Sapienza University of Rome

Master in Artificial Intelligence and Robotics
Master in Engineering in Computer Science

Machine Learning

A.Y. 2020/2021

Prof. L. Iocchi, F. Patrizi

4. Probability and Bayes Networks

L. Iocchi, F. Patrizi

Outline

- Uncertainty
- Probability
- Syntax and Semantics
- Inference
- Independence and Bayes' Rule

Uncertainty

Consider action A_t = leave for airport t minutes before flight.

Will A_t get me there on time? this is a very difficult problem to predict

Problems:

- partial observability (road state, other drivers' plans, etc.)
- noisy sensors (traffic reports)
- uncertainty in action outcomes (flat tire, etc.)
- complexity of modelling and predicting traffic

Uncertainty

Hence a purely logical approach either

- risks falsehood: "A_{25} will get me there on time"
- leads to conclusions that are too weak for decision making: "A_{25} will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact etc etc."
- leads to non-optimal decisions (A_{1440} might reasonably be said to get me there on time, but I'd have to stay overnight in the airport ...)

Probability

Representation of uncertainty with probabilities.

Given the available evidence, A_{25} will get me at the airport on time with probability 0.04

Given the available evidence, A_{60} will get me at the airport on time with probability 0.85

Given the available evidence, A_{1440} will get me at the airport on time with probability 0.999

Probability

Sample space

- Ω *sample space* (set of possibilities)
- $\omega \in \Omega$ is a *sample point/possible world/atomic event/outcome of a random process/...*

Probability space (or *probability model*)

- Function $P : \Omega \mapsto \Re$, such that
  - $0 \leq P(\omega) \leq 1$
  - $\sum_{\omega \in \Omega} P(\omega) = 1$

Example: rolling a die

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$P(\omega) = \{1/6, 1/6, 1/6, 1/6, 1/6, 1/6\}$$

Event

An *event* A is any subset of Ω

Probability of an event A is a function assigning to A a value in $[0, 1]$

$$P(A) = \sum_{\{\omega \in A\}} P(\omega)$$

Example 1: A_1 = "die roll < 4", $A_1 = \{1, 2, 3\} \subset \Omega$

$$P(A_1) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2$$

Example 2: A_2 = "die roll $= 4$", $A_2 = \{4\}$, $P(A_2) = 1/6$

Example 3: A_3 = "die roll > 6", $A_3 = \emptyset$, $P(A_3) = 0$

Example 4: A_4 = "die roll ≤ 6", $A_4 = \Omega$, $P(A_4) = 1$

Random variables

A *random variable* (outcome of a random phenomenon) is a function from the sample space Ω to some range (e.g., the reals or Booleans) $X : \Omega \mapsto B$.

Example: $Odd : \Omega \mapsto Boolean$.

X is a variable and a function !

$X = x_i$: the random variable X has the value $x_i \in B$
$X = x_i$ is equivalent to $\{\omega \in \Omega | X(\omega) = x_i\}$

Example: $Odd = true \equiv \{1, 3, 5\}$

Sapienza University of Rome, Italy - Machine Learning (2020/2021)

Random variables

P induces a *probability distribution* for a random variable X:

$$P(X = x_i) = \sum_{\{\omega \in \Omega | X(\omega) = x_i\}} P(\omega)$$

Example
$P(Odd = true) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 1/2$

Propositions

A proposition is the event (subset of Ω) where an assignment to a random variable holds.

- event $a \equiv \; A = true \; \equiv \{\omega \in \Omega | A(\omega) = true\}$

Propositions can be combined using standard logical operators, e.g.:

- event $\neg a \equiv \; A = false \; \equiv \{\omega \in \Omega | A(\omega) = false\}$
- event $a \wedge b$ = points ω where $A(\omega) = true$ and $B(\omega) = true$
- event $\neg a \vee b$ = points ω where $A(\omega) = false$ or $B(\omega) = true$

$$P(\neg a \vee b) = \sum_{\{\omega \in \Omega | A(\omega) = false \vee B(\omega) = true\}} P(\omega)$$

Syntax for propositions

- *Propositional* or *Boolean* random variables
  e.g., *Cavity* (do I have a cavity?).
  $Cavity = true$ is a proposition, also written *cavity*
- *Discrete* random variables (*finite* or *infinite*)
  e.g., *Weather* is one of $< sunny, rain, cloudy, snow >$.
  $Weather = rain$ is a proposition
  Values must be exhaustive and mutually exclusive
- *Continuous* random variables (*bounded* or *unbounded*)
  e.g., $Temp = 21.6$, $Temp < 22.0$.
- Arbitrary Boolean combinations of basic propositions
  e.g., $cavity \wedge Weather = rain \wedge Temp < 22.0$.

Prior Probability

P(h) to denote the initial probability that hypothesis h holds, before we have observed the training data. P(h) is often called the priorprobability of h and may reflect any backgroundknowledge we have about the chance that h is a correct hypothesis. If we have no such prior knowledge, then we might simply assign the same prior probability to each candidate hypothesis.

P(D) to denote the prior probability that training data D will be observed

Prior or *unconditional probabilities* of propositions correspond to belief prior to arrival of any (new) evidence.

Examples:
$P(Odd = true) = 0.5$
$P(Cavity = true) = 0.1$
$P(Weather = sunny) = 0.72$

Probability distribution

A *probability distribution* is a function assigning a probability value to all possible assignments of a random variable.
Note: sum of all values must be 1.

Examples:
$P(Odd) =< 0.5, 0.5 >$
$P(Cavity) =< 0.1, 0.9 >$
$P(Weather) =< 0.72, 0.1, 0.08, 0.1 >$

Note: for real valued random variable X, $P(X)$ is a continuous function.

Joint probability distribution

Joint probability distribution for a set of random variables gives the probability of every atomic joint event on those random variables (i.e., every sample point in the joint sample space).

Joint probability distribution of the random variables *Weather* and *Cavity*:
$P(Weather, Cavity)$ = a 4×2 matrix of values:

$Weather =$	$sunny$	$rain$	$cloudy$	$snow$
$Cavity = true$	0.144	0.02	0.016	0.02
$Cavity = false$	0.576	0.08	0.064	0.08

Every question about a domain can be answered by the joint distribution because every event is a set of sample points

Conditional/Posterior Probability

P(D|h) to denote the probability of observing data D given some world in which hypothesis h holds. More generally, we write P(xly) to denote the probability of x given y. In machine learning problems we are interested in the probability P (h|D) that h holds given the observed training data D. P(h|D) is called the posterior probability of h, because it reflects our confidence that h holds after we have seen the training data D.

N.B: Notice the posterior probability P(h|D) reflects the influence of the training data D, in contrast to the prior probability P(h),which is independent of D.

Belief after the arrival of some evidence.
I know the outcome of a random variable, how does this affect probability of other random variables?

Example:
I know that today $Weather = sunny$, how this information affects the random variable *Cavity*?

Notation:
$P(Cavity = true | Weather = sunny)$: conditional/posterior probability

Conditional/Posterior Probability

In general, conditional/posterior probabilities are different from joint probabilities and from prior probabilities.

$$P(Cavity = true | Weather = sunny) \neq$$
$$P(Cavity = true, Weather = sunny) \neq$$
$$P(Cavity = true)$$

Conditional/Posterior Probability

Consider another Boolean random variable Toothache.
Given that I have a toothache, how this affects the event of having a cavity?

Example:
$P(cavity) = 0.2$: prior
$P(cavity|toothache) = 0.6$: posterior

Conditional Probability Distributions

Conditional probability distributions: representation of all the values of conditional probabilities of random variables.

Example:
$P(Cavity|Toothache) =$ 2-element vector of 2-element vectors

Conditional probability

Definition of conditional probability:

$$P(a|b) \equiv \frac{P(a \wedge b)}{P(b)} \text{ if } P(b) \neq 0$$

Product rule

$$P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$$

A general version holds for whole distributions, e.g.,
$P(Weather, Cavity) = P(Weather|Cavity)P(Cavity)$

Total probabilities

For a Boolean random variable B

$$P(a) = P(a|b)P(b) + P(a|\neg b)P(\neg b)$$

In general, for a random variable Y accepting mutually exclusive values y_i

$$P(X) = \sum_{y_i \in \mathcal{D}(Y)} P(X|Y = y_i)P(Y = y_i)$$

$\mathcal{D}(Y)$: set of values for variable Y

Chain Rule

- *Chain rule* is derived by successive application of product rule:

$$\begin{aligned}
P(X_1, X_2) &= P(X_1)P(X_2|X_1) \\
P(X_1, \ldots, X_n) &= P(X_1, \ldots, X_{n-1})P(X_n|X_1, \ldots, X_{n-1}) \\
&= P(X_1, \ldots, X_{n-2})P(X_{n-1}|X_1, \ldots, X_{n-2})P(X_n|X_1, \ldots, X_{n-1}) \\
&= \ldots \\
&= \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1})
\end{aligned}$$

Inference by Enumeration

Start with the joint distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:
$P(\phi) = \sum_{\omega:\omega \models \phi} P(\omega)$

Inference by Enumeration

Start with the joint distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:
$P(\phi) = \sum_{\omega:\omega \models \phi} P(\omega)$

$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

Inference by Enumeration

Start with the joint distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:
$P(\phi) = \sum_{\omega:\omega\models\phi} P(\omega)$

$P(cavity \lor toothache) = 0.108+0.012+0.072+0.008+0.016+0.064 = 0.28$

Inference by Enumeration

Start with the joint distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

Can also compute conditional probabilities:

$$
\begin{aligned}
P(\neg cavity | toothache) &= \frac{P(\neg cavity \land toothache)}{P(toothache)} \\
&= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4
\end{aligned}
$$

Normalization

Start with the joint distribution:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	$\boxed{.108}$	$\boxed{.012}$	.072	.008
¬ cavity	$\boxed{.016}$	$\boxed{.064}$	.144	.576

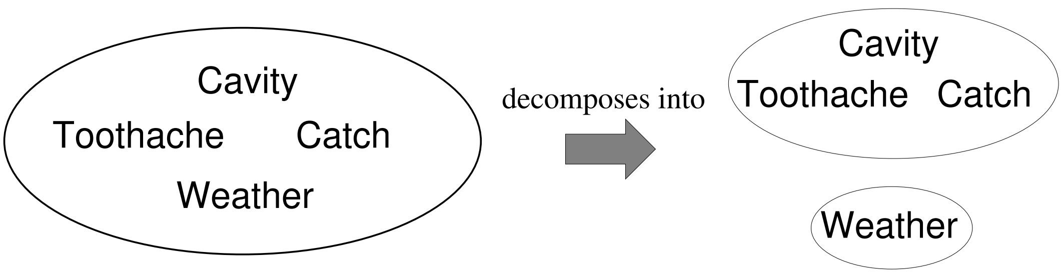
Denominator can be viewed as a *normalization constant* α

$$P(Cavity|toothache) = \alpha\, P(Cavity, toothache)$$
$$= \alpha\, [P(Cavity, toothache, catch) + P(Cavity, toothache, \neg catch)]$$
$$= \alpha\, [\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle]$$
$$= \alpha\, \langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle$$

General idea: compute distribution on query variable
by fixing *evidence variables* and summing over *hidden variables*

Independence

A and B are *independent* iff
$P(A|B) = P(A)$ or $P(B|A) = P(B)$ or $P(A, B) = P(A)P(B)$



$$P(Toothache, Catch, Cavity, Weather) =$$
$$P(Toothache, Catch, Cavity)P(Weather)$$

Independence

$P(Toothache, Catch, Cavity, Weather)$ has 32 entries
$P(Toothache, Catch, Cavity)$ and $P(Weather)$ have $8 + 4 = 12$ entries

Example: n independent biased coins, reduced size from 2^n to n

Absolute independence powerful, but rare.
Complex systems have hundreds of variables, none of which are independent.

Conditional independence

- $P(Toothache, Cavity, Catch)$ has $2^3 - 1 = 7$ independent entries
- If I have a cavity, the probability that the probe catches in it does not depend on whether I have a toothache:
  (1) $P(catch|toothache, cavity) = P(catch|cavity)$
- The same independence holds if I haven't got a cavity:
  (2) $P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$
- *Catch* is *conditionally independent* of *Toothache* given *Cavity*:
  $P(Catch|Toothache, Cavity) = P(Catch|Cavity)$
- Equivalent statements:
  $P(Toothache|Catch, Cavity) = P(Toothache|Cavity)$
  $P(Toothache, Catch|Cavity) = P(Toothache|Cavity)P(Catch|Cavity)$

Thus representation of $P(Toothache, Catch|Cavity)$ is simplified

Conditional independence

General formulation:

X conditionally independent from Y given Z iff $P(X|Y,Z) = P(X|Z)$

$$P(X,Y|Z) = P(X|Y,Z)P(Y|Z) = P(X|Z)P(Y|Z)$$

$$P(Y_1,\ldots,Y_n|Z) = P(Y_1|Y_2,\ldots Y_n,Z)P(Y_2|Y_3\ldots Y_n,Z)\cdots P(Y_n|Z)$$

Y_i conditionally independent from Y_j given Z

$$P(Y_1,\ldots,Y_n|Z) = P(Y_1|Z)P(Y_2|Z)\cdots P(Y_n|Z)$$

Sapienza University of Rome, Italy - Machine Learning (2020/2021)

Conditional independence

Chain rule + Conditional independence

$$\begin{aligned} P(X,Y,Z) &= P(X|Y,Z)P(Y,Z) = P(X|Y,Z)P(Y|Z)P(Z) \\ &= P(X|Z)\,P(Y|Z)\,P(Z) \end{aligned}$$

$P(Toothache, Catch, Cavity)$
$= P(Toothache|Catch, Cavity)P(Catch, Cavity)$
$= P(Toothache|Catch, Cavity)P(Catch|Cavity)P(Cavity)$
$= P(Toothache|Cavity)P(Catch|Cavity)P(Cavity)$
$2 + 2 + 1 = 5$ independent numbers (instead of $2^3 - 1$)

In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n.

Bayes' Rule

- Product rule $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$$\Rightarrow \textit{Bayes' rule } P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

or in distribution form

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \alpha P(X|Y)P(Y)$$

Useful for assessing *diagnostic* probability from *causal* probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

Bayes' Rule and conditional independence

Bayes rule

$$P(Z|Y_1, \ldots, Y_n) = \alpha P(Y_1, \ldots, Y_n|Z)\, P(Z)$$

$Y_i, \ldots Y_n$ conditionally independent each other given Z

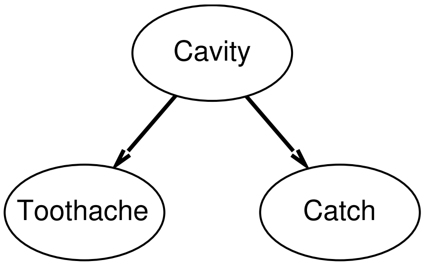
$$P(Z|Y_1, \ldots, Y_n) = \alpha P(Y_1|Z) \cdots P(Y_n|Z)\, P(Z)$$

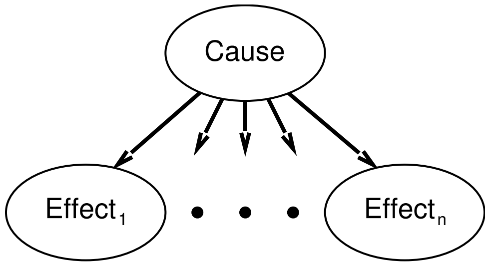
Effects conditionally independent each other given a cause.

$$P(Cause|Effect_1, \ldots, Effect_n) = \alpha P(Cause) \prod_i P(Effect_i|Cause)$$

Total number of parameters is *linear* in n

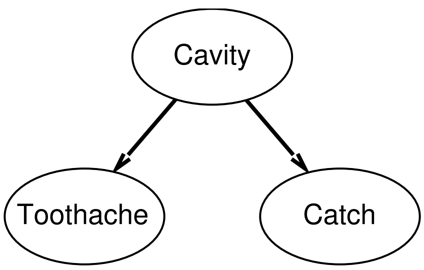
Bayesian networks

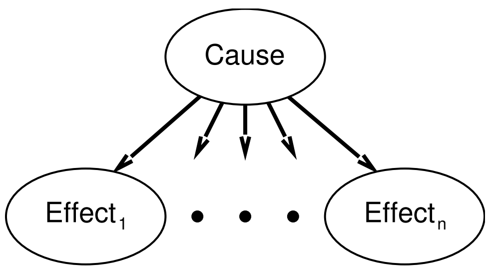




A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions.

Bayesian networks





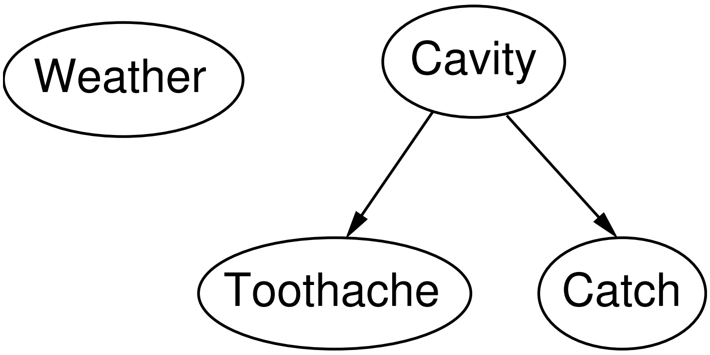
Syntax:

- a set of nodes, one per variable
- a directed, acyclic graph (link ≈ "directly influences")
- a conditional distribution for each node given its parents: $P(X_i|\mathrm{Parents}(X_i))$

In the simplest case, conditional distribution represented as a *conditional probability table* (CPT) giving the distribution over X_i for each combination of parent values.

Dentist BN Example

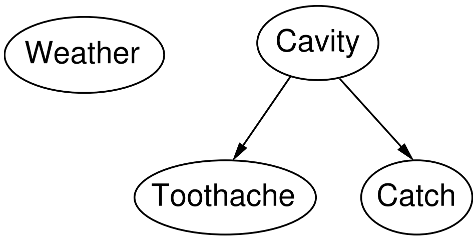
Topology of network encodes conditional independence assertions:



Weather is independent of the other variables
Toothache and *Catch* are conditionally independent given *Cavity*

Dentist BN Example

BN model given by the set of CPT $P(X_i|\mathrm{Parents}(X_i))$ for each variable X_i



$P(Weather)$
$P(Cavity)$
$P(Toothache|Cavity)$
$P(Catch|Cavity)$

All the joint probabilities can be computed from this model.
How many independent values?

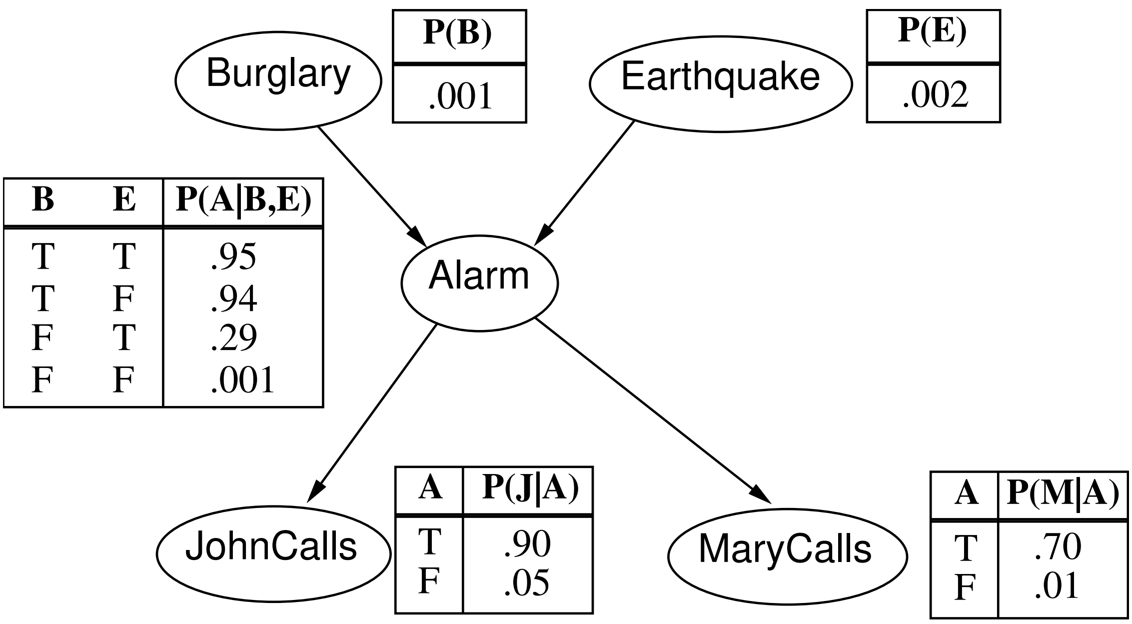
Burglar BN Example

I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes the alarm is set off by minor earthquakes. Is there a burglar?

Variables: *Burglar*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
Network topology reflects "causal" knowledge:

- A burglar can set the alarm
- An earthquake can set the alarm
- The alarm can cause Mary to call
- The alarm can cause John to call

Burglar BN Example



P(B)
.001

P(E)
.002

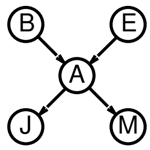
B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Compactness

A CPT for Boolean variable X_i with k Boolean parents has 2^k rows for the combinations of parent values



Each row requires one number p for $X_i = true$ (the number for $X_i = false$ is just $1 - p$)

If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers

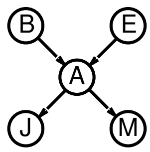
I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution

For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)

Computing joint probabilities

All joint probabilities computed with the chain rule:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | \mathrm{Parents}(X_i))$$



e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$$\begin{aligned} &= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \\ &= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ &\approx 0.00063 \end{aligned}$$