x and y will be numerically dependent in all distributions consistent with the given BN.
- BNs cannot represent all kinds of independence we may wish to encode.
- BNs are just one form of Graphical Model. Each class of Graphical Model is able to encode different properties of the numerical distriubution; BNs focus on representing conditional independence.