

Introduction to Belief Networks [Redes Bayesianas] (Chapter 3)¹

David Barber

(adapted by Miguel Palomino)

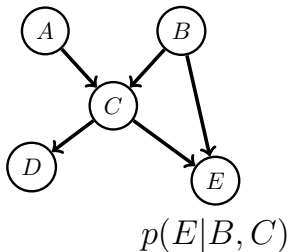
¹These slides accompany the book *Bayesian Reasoning and Machine Learning*. The book and demos can be downloaded from www.cs.ucl.ac.uk/staff/D.Barber/brml. Feedback and corrections are also available on the site. Feel free to adapt these slides for your own purposes, but please include a link the above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

Belief Networks (Bayesian Networks)

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

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

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



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

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

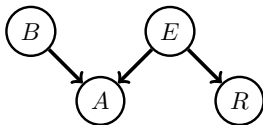
Assumptions:

- 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' ("explica alternativamente") to an extent the fact that the alarm is ringing.

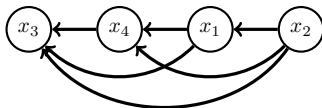
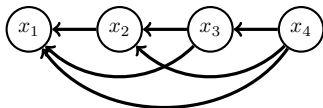
Redes bayesianas

Una red bayesiana es una distribución de la forma

$$p(x_1, \dots, x_D) = \prod_i^D p(x_i | \text{pa}(x_i))$$

Toda distribución se puede escribir como una RB, de manera no única.

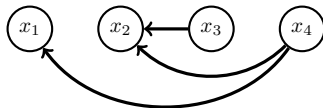
$$\begin{aligned} p(x_1, x_2, x_3, x_4) &= p(x_1 | x_2, x_3, x_4) p(x_2 | x_3, x_4) p(x_3 | x_4) p(x_4) \\ &= p(x_3 | x_4, x_1, x_2) p(x_4 | x_1, x_2) p(x_1 | x_2) p(x_2) \end{aligned}$$



Estos grafos “en cascada” son unilateralmente conexos.

Redes bayesianas: independencia condicional

El grafo subyacente en la RB codifica asunciones de independencia condicional entre los nodos. (¡Pero no la dependencia!)



Operando:

$$\begin{aligned} p(x_1, x_2 | x_4) &= \frac{1}{p(x_4)} \sum_{x_3} p(x_1, x_2, x_3, x_4) = \frac{1}{p(x_4)} \sum_{x_3} p(x_1 | x_4) p(x_2 | x_3, x_4) p(x_3) p(x_4) \\ &= p(x_1 | x_4) \sum_{x_3} p(x_2 | x_3, x_4) p(x_3) \end{aligned}$$

$$p(x_2 | x_4) = \sum_{x_3} p(x_2 | x_3, x_4) p(x_3)$$

por lo que $p(x_1, x_2 | x_4) = p(x_1 | x_4) p(x_2 | x_4)$ y $x_1 \perp\!\!\!\perp x_2 \mid x_4$.

Queremos un algoritmo que nos permita estudiar la independencia condicional sin cálculos tediosos.

Examples of Belief Networks in Machine Learning

Prediction (discriminative)

$$p(class|input)$$

Prediction (generative)

$$p(class|input) \propto p(input|class)p(class)$$

Time-series

Markov chains, Hidden Markov Models.

Unsupervised learning

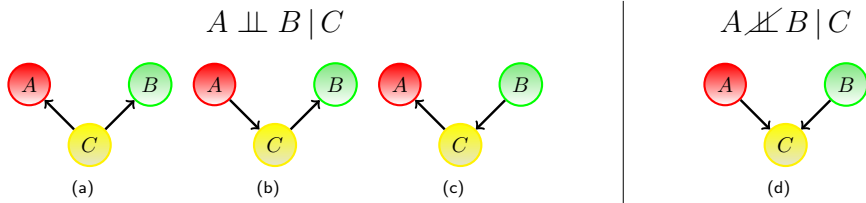
$$p(data) = \sum_{latent} p(data|latent)p(latent).$$

And many more

Personally I find the framework very useful for understanding and rationalising the many different approaches in machine learning and related areas.

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

All belief networks with three nodes and two links:



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

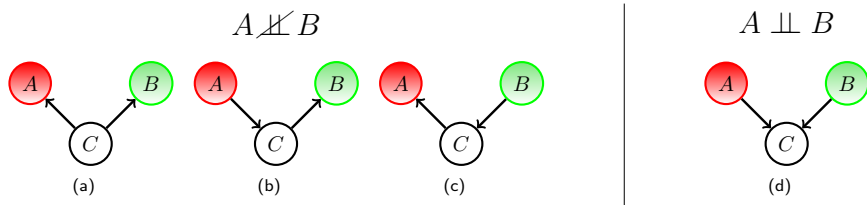
$$(a) \quad p(A, B|C) = \frac{p(A, B, C)}{p(C)} = \frac{p(A|C)p(B|C)p(C)}{p(C)} = p(A|C)p(B|C)$$

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

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

- 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

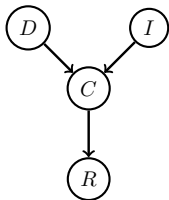


- In (a), (b) and (c), the variables A, B are marginally dependent.
- In (d) the variables A, B are marginally independent.

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

Collider

Dado un camino \mathcal{P} , un nodo **colisionador** es un nodo c en \mathcal{P} con vecinos a y b en \mathcal{P} tales que $a \rightarrow c \leftarrow b$.



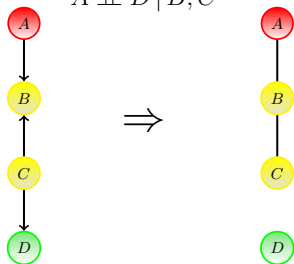
D : dificultad
 I : inteligencia
 C : calificación
 R : carta de recomendación

Se tiene que $D \perp\!\!\!\perp I$ pero ~~$D \perp\!\!\!\perp I \mid C$~~ y ~~$D \perp\!\!\!\perp I \mid R$~~ .

The 'connection'-graph

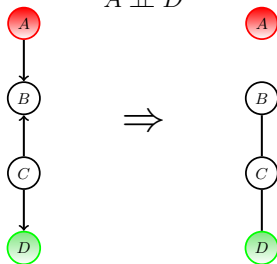
All paths in the connection graph need to be blocked to obtain $\perp\!\!\!\perp$:

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



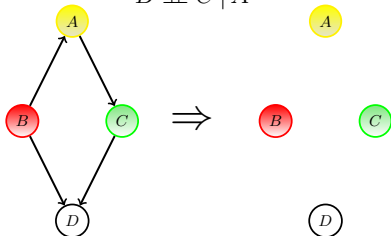
non-collider in the conditioning set blocks a path

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

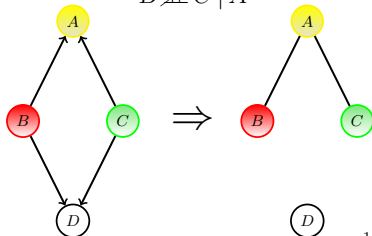


collider outside the conditioning set blocks a path

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




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



General Rule for Independence in Belief Networks

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} .
-

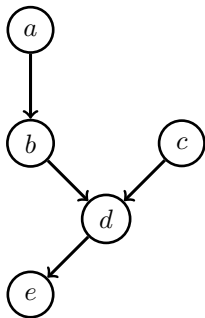
d-connected/separated

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'.

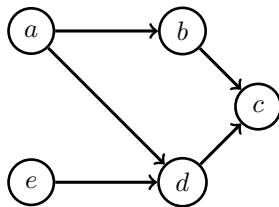
Theorem. Given three sets of nodes $\mathcal{X}, \mathcal{Y}, \mathcal{C}$, if \mathcal{X} is d-separated from \mathcal{Y} by \mathcal{C} , then \mathcal{X} and \mathcal{Y} are conditionally independent given \mathcal{C} .

Note that d-separation implies that $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$, but d-connection does not necessarily imply conditional dependence.

d-separación



$a \perp\!\!\!\perp e \mid b$
(d es colisionador en $a-b-d-c$,
pero no en $a-b-d-e$)



~~$a \perp\!\!\!\perp e \mid c$~~

Markov Equivalence

skeleton

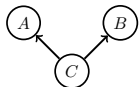
Formed from a graph by removing the arrows

immorality

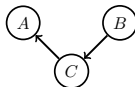
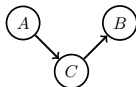
An immorality in a DAG is a configuration of three nodes, A, B, C such that C is a child of both A and B , with A and B not directly connected.

Markov equivalence

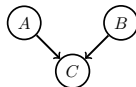
Two graphs represent the same set of independence assumptions if and only if they have the same skeleton and the same set of immoralities.



no immoralities

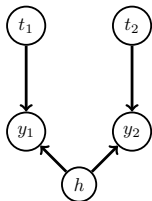


an immorality



beware the causal interpretation!

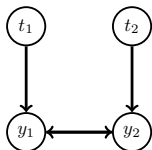
Limitations of expressibility



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

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

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



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 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!

Resolución de la paradoja

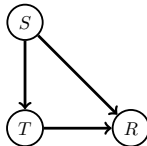
Los datos anteriores se pueden resumir así:

$$p(R = 1|S = h, T = 1) = 0.6 < 0.7 = p(R = 1|S = h, T = 0)$$

$$p(R = 1|S = m, T = 1) = 0.2 < 0.3 = p(R = 1|S = h, T = 0)$$

$$p(R = 1|T = 1) = 0.5 > 0.4 = p(R = 1|T = 0)$$

cálculo observacional



En particular:

$$p(R|T) = \sum_S p(S, R|T) = \sum_S p(R|S, T)p(S|T)$$

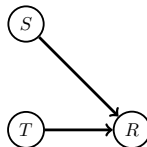
Intervenciones

cálculo de intervenciones

En realidad nos interesa

$$p(R = 1 | \text{do}(T = 1)) \quad \text{intervenimos y hacemos } T = 1$$

Si $T = 1$, T deja de depender de S y tenemos una nueva distribución \tilde{p} :



$$p(R | \text{do}(T)) = \sum_S \tilde{p}(S, R | T) = \sum_S p(R | S, T) p(S)$$

Esto resulta en el siguiente resultado, no paradójico:

$$p(R = 1 | \text{do}(T = 1)) = 0.6 \times 0.5 + 0.2 \times 0.5 = 0.4$$

$$p(R = 1 | \text{do}(T = 0)) = 0.7 \times 0.5 + 0.3 \times 0.5 = 0.5$$

Moraleja: las redes bayesianas, *per se*, no codifican relaciones causales.