Artificial Intelligence

academic year 2025/2026

Giorgio Fumera

Pattern Recognition and Applications Lab
Department of Electrical and Electronic Engineering
University of Cagliari (Italy)

Knowledge Representation and Inference under Uncertainty: Bayesian Networks

Suggested textbooks

This course requires a basic knowledge of probability theory

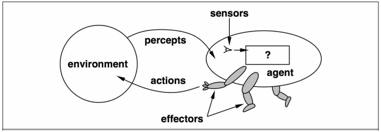
Bayesian neworks are covered by the course textbook, which provides also an informal introduction to probability theory:
S. Russell, P. Norvig, *Artificial Intelligence – A Modern Approach*, 4th Ed., Pearson, 2021 (or a previous edition)

For a formal and comprehensive introduction to probability theory and statistics, students are referred to textbooks such as:
A.M. Mood, F.A. Graybill, D.C. Boes, *Introduction to the Theory of Statistics*, McGraw-Hill, 1991 / 1998

Introduction

Rational agents: acting under uncertainty

Often rational agents must make decisions and act under **uncertainty** about their environment, e.g.:

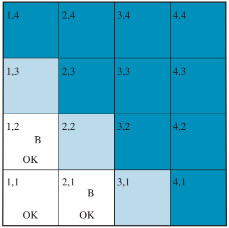


- medical diagnosis from patients' symptoms and outcomes of medical tests
- speech/image recognition from **noisy** audio/video signals
- self-driving vehicles: **incomplete** and **noisy** sensory data

Limitations of logical agents

Logic considers propositions that can only be either true, false, or unknown, but cannot associate them a degree of belief

Example: a logical agent's possible knowledge about pits in the wumpus world, after visiting squares $(1,1),(1,2)$ and $(2,1)$: the truth value of $P_{1,3}$, $P_{2,2}$ and $P_{3,1}$ is **unknown** (sensors report only **local** information)



The next move can only be decided **randomly**

Limitations of logical agents

Diagnosis problems (e.g., in medicine): a good example of the limitations of the logical approach in dealing with uncertainty

Example: the following dental diagnosis rule (where p denotes a patient) is wrong, since toothache can have other causes besides cavity:

$$\forall p \; Symptom(p, Toothache) \Rightarrow Disease(p, Cavity)$$

Trying to specify **all** possible causes is pointless, e.g.:

$$\begin{aligned} &\forall p \; Symptom(p, Toothache) \Rightarrow \\ &\quad Disease(p, Cavity) \lor Disease(p, Abscess) \lor \ldots \end{aligned}$$

Turning the first rule into a **causal** one does not work, either, since not all cavities cause toothache:

$$\forall p \; Disease(p, Cavity) \Rightarrow Symptom(p, Toothache)$$

Making rational decisions under uncertainty

Uncertainty can be due to **laziness**, or by **theoretical** or **practical ignorance**

In domains involving uncertainty (medicine, business, law, etc.), only a degree of belief in sentences of interest can be provided

Probability theory is a widely used tool to summarise this kind of uncertainty into a numerical value, conventionally in the range $[0, 1]$

For instance, stating that the probability that a patient with a toothache has a cavity is 0.8 means that

- in all possible situations indistinguishable by the agent from the current one, 80% of patients suffering from toothache have a cavity
- the missing 20% summarises **all** the other possible causes of toothache the agent is too lazy or ignorant to confirm or deny

Making rational decisions under uncertainty

Effective decision-making agents under uncertainty should have preferences about the possible **outcomes** of their actions

Representing and reasoning with preferences is the subject of utility theory

Rational decision-making under uncertainty can be achieved by **combining** the likelihood (probability) and the preference about actions' outcomes, which in turn is the subject of decision theory:

$$\textit{decision theory} = \textit{probability theory} + \textit{utility theory}$$

Basically, a **rational agent** chooses the actions that yields the highest **expected** utility, averaged over all its possible outcomes

Probabilistic modelling and inference

Random variables

Random variables are the basic element of a language for describing the "state of the world" of interest to a probabilistic agent

- notation: symbols with uppercase initial, e.g.: M, X, $\mathit{Weather}$, $\mathit{Toothache}$
- domain (set of **mutually exclusive** values)
  - Boolean, e.g.: $\mathit{Toothache} \in \{\mathrm{true},\ \mathrm{false}\}$
  - discrete, e.g.: $\mathit{Weather} \in \{\mathrm{sunny},\ \mathrm{rainy},\ \mathrm{cloudy},\ \mathrm{snow}\}$, $M \in \mathbb{N}$
  - continuous, e.g.: $X \in \mathbb{R}$

For convenience, in the following the domain of a discrete random variable will be represented as an ordered **tuple** instead of a set, e.g.: $\langle\mathrm{true},\ \mathrm{false}\rangle$, $\langle\mathrm{sunny}, \mathrm{rain}, \mathrm{cloudy}, \mathrm{snow}\rangle$

Probability distribution function (PDF)

Example: $\mathit{Weather} \in \langle \mathrm{sunny}, \mathrm{rain}, \mathrm{cloudy}, \mathrm{snow} \rangle$

$$\begin{aligned} P(\mathit{Weather} &= \mathrm{sunny}) = 0.7 \\ P(\mathit{Weather} &= \mathrm{rain}) = 0.2 \\ P(\mathit{Weather} &= \mathrm{cloudy}) = 0.08 \\ P(\mathit{Weather} &= \mathrm{snow}) = 0.02 \end{aligned}$$

Vector notation: $\mathbf{P}(\mathit{Wheater}) = \langle 0.7, 0.2, 0.08, 0.02 \rangle$

Axioms and theorems of probability theory:

- $0 \leq P(\cdot) \leq 1$ (axiom)
- $\sum_{x \in \mathcal{D}(X)} P(X = x) = 1$ (theorem)

(Full) joint PDF

Example: Boolean random variables describing a dentist's patient

- *Toothache* (the patient has a toothache)
- *Cavity* (the patient has a cavity)
- *Catch* (the dentist's steel probe catches in a tooth)

(Full) joint PDF $\mathbf{P}(\mathit{Toothache}, \mathit{Cavity}, \mathit{Catch})$:

	$\mathit{Toothache} = t$		$\mathit{Toothache} = f$	
	$\mathit{Catch} = t$	$\mathit{Catch} = f$	$\mathit{Catch} = t$	$\mathit{Catch} = f$
$\mathit{Cavity} = t$	0.108	0.012	0.072	0.008
$\mathit{Cavity} = f$	0.016	0.064	0.144	0.576

Theorem:

$$\sum_{a,b,c \in \{\mathrm{true},\mathrm{false}\}} P(\mathit{Toothache} = a, \mathit{Cavity} = b, \mathit{Catch} = c) = 1$$

Events

Event: any combination of values of (a subset of) the random variables, e.g.:

$$\mathit{Cavity} = \mathit{true} \;\vee\; \mathit{Toothache} = \mathit{false}$$

Atomic event: any combination of values of **all** the random variables (a **complete** description of the “state of the world”), e.g.:

$$\mathit{Cavity} = \mathrm{true} \;\wedge\; \mathit{Toothache} = \mathrm{false} \;\wedge \mathit{Catch} = \mathrm{true}$$

Atomic events are **mutually esclusive** and **exhaustive**

Any event is a **disjunction** of atomic events, e.g.:

$$\begin{aligned}&\mathit{Cavity} = \mathrm{true} \;\vee\; \mathit{Toothache} = \mathrm{false} \equiv \\ &(\mathit{Cavity} = \mathrm{true} \;\wedge\; \mathit{Toothache} = \mathrm{true} \;\wedge \mathit{Catch} = \mathrm{true}) \vee \ldots\end{aligned}$$

Probabilistic inference

Computing the probability of an event of interest

A dentist may be interested in

- $P(\mathit{Cavity} = \mathrm{true})$
- $P(\mathit{Cavity} = \mathrm{true} \wedge \mathit{Toothache} = \mathrm{true})$

Marginal probability: joint probability of any **subset** of random variables

Probabilistic inference: marginalisation (sum rule)

(Full) joint PDF $\mathbf{P}(\mathit{Toothache}, \mathit{Cavity}, \mathit{Catch})$:

	$\mathit{Toothache} = t$		$\mathit{Toothache} = f$	
	$\mathit{Catch} = t$	$\mathit{Catch} = f$	$\mathit{Catch} = t$	$\mathit{Catch} = f$
$\mathit{Cavity} = t$	0.108	0.012	0.072	0.008
$\mathit{Cavity} = f$	0.016	0.064	0.144	0.576

Marginal PDF $\mathbf{P}(\mathit{Cavity}, \mathit{Toothache})$:

		Tootache	
		true	false
Cavity	true	0.120	0.080
	false	0.080	0.720

Marginalisation, or sum rule ($\mathcal{X}_k$ denotes the domain of X_k):

$$\mathbf{P}(X_1, \ldots, X_p) = \sum_{x_{p+1} \in \mathcal{X}_{p+1}, \ldots, x_n \in \mathcal{X}_n} \mathbf{P}(X_1, \ldots, X_p, X_{p+1} = x_{p+1}, \ldots, X_n = x_n)$$

Prior and posterior (conditional) probability

A dentist may be also interested in posterior (conditional) probabilities

- $P(\mathit{Cavity} = \mathrm{true} | \mathit{Toothache} = \mathrm{true}) \triangleq \frac{P(\mathit{Cavity}=\mathrm{true}, \mathit{Toothache}=\mathrm{true})}{P(\mathit{Toothache}=\mathrm{true})}$
- $P(\mathit{Cavity} = \mathrm{true} | \mathit{Catch} = \mathrm{true}) \triangleq \frac{P(\mathit{Cavity}=\mathrm{true}, \mathit{Catch}=\mathrm{true})}{P(\mathit{Catch}=\mathrm{true})}$
- $P(\mathit{Cavity} = \mathrm{true} | \mathit{Toothache} = \mathrm{true}, \mathit{Catch} = \mathrm{false}) \triangleq$ $\frac{P(\mathit{Cavity}=\mathrm{true}, \mathit{Toothache}=\mathrm{true}, \mathit{Catch}=\mathrm{false})}{P(\mathit{Toothache}=\mathrm{true}, \mathit{Catch}=\mathrm{false})}$
- $P(\mathit{Cavity} = \mathrm{true}, \mathit{Catch} = \mathrm{false} | \mathit{Toothache} = \mathrm{true}) \triangleq$ $\frac{P(\mathit{Cavity}=\mathrm{true}, \mathit{Catch}=\mathrm{false}, \mathit{Toothache}=\mathrm{true})}{P(\mathit{Toothache}=\mathrm{true})}$

Conditional PDF

Conditional probability: $P(\mathit{Cavity} = \mathrm{true} | \mathit{Toothache} = \mathrm{true}) = 0.8$

Conditional PDF: $\mathbf{P}(\mathit{Cavity} | \mathit{Toothache} = \mathrm{true}) = \langle 0.8, 0.2 \rangle$

Also the values of a conditional PDF sum to 1

Every value of the conditioning event corresponds to a **distinct** conditional PDF, e.g.:

- $\mathbf{P}(\mathit{Cavity} | \mathit{Toothache} = \mathrm{true}) = \langle 0.8, 0.2 \rangle$
- $\mathbf{P}(\mathit{Cavity} | \mathit{Toothache} = \mathrm{false}) = \langle 0.05, 0.95 \rangle$

Probabilistic inference

Also conditional probabilities can be computed from the full joint PDF

$$
\begin{aligned}
&P(\mathit{Cavity} = \mathrm{true} \mid \mathit{Toothache} = \mathrm{true}) \\
&= \frac{P(\mathit{Cavity} = \mathrm{true}, \mathit{Toothache} = \mathrm{true})}{P(\mathit{Toothache} = \mathrm{true})} \quad \text{(by definition)}
\end{aligned}
$$

Full joint PDF:

	$\mathit{Toothache} = \mathrm{t}$		$\mathit{Toothache} = \mathrm{t}$	
	$\mathit{Catch} = \mathrm{t}$	$\mathit{Catch} = \mathrm{f}$	$\mathit{Catch} = \mathrm{t}$	$\mathit{Catch} = \mathrm{f}$
$\mathit{Cavity} = \mathrm{t}$	0.108	0.012	0.072	0.008
$\mathit{Cavity} = \mathrm{f}$	0.016	0.064	0.144	0.576

Sum rule:

$$
\begin{aligned}
&P(\mathit{Cavity} = \mathrm{t}, \mathit{Toothache} = \mathrm{t}) = 0.108 + 0.012 = 0.120 \\
&P(\mathit{Toothache} = \mathrm{t}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.200
\end{aligned}
$$

Probabilistic inference: summary

If the full joint PDF of random variables $X_1, \ldots, X_n$ is **known**

- computing a marginal (prior) PDF: sum rule

$$\mathbf{P}(X_1, \ldots, X_p) = \sum_{x_{p+1} \in \mathcal{X}_{p+1}, \ldots, x_n \in \mathcal{X}_n} \mathbf{P}(X_1, \ldots, X_p, X_{p+1} = x_{p+1}, \ldots, X_n = x_n)$$

- computing a conditional (posterior) PDF: definition + sum rule

$$\mathbf{P}(X_1, \ldots, X_p | X_{p+1}, \ldots, X_q) = \frac{\mathbf{P}(X_1, \ldots, X_p, X_{p+1}, \ldots, X_q)}{\mathbf{P}(X_{p+1}, \ldots, X_q)} = \ldots$$

Probabilistic inference using the full joint PDF: issues

How to compute or estimate the full joint PDF?

- $Coin \in \langle \text{heads}, \text{tails} \rangle$: $\mathbf{P}(Coin) = \langle ?, ? \rangle$
- $Dice \in \langle$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\rangle$: $\mathbf{P}(Dice) = \langle ?, ?, ?, ?, ?, ? \rangle$

Classical (a priori) probability: mutually exclusive, equally likely, random atomic events

- $Coin \in \langle \text{heads}, \text{tails} \rangle$: $\mathbf{P}(Coin) = \langle \frac{1}{2}, \frac{1}{2} \rangle$
- $Dice \in \langle$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\rangle$: $\mathbf{P}(Dice) = \langle \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6} \rangle$

Same approach for the probability of **any** related event, e.g.:

- getting an even face up (⚁ or ⚃ or ⚅)
- getting a sum of the faces up equal to 6 after throwing two dice (⚀, ⚄ or ⚁, ⚃, . . .)

Probabilistic inference using the full joint PDF: issues

How to compute or estimate the full joint PDF?

	$\mathit{Toothache} = \mathrm{t}$		$\mathit{Toothache} = \mathrm{f}$	
	$\mathit{Catch} = \mathrm{t}$	$\mathit{Catch} = \mathrm{f}$	$\mathit{Catch} = \mathrm{t}$	$\mathit{Catch} = \mathrm{f}$
$\mathit{Cavity} = \mathrm{t}$	?	?	?	?
$\mathit{Cavity} = \mathrm{f}$	?	?	?	?

Frequentist (a posteriori) probability: when events of interest can be observed under **similar** and **uniform** conditions

Examples

- getting ⚅ after throwing a **loaded** dice
- a chip manufactured by company *XYZ* is faulty
- a student in CECAI graduates within 2 years

Probabilistic inference using the full joint PDF: issues

How to compute or estimate the full joint PDF?

- the piano player John Smith will break one or both of his hands within the next 10 years
- the third World War will start within 2026

Subjective probability (e.g., domain experts' judgement)

Probabilistic inference using the full joint PDF: issues

Effort required to estimate the full joint PDF

Example: full joint PDF of n Boolean random variables

	$\mathit{Toothache} = \mathrm{t}$		$\mathit{Toothache} = \mathrm{f}$	
	$\mathit{Catch} = \mathrm{t}$	$\mathit{Catch} = \mathrm{f}$	$\mathit{Catch} = \mathrm{t}$	$\mathit{Catch} = \mathrm{f}$
$\mathit{Cavity} = \mathrm{t}$	?	?	?	?
$\mathit{Cavity} = \mathrm{f}$	?	?	?	?

$2^n - 1$ probability values must be specified; many of them may be not easy to estimate, too

Probabilistic inference using the full joint PDF: issues

Computational complexity of inference

Example: inference over n Boolean random variables:

$$\mathbf{P}(X_1, \ldots, X_p) = \text{(sum rule)}$$
$$\sum_{x_{p+1} \in \mathcal{X}_{p+1}, \ldots, x_n \in \mathcal{X}_n} \mathbf{P}(X_1, \ldots, X_p, X_{p+1} = x_{p+1}, \ldots, X_n = x_n)$$

sum of 2^n probability values (2^{n-p} values for **each** of the 2^p values of $X_1, \ldots, X_p$)

Simplifying probabilistic modelling and inference

A useful tool: the product rule

$$\underbrace{\mathbf{P}(X|Y) \triangleq \frac{\mathbf{P}(X,Y)}{\mathbf{P}(Y)}}_{\text{definition of conditional probability}} \quad \Rightarrow \quad \underbrace{\mathbf{P}(X,Y) = \mathbf{P}(X|Y)\mathbf{P}(Y)}_{\text{product rule}}$$

Extension to **groups** of random variables, e.g.:

$$\mathbf{P}(X_1,\ldots,X_q) = \mathbf{P}(X_1,\ldots,X_p|X_{p+1},\ldots,X_q)\mathbf{P}(X_{p+1},\ldots,X_q)$$

Extension to **conditional** probabilities, e.g.:

$$\mathbf{P}(X_1,X_2|X_3) = \mathbf{P}(X_1|X_2,X_3)\mathbf{P}(X_2|X_3)$$

Simplifying probabilistic modelling and inference

Besides the Boolean variables *Toothache*, *Catch* and *Cavity*, our dentist would like to use $Weather \in \langle \mathrm{sunny}, \mathrm{rain}, \mathrm{cloudy}, \mathrm{snow} \rangle$

How many probability values must be estimated to specify the full joint PDF $\mathbf{P}(Toothache, Catch, Cavity, Weather)$?
$(2 \times 2 \times 2 \times 4) - 1 = 31 \ldots$

However, it is reasonable to assume that there is no cause–effect relation between weather and dental problems. . .

Simplifying probabilistic modelling and inference

Using the product rule:

$$\mathbf{P}(Toothache, Catch, Cavity, Weather) =$$
$$\mathbf{P}(Toothache, Catch, Cavity|Weather)\mathbf{P}(Weather)$$

Weather can be **assumed** to be (statistically) independent on *Toothache*, *Catch* and *Cavity*:

$$\mathbf{P}(Toothache, Catch, Cavity|Weather) = \mathbf{P}(Toothache, Catch, Cavity)$$

It follows that:

$$\mathbf{P}(Toothache, Catch, Cavity, Weather) =$$
$$\mathbf{P}(Toothache, Catch, Cavity)\mathbf{P}(Weather)$$

How many probability values must be estimated to specify $\mathbf{P}(Toothache, Catch, Cavity)$ and $\mathbf{P}(Weather)$? $(2^3 - 1) + (4 - 1) = 10$!

Independence: formal definition

Two variables X and Y are independent, if and only if

$$\mathbf{P}(X|Y) = \mathbf{P}(X)$$

Equivalent conditions (notice the symmetry):

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

Extension to **groups** of random variables, e.g. (**equivalent** conditions):

$$\begin{aligned} \mathbf{P}(X_1, \dots, X_p|X_{p+1}, \dots, X_q) &= \mathbf{P}(X_1, \dots, X_p) \\ \mathbf{P}(X_{p+1}, \dots, X_q|X_1, \dots, X_p) &= \mathbf{P}(X_{p+1}, \dots, X_q) \\ \mathbf{P}(X_1, \dots, X_p, X_{p+1}, \dots, X_q) &= \mathbf{P}(X_1, \dots, X_p)\mathbf{P}(X_1, \dots, X_p) \end{aligned}$$

Why is independence useful?

Example: n Boolean variables

Full joint PDF $\mathbf{P}(X_1, \ldots, X_n)$: specified by $2^n - 1$ probability values...

...if the variables were **all** independent:

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i)$$

only $n \times (2^1 - 1) = n$ probability values are required!

Unfortunately, absolute independence is rare in practice...

Simplifying probabilistic modelling and inference

A weaker form of independence can be exploited, considering the cause–effect relation between variables

Toothache and *Catch* are possible **effects** of *Cavity*: using the product rule, the joint PDF can be rewritten into a causal form $P(effects|cause)$:

$$\mathbf{P}(Toothache, Catch, Cavity) = \underbrace{\mathbf{P}(Toothache, Catch|Cavity)}_{P(effects|cause)}\mathbf{P}(Cavity)$$

What can be said about $\mathbf{P}(Toothache, Catch|Cavity)$?

Simplifying probabilistic modelling and inference

Note first that *Toothache* and *Catch* cannot be considered absolutely independent, i.e.:

$$\mathbf{P}(\mathit{Toothache} | \mathit{Catch}) \neq \mathbf{P}(\mathit{Toothache})$$

Example: patients with a “hollow” in a tooth are more likely to have toothache than patients without any “hollow”, thus:

$$\mathbf{P}(\mathit{Toothache} | \mathit{Catch} = \mathrm{true}) \neq \mathbf{P}(\mathit{Toothache} | \mathit{Catch} = \mathrm{false})$$

However, *Toothache* and *Catch* can be **assumed** to be conditionally independent **given** *Cavity*:

$$\mathbf{P}(\mathit{Toothache} | \mathit{Catch}, \mathit{Cavity}) = \mathbf{P}(\mathit{Toothache} | \mathit{Cavity}) \ ,$$

since, e.g., among patients with a cavity, the fraction with toothache can be considered constant, regardless of the presence of absence of “hollows” in some tooth

Simplifying probabilistic modelling and inference

Note that the expression:

$$\mathbf{P}(Toothache|Catch, Cavity) = \mathbf{P}(Toothache|Cavity)$$

is **equivalent** to:

$$\begin{aligned}\mathbf{P}(Catch|Toothache, Cavity) &= \mathbf{P}(Catch|Cavity)\\ \mathbf{P}(Toothache, Catch|Cavity) &= \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)\end{aligned}$$

Conditional independence: formal definition

Two variables X and Y are conditionally independent **given** another variable Z, if and only if

$$\mathbf{P}(X|Y,Z) = \mathbf{P}(X|Z)$$

Equivalent conditions:

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

Extension to **groups** of random variables (**equivalent** conditions):

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

Usefulness of conditional independence

From the above conditional independence assumption it follows that:

$$\mathbf{P}(\mathit{Toothache}, \mathit{Catch}, \mathit{Cavity}) = \text{ (product rule)}$$
$$\mathbf{P}(\mathit{Toothache}, \mathit{Catch} | \mathit{Cavity})\mathbf{P}(\mathit{Cavity}) = \text{ (conditional independence)}$$
$$\mathbf{P}(\mathit{Toothache} | \mathit{Cavity})\mathbf{P}(\mathit{Catch} | \mathit{Cavity})\mathbf{P}(\mathit{Cavity})$$

To specify the full joint distribution, $2^3 - 1 = 7$ probability values must be estimated...

To specify the other distributions one needs to estimate

- $\mathbf{P}(\mathit{Toothache} | \mathit{Cavity})$: $2 \times (2^1 - 1) = 2$ values
- $\mathbf{P}(\mathit{Catch} | \mathit{Cavity})$: $2 \times (2^1 - 1) = 2$ values
- $\mathbf{P}(\mathit{Cavity})$: $2^1 - 1 = 1$ value

for a total of 5 values – a **negligible** gain?

Usefulness of conditional independence

Example: n Boolean variables

The conditional PDF

$$\mathbf{P}(X_1, \ldots, X_{n-1} | X_n)$$

is specified by $2 \times (2^{n-1} - 1) = 2^n - 2$ probability values...

...if $X_1, \ldots, X_{n-1}$ were conditionally independent given X_n:

$$\mathbf{P}(X_1, \ldots, X_n | Y) = \prod_{i=1}^{n-1} \mathbf{P}(X_i | X_n) \ ,$$

only $(n-1) \times 2 \times (2^1 - 1) = 2(n-1)$ probability values are required!

Simplifying probabilistic modelling and inference

Potential advantage of causal inference in the form $P(effects|cause)$: distinct effects of a common cause may be conditionally independent, **given** the cause, e.g.:

$$\mathbf{P}(\underbrace{Toothache, Catch}_{\text{effects}} \mid \underbrace{Cavity}_{\text{cause}}) = \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)$$

In practice, diagnostic inference $P(cause|effects)$ is often required, e.g.:

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

Unfortunately, causal knowledge is usually easier to obtain than diagnostic. . .

Diagnostic inference by means of causal inference

A useful tool: Bayes' rule

Two **equivalent** expressions of the product rule:

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

it follows that:

$$\underbrace{\mathbf{P}(Y|X) = \frac{\mathbf{P}(X|Y)\mathbf{P}(Y)}{\mathbf{P}(X)}}_{\mathrm{Bayes'\ rule}}$$

Also the denominator $\mathbf{P}(X)$ can be rewritten as a function of $\mathbf{P}(Y|X)$, through the sum and product rules:

$$\mathbf{P}(X) = \sum_{y\in\mathcal{Y}} \mathbf{P}(X, Y = y) = \sum_{y\in\mathcal{Y}} \mathbf{P}(X|Y = y)P(Y = y)$$

Diagnostic inference by means of causal inference

An example of the application of Bayes' rule:

$$\mathbf{P}(\mathit{Cavity}|\mathit{Toothache}, \mathit{Catch})$$

$$= \frac{\mathbf{P}(\mathit{Toothache}, \mathit{Catch}|\mathit{Cavity})\ \mathbf{P}(\mathit{Cavity})}{\mathbf{P}(\mathit{Toothache}, \mathit{Catch})}$$

$$= \frac{\mathbf{P}(\mathit{Toothache}, \mathit{Catch}|\mathit{Cavity})\ \mathbf{P}(\mathit{Cavity})}{\sum_{c \in \{\mathrm{t,f}\}} \mathbf{P}(\mathit{Toothache}, \mathit{Catch}|\mathit{Cavity} = c) P(\mathit{Cavity} = c)}$$

using the conditional independence assumption previously made:

$$= \frac{\mathbf{P}(\mathit{Toothache}|\mathit{Cavity})\mathbf{P}(\mathit{Catch}|\mathit{Cavity})\ \mathbf{P}(\mathit{Cavity})}{\sum_{c \in \{\mathrm{t,f}\}} \mathbf{P}(\mathit{Toothache}|\mathit{Cavity} = c)\mathbf{P}(\mathit{Catch}|\mathit{Cavity} = c) P(\mathit{Cavity} = c)}$$

Probabilistic modelling and inference: summary

- Full joint PDF + definition of conditional probability + sum rule allow to compute **any** probability... but the effort for estimating the full joint PDF and the **computational complexity** of inference using the sum rule are too high
- Absolute independence (rare) and conditional independence (common) reduce effort and computational complexity
- Conditional independence can be exploited by considering causal knowledge $P(effects|cause)$ (cause–effect relations)
- Diagnostic inference can be carried out from causal knowledge by means of Bayes' rule

The *chain rule* to represent full joint PDFs

Estimating the full joint PDF is usually easier if it is expressed in causal form, $P(\mathit{effects} | \mathit{cause})$, since this often allows to exploit conditional independence assumptions.

To this aim, a **general** strategy is to first sort the variables from the "root causes" to the "end effects", e.g.:

$$\text{end effect}\,\underleftarrow{\mathit{Catch},\ \mathit{Toothache},\ \mathit{Cavity}}\,\text{root cause}$$

Note that

- cavity is a (possible) **direct** cause of both toothache and of "hollows" in a tooth, thus it is the "root cause"
- toothache and "hollows" in a tooth do not **directly** influence each other, instead (i.e., there is no **direct** cause–effect relation between them), thus they can be listed in any order

The *chain rule* to represent full joint PDFs

The full joint PDF can be rewritten using the product rule by conditioning on all the variables except for the **first** one (in the considered order):

$$\mathbf{P}(\mathit{Catch}, \mathit{Toothache}, \mathit{Cavity}) = \\ \mathbf{P}(\mathit{Catch} | \mathit{Toothache}, \mathit{Cavity})\mathbf{P}(\mathit{Toothache}, \mathit{Cavity})$$

The product rule can be applied again in the same way to the marginal PDF $\mathbf{P}(\mathit{Toothache}, \mathit{Cavity})$, leading to:

$$\mathbf{P}(\mathit{Catch}, \mathit{Toothache}, \mathit{Cavity}) = \\ \mathbf{P}(\mathit{Catch} | \mathit{Toothache}, \mathit{Cavity})\mathbf{P}(\mathit{Toothache} | \mathit{Cavity})\mathbf{P}(\mathit{Cavity})$$

The *chain rule* to represent full joint PDFs

The last expression of the full joint PDF:

$$\mathbf{P}(Catch|Toothache, Cavity)\mathbf{P}(Toothache|Cavity)\mathbf{P}(Cavity)$$

is the product of two **conditional** distributions of a **single** variable, conditioned on **all** the variables that **follow** it in the chosen order, and of the **prior** distribution of the **last** variable (the "root cause").

By construction, all the conditional distributions are in the causal form $P(effect|causes)$. This allows the previous conditional independence assumption on $\mathbf{P}(Catch|Toothache, Cavity)$ to be exploited, leading to:

$$\mathbf{P}(Catch, Toothache, Cavity) =$$
$$\mathbf{P}(Catch|Cavity)\mathbf{P}(Toothache|Cavity)\mathbf{P}(Cavity)$$

The *chain rule* to represent full joint PDFs

The previous one is an example of the chain rule, a specific way of rewriting the full joint PDF by means of **repeated** applications of the product rule.

In the general case of n variables $X_1, \ldots, X_n$:

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

In a conditional PDF $\mathbf{P}(X_k | X_{k-1}, \ldots, X_1)$ the conditioning variables $X_{k-1}, \ldots, X_1$ are called the **parents** of X_k; denoting them by $pa(X_k)$:

$$\mathbf{P}(X_1, X_2, \ldots, X_n) = \prod_{k=1}^{n} \mathbf{P}(X_k | pa(X_k))$$

The *chain rule* to represent full joint PDFs

Note that for n variables there are $n!$ distinct but **equivalent** ways to represent their joint PDF using the chain rule (one for each of the possible $n!$ ways of sorting them), e.g.:

$$\begin{aligned}
\mathbf{P}(X_1, X_2, X_3) &= \mathbf{P}(X_1|X_2, X_3)\mathbf{P}(X_2|X_3)\mathbf{P}(X_3) \\
&= \mathbf{P}(X_1|X_3, X_2)\mathbf{P}(X_3|X_2)\mathbf{P}(X_2) \\
&= \mathbf{P}(X_2|X_1, X_3)\mathbf{P}(X_1|X_3)\mathbf{P}(X_3) \\
&= \mathbf{P}(X_2|X_3, X_1)\mathbf{P}(X_3|X_1)\mathbf{P}(X_1) \\
&= \mathbf{P}(X_3|X_1, X_2)\mathbf{P}(X_1|X_2)\mathbf{P}(X_2) \\
&= \mathbf{P}(X_3|X_2, X_1)\mathbf{P}(X_2|X_1)\mathbf{P}(X_1)
\end{aligned}$$

The **most convenient** representaiton is the one that corresponds to one of the possible orders from “root causes” to “end effects.”

Bayesian networks

Probabilistic graphical models

Probabilistic graphical models are a framework for graphical representation of **structured** PDFs

Main classes of graphical models

- Bayesian networks (BNa): **directed acyclic** graphs (DAGs)
- Markov random fields: **undirected** graphs

This course focuses on BNs, that are useful for

- representing full joint PDFs with causal dependencies and conditional independence relations
- developing efficient probabilistic inference algorithms

Bayesian network structure

The expression of the joint PDF obtained using the chain rule is represented by a BN where

- each node represents one random variable, and is associated with its distribution in the expression of the chain rule
- oriented edges represent conditional dependencies, linking each variable (node) with the ones on which its distribution is conditioned

To represent the joint PDF

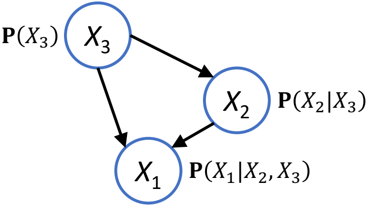
$$\begin{aligned} \mathbf{P}(X_1, X_2, \ldots, X_n) &= \prod_{k=1}^{n} \mathbf{P}(X_k | X_{k-1}, \ldots, X_1) \\ &= \prod_{k=1}^{n} \mathbf{P}(X_k | pa(X_k)) \end{aligned}$$

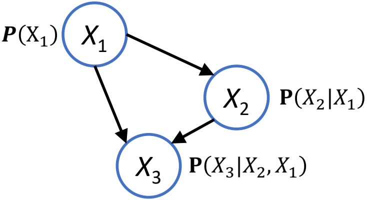
- nodes are drawn top-down, from the “root cause” X_1 to the “end effect” X_n
- for each pair (X_p, X_q), an edge $X_p \rightarrow X_q$ is drawn, if $X_p \in pa(X_q)$

Bayesian network structure

Example: two possible, **equivalent** expressions of $\mathbf{P}(X_1, X_2, X_3)$, and their representation through a BN

$$\mathbf{P}(X_1|X_2,X_3)\mathbf{P}(X_2|X_3)\mathbf{P}(X_3) \qquad \mathbf{P}(X_3|X_2,X_1)\mathbf{P}(X_2|X_1)\mathbf{P}(X_1)$$



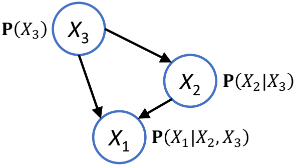


Bayesian network structure

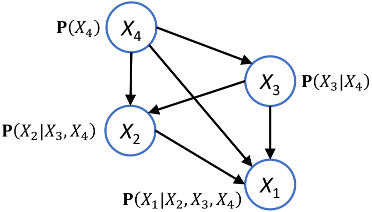
By definition a BN is a fully connected DAG, i.e., there is an edge (regardless of the direction) between **every** pair of nodes

Examples:

$$\mathbf{P}(X_1, X_2, X_3) = \mathbf{P}(X_1|X_2, X_3)\mathbf{P}(X_2|X_3)\mathbf{P}(X_3)$$



$$\mathbf{P}(X_1, X_2, X_3, X_4) = \mathbf{P}(X_1|X_2, X_3, X_4)\mathbf{P}(X_2|X_3, X_4) \times \mathbf{P}(X_3|X_4)\mathbf{P}(X_4)$$



Probabilistic inference with Bayesian networks

We know that any probabilistic inference can be carried out using the full joint PDF; if it is rewritten using the chain rule, the corresponding conditional PDFs can be used, instead

Example: starting from $\mathbf{P}(X_1, X_2, X_3) = \mathbf{P}(X_1|X_2, X_3)\mathbf{P}(X_2|X_3)\mathbf{P}(X_3)$, compute $\mathbf{P}(X_1|X_2)$

1. using the definition of conditional probability and the sum rule:

$$\mathbf{P}(X_1|X_2) = \frac{\mathbf{P}(X_1, X_2)}{\mathbf{P}(X_2)} = \frac{\sum_{x_3} \mathbf{P}(X_1, X_2, X_3 = x_3)}{\sum_{x_1, x_3} \mathbf{P}(X_1 = x_1, X_2, X_3 = x_3)} = \dots$$

2. using the above expression of the chain rule:

$$\dots = \frac{\sum_{x_3} \mathbf{P}(X_1|X_2, X_3 = x_3)\mathbf{P}(X_2|X_3 = x_3)P(X_3 = x_3)}{\sum_{x_1, x_3} \mathbf{P}(X_1 = x_1|X_2, X_3 = x_3)\mathbf{P}(X_2|X_3 = x_3)P(X_3 = x_3)}$$

Probabilistic inference with Bayesian networks

Only using the chain rule to represent a joint PDF does **not** reduce the effort required to estimate it

Example: to estimate the joint PDF $\mathbf{P}(X_1, X_2, X_3)$, one needs to specify $2^3 - 1 = 7$ probability values

Considering any expression of the chain rule, e.g.:

$$\mathbf{P}(X_1, X_2, X_3) = \mathbf{P}(X_1 | X_2, X_3)\mathbf{P}(X_2 | X_3)\mathbf{P}(X_3) \ ,$$

estimating the corresponding distributions requires:

- $\mathbf{P}(X_1 | X_2, X_3)$: $2^2 \times (2^1 - 1) = 4$ values,
- $\mathbf{P}(X_2 | X_3)$: $2^1 \times (2^1 - 1) = 2$ values,
- $\mathbf{P}(X_3)$: $2^1 - 1 = 1$ value,

for the same total of 7 probability values

Conditional independence relations in Bayesian networks

On the other hand, we have also seen that conditional independence relations reduce the effort of for estimating conditional distributions

Example: considering again a possible expression of the joint PDF

$$\mathbf{P}(X_1, X_2, X_3) = \mathbf{P}(X_1|X_2, X_3)\mathbf{P}(X_2|X_3)\mathbf{P}(X_3) \ ,$$

if X_1 is conditionally independent of X_2 given X_3, i.e.,

$$\mathbf{P}(X_1|X_2, X_3) = \mathbf{P}(X_1|X_3) \ ,$$

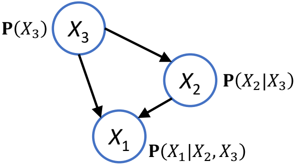
to estimate $\mathbf{P}(X_1|X_3)$ one needs to specify $2 \times (2^1 - 1) = 2$ probability values instead of $2^2 \times (2^1 - 1) = 4$

Conditional independence relations in Bayesian networks

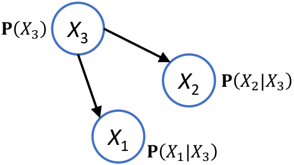
In a BN conditional independence relations are expressed by the **absence** of the corresponding edges: the resulting DAG is **no more** fully connected

Example: left: a possible, general expression of the joint PDF $\mathbf{P}(X_1, X_2, X_3)$; right: the expression obtained by the conditional independence assumption $\mathbf{P}(X_1|X_2, X_3) = \mathbf{P}(X_1|X_3)$: notice the absence of the corresponding edge from X_2 to X_1

$\mathbf{P}(X_1|X_2, X_3)\mathbf{P}(X_2|X_3)\mathbf{P}(X_3)$



$\mathbf{P}(X_1|X_3)\mathbf{P}(X_2|X_3)\mathbf{P}(X_3)$

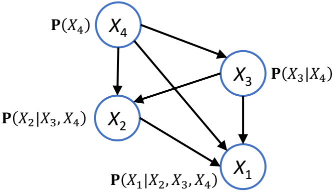


Conditional independence relations in Bayesian networks

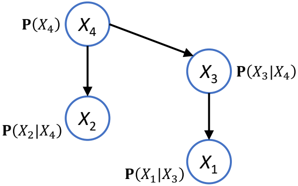
Example: left: a possible, general expression of the joint PDF $\mathbf{P}(X_1, X_2, X_3, X_4)$; right: the one obtained by the following conditional independence assumptions, with the corresponding BNs

- X_1 is conditionally independent of X_2 and X_4 given X_3:
  $\mathbf{P}(X_1|X_2, X_3, X_4) = \mathbf{P}(X_1|X_3)$
- X_2 is conditionally independent of X_3 given X_4:
  $\mathbf{P}(X_2|X_3, X_4) = \mathbf{P}(X_2|X_4)$

$$\mathbf{P}(X_1|X_2, X_3, X_4)\mathbf{P}(X_2|X_3, X_4) \times \mathbf{P}(X_3|X_4)\mathbf{P}(X_4)$$



$$\mathbf{P}(X_1|X_3)\mathbf{P}(X_2|X_4)\mathbf{P}(X_3|X_4)\mathbf{P}(X_4)$$



Conditional independence relations in Bayesian networks

How many proability values need to be specified to estimate the joint PDF of the previous example?

- general form: $\mathbf{P}(X_1|X_2, X_3, X_4)\mathbf{P}(X_2|X_3, X_4)\mathbf{P}(X_3|X_4)\mathbf{P}(X_4)$
  - $\mathbf{P}(X_1|X_2, X_3, X_4)$: $2^3 \times (2^1 - 1) = 8$ values,
  - $\mathbf{P}(X_2|X_3, X_4)$: $2^2 \times (2^1 - 1) = 4$ values,
  - $\mathbf{P}(X_3|X_4)$: $2 \times (2^1 - 1) = 2$ values,
  - $\mathbf{P}(X_4)$: $2^1 - 1 = 1$ value,

  for a total of **15** values

- taking into account conditional independence relations: $\mathbf{P}(X_1|X_3)\mathbf{P}(X_2|X_4)\mathbf{P}(X_3|X_4)\mathbf{P}(X_4)$
  - $\mathbf{P}(X_1|X_3)$: $2 \times (2^1 - 1) = 2$ values,
  - $\mathbf{P}(X_2|X_4)$: $2 \times (2^1 - 1) = 2$ values,
  - $\mathbf{P}(X_3|X_4)$: $2 \times (2^1 - 1) = 2$ values,
  - $\mathbf{P}(X_4)$: $2^1 - 1 = 1$ value,

  for a total of **7** values

Conditional independence relations in Bayesian networks

Note that conditional independence relations can be exploited to simplify the expression of a joint PDF, and the corresponding BN, **only** if the chain rule is applied in the "correct" order

Example: two possible expressions of the joint PDF $\mathbf{P}(X_1, X_2, X_3, X_4)$:

$$\mathbf{P}(X_1|X_2, X_3, X_4)\mathbf{P}(X_2|X_3, X_4)\mathbf{P}(X_3|X_4)\mathbf{P}(X_4) \tag{1}$$

$$\mathbf{P}(X_4|X_3, X_2, X_1)\mathbf{P}(X_3|X_2, X_1)\mathbf{P}(X_2|X_1)\mathbf{P}(X_1) \tag{2}$$

Under the following conditional independence relations:

- $\mathbf{P}(X_1|X_2, X_3, X_4) = \mathbf{P}(X_1|X_3)$,
- $\mathbf{P}(X_2|X_3, X_4) = \mathbf{P}(X_2|X_4)$,

expression (1) can be simplified into

$$\mathbf{P}(X_1|X_3)\mathbf{P}(X_2|X_4)\mathbf{P}(X_3|X_4)\mathbf{P}(X_4) \ ,$$

whereas expression (2) **cannot**

Conditional independence relations in Bayesian networks

Example: we have seen that *Catch* and *Toothache* can be assumed to be conditionally independent given *Cavity*:

$$\mathbf{P}(Catch|Toothache, Cavity) = \mathbf{P}(Catch|Cavity)$$

If the chain rule is applied to their joint PDF as:

$$\begin{aligned}&\mathbf{P}(Catch, Toothache, Cavity)\\&= \mathbf{P}(Catch|Toothache, Cavity)\mathbf{P}(Toothache|Cavity)\mathbf{P}(Cavity)\ ,\end{aligned}$$

it can be simplified, thanks to the above assumption, into:

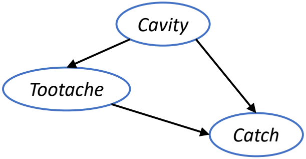
$$\mathbf{P}(Catch|Cavity)\mathbf{P}(Toothache|Cavity)\mathbf{P}(Cavity)$$

Conditional independence relations in Bayesian networks

BNs corresponding to the joint PDFs of the previous example:

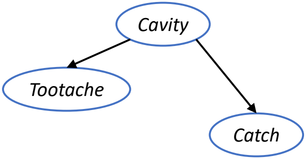
general expression:

$\mathbf{P}(Catch|Toothache, Cavity)$
$\times\mathbf{P}(Toothache|Cavity)\mathbf{P}(Cavity)$



simpler expression:

$\mathbf{P}(Catch|Cavity)$
$\times\mathbf{P}(Toothache|Cavity)\mathbf{P}(Cavity)$



Defining Bayesian networks

A **general** approach to choose the “best” order between variables when applying the chain rule:

> *select the “root cause” variables first, then the ones they **directly** influence, and so on, until reaching the variables which have **no direct causal influence** on the others, i.e., the “end effect” variables*

The same approach allows to **directly** define the structure of the corresponding BN, by adding nodes one at a time in the same order suggested above, together with the corresponding edges, **without** the need of deriving it from the expression of the joint PDF obtained by the chain rule

Defining Bayesian networks: an example

You have a new **burglar alarm** installed at home.

It is fairly reliable at detecting a burglary, but also responds on occasion to minor **earthquakes**.

You also have two neighbors, **John** and **Mary**, who have promised to **call** you at work when they **hear the alarm**.

John always calls when he hears the alarm, but sometimes **confuses** the telephone ringing with the alarm and calls then, too. Mary, on the other hand, likes rather loud music and sometimes **misses** the alarm altogether.

You may be interested in different probabilistic inferences, e.g., estimating the probability of a burglary given the evidence of who has or has not called.

taken from Russell and Norvig, *Artificial Intelligence – A modern Approach*, 2nd Ed., Pearson, 2003

Defining Bayesian networks: an example

First step: what **random variables** should be used to represent the events of interest?

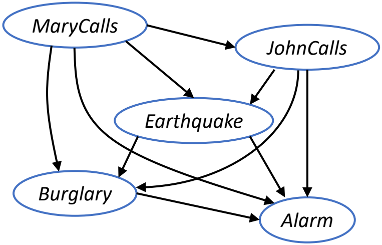
A possible choice is the following set of Boolean variables, describing the occurrence of certain events, e.g., over a given day

- *Alarm* (*A* for brevity): whether the alarm sounded or not
- *Burglary* (*B*): whether a burglary occurred or not
- *Earthquake* (*E*): whether an earthquake occurred or not
- *JohnCalls* (*J*): whether John called or not
- *MaryCalls* (*M*): whether Mary called or not

Defining Bayesian networks: an example

One of the possible expressions of the full joint PDF obtained using the chain rule, and the corresponding (fully connected) BN:

$$
\begin{aligned}
&\mathbf{P}(A, B, E, J, M) \\
&= \mathbf{P}(A|B, E, J, M)\mathbf{P}(B|E, J, M)\mathbf{P}(E|J, M)\mathbf{P}(J|M)\mathbf{P}(M)
\end{aligned}
$$



Defining Bayesian networks: an example

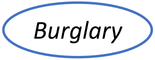
To identify conditional independence relations (if any), causal dependencies among the variables should be first identified

The structure of the BN can be directly defined, together with the identification of causal dependencies

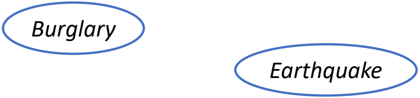
Burglaries and earthquakes can be considered as “root causes”. It can also be assumed that there are no causal dependencies among them. Note that this is only a (reasonable) **assumption**, useful to **simplify** the probabilistic model

Defining Bayesian networks: an example

Since there is no causal dependency between burglaries and earthquakes, any of them can be the first node of the BN, e.g., *Burglary*:



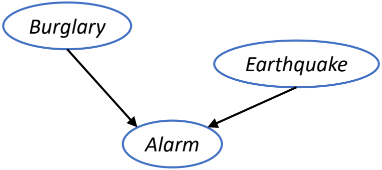
The next node to add is *Earthquake*, with **no** incoming edge from *Burglary* for the same reason above:



Defining Bayesian networks: an example

Both burglaries and earthquakes **directly** influence the state of the alarm, but **not** the fact that Mary or John calls: one can assume they will call only if they hear (or believe to hear) the alarm sounding

The next node to add is therefore *Alarm*, with incoming edges from both *Burglary* and *Earthquake*:

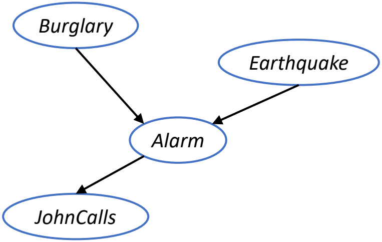


Defining Bayesian networks: an example

What about *JohnCalls* and *MaryCalls*?

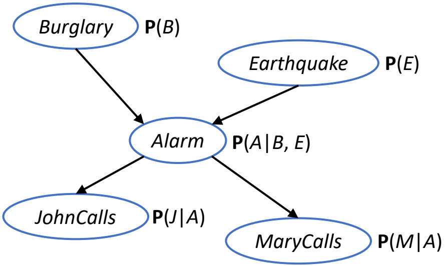
- both are **directly** influenced by the state of the alarm, but **not** by burglaries and earthquakes
- we can assume John and Mary do not communicate with each other, so neither variable **directly** influences the other

One can therefore add any of these variables first, e.g., *JohnCalls*, with an incoming edge only from *Alarm*:



Defining Bayesian networks: an example

One finally adds *MaryCalls*, again with an incoming edge only from *Alarm*, obtaining the complete BN:



Note that the above BN is **not** fully connected as the previous one: it contains only 4 edges instead of 10

Note also that **all** the conditional PDFs associated to the nodes of the BN are in causal form, $P(\mathit{effect}|\mathit{causes})$, which usually are relatively simpler to estimate

Defining Bayesian networks: an example

The considered order among the variables, from “root causes” to “end effects,” is:

$$Burglary, Earthquake, Alarm, JohnCalls, MaryCalls$$

The corresponding, **general** expression of the full joint PDF (**without** conditional independence assumptions) is:

$$\mathbf{P}(M|J, A, E, B)\mathbf{P}(J|A, E, B)\mathbf{P}(A|E, B)\mathbf{P}(E|B)\mathbf{P}(B)$$

The **equivalent** expression represented by the BN (**including** conditional independence assumptions) is:

$$\mathbf{P}(B, E, A, J, M) = \mathbf{P}(M|A)\mathbf{P}(J|A)\mathbf{P}(A|E, B)\mathbf{P}(E)\mathbf{P}(B)$$

Defining Bayesian networks: an example

Comparing the **corresponding** conditional distributions of the above expressions of the full joint PDF allows one to formally express the conditional independence assumptions made during the definition of the BN structure. The comparison shows that:

1. $\mathbf{P}(E|B) = \mathbf{P}(E)$
   Earthquake is (absolutely) independent of *Burglary*
2. $\mathbf{P}(J|A, E, B) = \mathbf{P}(J|A)$
   JohnCalls is conditionally independent of *Burglary* and *Earthquake* given *Alarm*
3. $\mathbf{P}(M|J, A, E, B) = \mathbf{P}(M|A)$
   MaryCalls is conditionally independent of *Burglary*, *Earthquake* and *JohnCalls* given *Alarm*

Defining Bayesian networks: an example

In general a BN implies other conditional independence relations besides the ones identified as in the previous example (see the textbook)

Attention should be paid to avoid drawing **wrong** conclusions about independence relations represented by a given BN, e.g.:

- the absence of edges pointing to *Burglary* does **not** mean it is independent of all the other variables: it is just due to the choice of adding it as the first node of the BN
- the absence of an edge between *MaryCalls* and *Burglary* does **not** mean they are independent of each other (i.e., $\mathbf{P}(M|B) \neq \mathbf{P}(M)$): it is due to the fact that *Alarm*, a **direct effect** of a burglar entering home and a **direct cause** of Mary calling, has been added to the BN **after** *Burglary* and **before** *MaryCalls*

Defining Bayesian networks: an example

To define the full joint PDF $\mathbf{P}(B, E, A, J, M)$, $2^5 - 1 = 31$ probability values must be specified

The expression represented by the BN requires, instead:

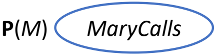
- $\mathbf{P}(M|A)$: $2 \times (2^1 - 1) = 2$ values,
- $\mathbf{P}(J|A)$: $2 \times (2^1 - 1) = 2$ values,
- $\mathbf{P}(A|E, B)$: $2^2 \times (2^1 - 1) = 4$ values,
- $\mathbf{P}(E)$: $2^1 - 1 = 1$ value,
- $\mathbf{P}(B)$: $2^1 - 1 = 1$ value,

for a total of 10 probability values

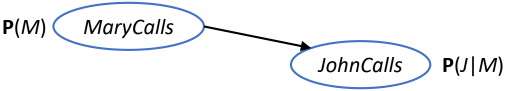
Defining Bayesian networks: an example

What if a different order among the variables is chosen, e.g.: *MaryCalls*, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*?

- adding *MaryCalls*:

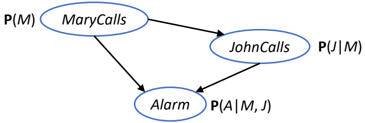


- adding *JohnCalls*: if Mary calls, it is likely the alarm has sounded, which makes it more likely that John calls: *JohnCalls* needs *MaryCalls* as a parent

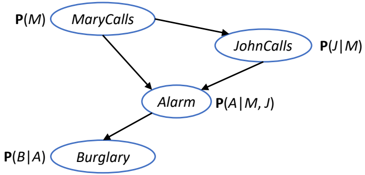


Defining Bayesian networks: an example

- adding *Alarm*: if both call, it is more likely that the alarm has sounded than if just one or neither call: both *MaryCalls* and *JohnCalls* are needed as parents

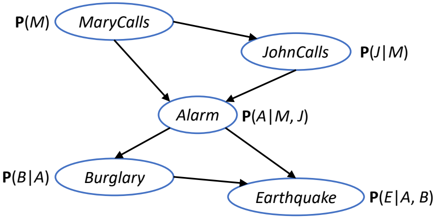


- adding *Burglary*: if we know the alarm is sounding, a call from John or Mary does not provide additional information about burglary: only *Alarm* is needed as parent



Constructing Bayesian networks: an example

- adding *Earthquake*: if the alarm is sounding, it is more likely that an earthquake occurred; if we know that also a burglary occurred, this explains the alarm, and the probability of an earthquake would be only slightly above normal; a call from John or Mary does not provide any additional information on earthquakes: only *Alarm* and *Burglary* are needed as parents



Constructing Bayesian networks: an example

The above BN represents the following expression of the full joint PDF:

$$\mathbf{P}(E, B, A, J, M) = \mathbf{P}(E|A, B)\mathbf{P}(B|A)\mathbf{P}(A|J, N)\mathbf{P}(J|M)\mathbf{P}(M)$$

It has **six** edges, **two more** than the previous BN: therefore it requires **more** probability values to be specified

Moreover, the associated conditional PDFs are more difficult to estimate: they require unnatural probability judgments, in particular **non-causal** ones, e.g., the probability of an earthquake given that a burglary did (not) occur and the alarm did (not) sound, $\mathbf{P}(E|B, A)$

Completeness and accuracy of probabilistic models

Are the previous BNs exact and complete probabilistic models?

- they are **approximate** rather than exact models, due to the underlying **assumptions**; if the assumptions are reasonable, the approximation might be accurate enough for **practical** purposes
- they may also appear **incomplete**: other possible causes of the considered events are not **explicitly** represented, e.g.:
  - no node for Mary's currently listening to loud music
  - no node for the telephone ringing and confusing John

  actually all such causes are **implicitly** summarised in the conditional PDFs, e.g., a non-zero value for $P(M = \mathrm{t}|A = \mathrm{f})$ accounts for **all** possible causes that make Mary call even if the alarm did not sound

Completeness and accuracy of probabilistic models

The choice of not representing **all** possible factors **explicitly** is an example of **laziness** and **ignorance**: it would be very difficult or even impossible to consider them all and assess their likelihood

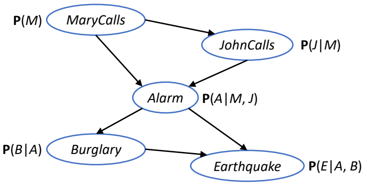
On the other hand, the conditional PDFs associated to a BN **summarise** a **potentially infinite** set of circumstances in which the considered events can happen or not, e.g.:

- $\mathbf{P}(A|E, B)$ summarise all the other causes, beside earthquakes and burglaries, that make the alarm sound or fail to sound (passing helicopter, high humidity, power failure, dead battery, cut wires, a dead mouse stuck inside the bell, etc.)
- $\mathbf{P}(J|A)$ and $\mathbf{P}(M|A)$ summarise all the other causes, besides the state of the alarm, that make John or Mary call or fail to call (out to lunch, on vacation, listening to loud music, passing helicopter, etc.)

This is the way in which probabilistic models can be kept small enough to be **practically useful**, albeit only **approximate**

Exercises

1. Write the expression of the full joint PDF represented by the BN shown on the right



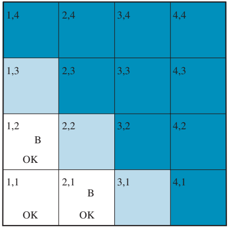
2. Identify the conditional independence relations explicitly represented by this BN
3. Determine the number of probability values to be specified to define the PDFs associated to this BN
4. Define a BN by introducing the variables in the following order: *MaryCalls*, *JohnCalls*, *Earthquake*, *Burglary*, *Alarm*; then repeat steps 1–3 for the new BN

Exercise: the wumpus world revisited

Two properties of the wumpus world

- a pit causes breezes in all neighboring squares
- each square other than $(1,1)$ can contain a pit with probability 0.2

If the agent has already explored squares $(1,1), (1,2)$ and $(2,1)$, and its current knowledge is the one represented on the right, each of the three reachable squares $(1,3)$, $(2,2)$ and $(3,1)$ can contain a pit



1. define a suitable BN to represent knowledge about pits
2. what is the probability that each of the three reachable squares contains a pit?
3. what conclusions about the presence of pits in these squares can be drawn using logic-based knowledge representation and inference?

Inference in Bayesian networks

Exact inference in Bayesian networks

Basic task of probabilistic inference systems: computing the **posterior** PDF of a subset of query variables $\mathbf{Q}$ given the observed values $\mathbf{e}$ of a distinct subset of evidence variables $\mathbf{E}$:

$$\mathbf{P}(\mathbf{Q}|\mathbf{E}=\mathbf{e})$$

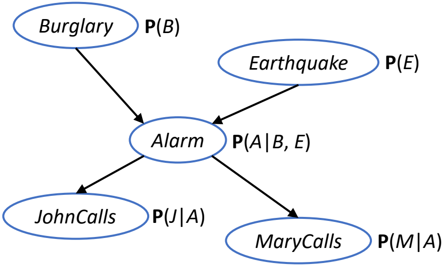
As previously shown, any probabilistic inference can be carried out from the full joint PDF, using the definition of posterior probability and the sum rule (let $\mathbf{Y}$ denote the variables that are not part of $\mathbf{Q}$ nor of $\mathbf{E}$):

$$\mathbf{P}(\mathbf{Q}|\mathbf{E}=\mathbf{e}) = \frac{\mathbf{P}(\mathbf{Q},\mathbf{E}=\mathbf{e})}{\mathbf{P}(\mathbf{E}=\mathbf{e})} = \frac{\sum_{\mathbf{y}} \mathbf{P}(\mathbf{Q},\mathbf{E}=\mathbf{e},\mathbf{Y}=\mathbf{y})}{\sum_{\mathbf{q},\mathbf{y}} P(\mathbf{Q}=\mathbf{q},\mathbf{E}=\mathbf{e},\mathbf{Y}=\mathbf{y})}$$

If the full joint PDF is represented by a BN, the numerator and denominator can be rewritten in terms of the corresponding chain rule

Exact inference in Bayesian networks: example

Consider the BN of the previous example:



(cont.)

Exact inference in Bayesian networks: example

After a call by both Mary and John, one may want to know the posterior PDF of a burglary:

$$\mathbf{P}(B|J = \mathrm{true}, M = \mathrm{true})$$

Note that this conditional PDF is not associated to one of the nodes of the BN. Using the definition of conditional probability and then the sum rule over the full joint PDF:

$$\mathbf{P}(B|J = \mathrm{t}, M = \mathrm{t}) = \frac{\mathbf{P}(B, J = \mathrm{t}, M = \mathrm{t})}{P(J = \mathrm{t}, M = \mathrm{t})} =$$

$$\frac{\sum_{e,a \in \{\mathrm{t},\mathrm{f}\}} \mathbf{P}(B, J = \mathrm{t}, M = \mathrm{t}, E = e, A = a)}{\sum_{b,e,a \in \{\mathrm{t},\mathrm{f}\}} \mathbf{P}(B = b, J = \mathrm{t}, M = \mathrm{t}, E = e, A = a)}$$

(cont.)

Exact inference in Bayesian networks: example

The numerator and denominator of the last expression can be finally rewritten using the expression of the chain rule represented by the BN:

$$\frac{\sum_{e,a\in\{\mathrm{t},\mathrm{f}\}} P(M{=}\mathrm{t}|A{=}a)P(J{=}\mathrm{t}|A{=}a)\mathbf{P}(A{=}a|E{=}e,B)P(E{=}e)\mathbf{P}(B)}{\sum_{b,e,a\in\{\mathrm{t},\mathrm{f}\}} P(M{=}\mathrm{t}|A{=}a)P(J{=}\mathrm{t}|A{=}a)P(A{=}a|E{=}e,B{=}b)P(E{=}e)P(B{=}b)}$$

ù Note that all the PDFs in the above expression are assumed to

be known, as part of the definition of the probabilistic model represented by the considered BN

Computational complexity of exact inference

The exact inference procedure presented above has an **exponential** worst-case time and space complexity in the number of variables

A practical solution is to trade exactness for a lower complexity

To this aim, approximate inference methods based on randomised sampling, known as Monte Carlo (MC) algorithms, can be used

MC algorithms have become widespread in computer science to estimate quantities difficult to calculate exactly

In the following, the simplest, direct sampling family of MC algorithms is considered. More efficient inference methods exist, like the Markov Chain MC algorithm

Direct sampling for probabilistic inference

Primitive element of sampling algorithms: direct sampling from a known **prior** probability distribution (i.e., with **no** evidence)

Example: the outcome of flipping an unbiased coin can be represented as a random variable $\mathit{Coin} \in \langle \mathrm{heads}, \mathrm{tails} \rangle$ with prior distribution $\mathbf{P}(\mathit{Coin}) = \langle 0.5, 0.5 \rangle$

Sampling from $\mathbf{P}(\mathit{Coin})$ is equivalent to flipping the coin

This can be achieved, e.g., by

- generating a random number $x \in [0, 1]$ from a uniform probability density function (e.g., using library functions of programming languages, such as Python's `random`)
- setting $\mathit{Coin} = \mathrm{heads}$, if $x < 0.5$, and $\mathit{Coin} = \mathrm{tails}$ otherwise

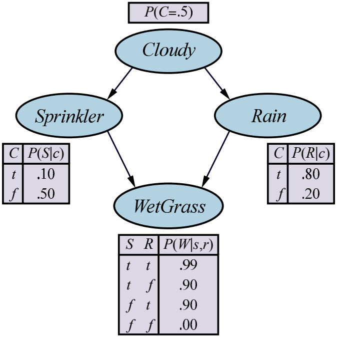
Direct sampling in Bayesian networks

The same method can be used for approximate inference in BNs, using a BN as a generative model

The direct sampling process consists of sampling a value for each variable X in topological order, i.e., after a sample $\mathbf{z}$ has already been drawn for **all** its parents $\mathbf{Z} = pa(X)$, from the corresponding PDF conditioned on $\mathbf{z}$: $\mathbf{P}(X|pa(X) = \mathbf{z})$

Direct sampling in Bayesian networks: example

A BN that describes a person's daily lawn routine: each morning, she checks the weather; if it is cloudy, she usually does not turn on the sprinkler; if the sprinkler is on, or if it rains during the day, the grass will be wet



$P(C{=}.5)$

C	$P(S\|c)$
t	.10
f	.50

C	$P(R\|c)$
t	.80
f	.20

S	R	$P(W\|s,r)$
t	t	.99
t	f	.90
f	t	.90
f	f	.00

Direct sampling in Bayesian networks: example

A topological order compatible with the above BN is: *Cloudy*, *Sprinkler*, *Rain*, *WetGrass*. Direct sampling proceeds as follows

1. sample from $\mathbf{P}(\mathit{Cloudy}) = \langle 0.5, 0.5\rangle$
   suppose this returns $\mathit{Cloudy} = \mathrm{true}$
2. sample from $\mathbf{P}(\mathit{Sprinkler}|\mathit{Cloudy} = \mathrm{t}) = \langle 0.1, 0.9\rangle$
   suppose this returns $\mathit{Sprinkler} = \mathrm{false}$
3. sample from $\mathbf{P}(\mathit{Rain}|\mathit{Cloudy} = \mathrm{t}) = \langle 0.8, 0.2\rangle$
   suppose this returns $\mathit{Rain} = \mathrm{true}$
4. sample from $\mathbf{P}(\mathit{WetGrass}|\mathit{Sprinkler} = \mathrm{f}, \mathit{Rain} = \mathrm{t}) = \langle 0.9, 0.1\rangle$
   suppose this returns $\mathit{WetGrass} = \mathrm{true}$

The generated sample is therefore:

$$\mathit{Cloudy} = \mathrm{true},\ \mathit{Sprinkler} = \mathrm{false},\ \mathit{Rain} = \mathrm{true},\ \mathit{WetGrass} = \mathrm{true}$$

Direct sampling: justification

Let $P_{\mathrm{DS}}(X_1 = x_1, \ldots, X_n = x_n)$ denote the probability that event $(x_1, \ldots, x_n)$ is generated by the direct sampling process from a BN

Considering the sampling process, it is easy to see that

$$P_{\mathrm{DS}}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{k=1}^{n} P(X_k = x_k | pa(X_k))$$

The right-hand side of the above expression equals by definition the full joint PDF of the n variables, therefore:

$$P_{\mathrm{DS}}(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$$

(cont.)

Direct sampling: justification

If N samples are drawn, denoting by $N_{\mathrm{DS}}(x_1, \ldots, x_n)$ the number of occurrences of any event $\mathbf{x} = (x_1, \ldots, x_n)$, the frequency of $\mathbf{x}$ is expected to converge to the probability of $\mathbf{x}$, as N increases:

$$\begin{aligned} \lim_{N \to \infty} \frac{N_{\mathrm{DS}}(x_1, \ldots, x_n)}{N} &= P_{\mathrm{DS}}(X_1 = x_1, \ldots, X_n = x_n) \\ &= P(X_1 = x_1, \ldots, X_n = x_n) \end{aligned}$$

For instance, in the previous example:

$$\begin{aligned} &P_{\mathrm{DS}}(\mathit{Cloudy} = \mathrm{t}, \mathit{Sprinkler} = \mathrm{f}, \mathit{Rain} = \mathrm{t}, \mathit{WetGrass} = \mathrm{t}) \\ &= P(\mathit{Cloudy} = \mathrm{t}) \times P(\mathit{Sprinkler} = \mathrm{f} | \mathit{Cloudy} = \mathrm{t}) \\ &\quad \times P(\mathit{Rain} = \mathrm{t} | \mathit{Cloudy} = \mathrm{t}) \times P(\mathit{WetGrass} = \mathrm{t} | \mathit{Sprinkler} = \mathrm{f}, \mathit{Rain} = \mathrm{t}) \\ &= 0.5 \times 0.9 \times 0.8 \times 0.9 = 0.324 \end{aligned}$$

Therefore, for large N one expects 32.4% samples to be of this event

Inference in Bayesian networks: rejection sampling

Rejection sampling generates samples from a hard-to-sample distribution given an easy-to-sample one

It can be used to compute a conditional probability $P(\mathbf{Q} = \mathbf{q} | \mathbf{E} = \mathbf{e})$ from the full joint PDF represented by a BN:

1. generate N samples from the BN using direct sampling
2. reject (discard) all the samples that do not match the evidence $\mathbf{e}$
3. among the $N' \leq N$ remaining samples, let $N'' \leq N'$ be the number of samples for which $\mathbf{Q} = \mathbf{q}$: estimate $P(\mathbf{Q} = \mathbf{q} | \mathbf{E} = \mathbf{e})$ as N''/N'

The larger the number of samples N, the more accurate the estimate

The justification of rejection sampling is similar to that of direct sampling

Rejection sampling: example

Assume one wishes to estimate $P(Rain = \text{true}|Sprinkler = \text{true})$ from $N = 100$ samples generated by the direct sampling algorithm

Among the generated samples, suppose 73 have $Sprinkler = \text{false}$: these samples are **rejected**

Among the remaining $N' = 27$ samples, assume $N'' = 8$ have $Rain = \text{true}$. Therefore:

$$P(Rain = \text{true}|Sprinkler = \text{true}) \approx \frac{N''}{N'} = \frac{8}{27} \simeq 0.296$$

Inference in Bayesian networks: likelihood weighting

Rejection sampling is **inefficient**: the lower the prior probability of the evidence, $P(\mathbf{E} = \mathbf{e})$, the larger the number of rejected samples before generating enough samples consistent with the evidence to **accurately** estimate $P(\mathbf{Q} = \mathbf{q} | \mathbf{E} = \mathbf{e})$

Likelihood weighting avoids this inefficiency: it fixes the evidence variables and generates only events that are consistent with the evidence

Among the N generated samples, each sample where $\mathbf{Q} = \mathbf{q}$ is not counted as one, but is weighted by the likelihood that it accords to the evidence

The value of $P(\mathbf{Q} = \mathbf{q} | \mathbf{E} = \mathbf{e})$ is finally estimated as the sum of such weights divided by N

Also the justification of the likelihood weighting algorithm is similar to that of direct sampling

Likelihood weighting: example

Assume one wishes to estimate $P(\mathit{Rain} = \mathrm{t} | \mathit{Sprinkler} = \mathrm{t}, \mathit{WetGrass} = \mathrm{t})$:

1. set the likelihood weight $w \leftarrow 1.0$
2. sample from $\mathbf{P}(\mathit{Cloudy}) = \langle 0.5, 0.5 \rangle$
   suppose this returns $\mathit{Cloudy} = \mathrm{true}$
3. *Sprinkler* is an evidence variable with value true:
   set $w \leftarrow w \times P(\mathit{Sprinkler} = \mathrm{t} | \mathit{Cloudy} = \mathrm{t}) = 0.1$
4. sample from $\mathbf{P}(\mathit{Rain} | \mathit{Cloudy} = \mathrm{t}) = \langle 0.8, 0.2 \rangle$
   suppose this returns $\mathit{Rain} = \mathrm{true}$
5. *WetGrass* is an evidence variable with value true:
   set $w \leftarrow w \times P(\mathit{WetGrass} = \mathrm{t} | \mathit{Sprinkler} = \mathrm{t}, \mathit{Rain} = \mathrm{t}) = 0.099$

The generated sample, with likelihood weight 0.099, is therefore:

$$\mathit{Cloudy} = \mathrm{true},\ \mathit{Sprinkler} = \mathrm{true},\ \mathit{Rain} = \mathrm{true},\ \mathit{WetGrass} = \mathrm{true}$$

Likelihood weighting: example

The sample generated in the previous example will be counted among the events for which *Rain* $= \mathrm{true}$ with a low likelihood weight of 0.099

This can be intuitively explained by the fact that the generated sample corresponds to a cloudy day, which makes the sprinkler unlikely to be on

In other words, among samples that accord with the evidence, one expects 9.9% samples to have *Rain* $= \mathrm{true}$ and *Cloudy* $= \mathrm{true}$