

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 networks 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)

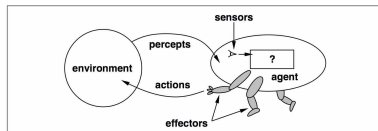
For a formal and comprehensive introduction to probability theory and statistics, students are referred to textbooks such as:

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)

1,4	2,4	3,4	4,4
1,3	2,3	3,3	4,3
1,2 B OK	2,2	3,2	4,2
1,1 OK	2,1 B OK	3,1	4,1

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:

$$\forall p \text{ Symptom}(p, \text{Toothache}) \Rightarrow \text{Disease}(p, \text{Cavity})$$

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

$$\forall p \text{ Symptom}(p, \text{Toothache}) \Rightarrow \\ \text{Disease}(p, \text{Cavity}) \vee \text{Disease}(p, \text{Abscess}) \vee \dots$$

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

$$\forall p \text{ Disease}(p, \text{Cavity}) \Rightarrow \text{Symptom}(p, \text{Toothache})$$

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 , *Weather*, *Toothache*
- ▶ **domain** (set of **mutually exclusive** values)
 - **Boolean**, e.g.: $Toothache \in \{\text{true}, \text{false}\}$
 - **discrete**, e.g.: $Weather \in \{\text{sunny}, \text{rainy}, \text{cloudy}, \text{snow}\}$,
 $M \in \mathbb{N}$
 - **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.: $\langle \text{true}, \text{false} \rangle$, $\langle \text{sunny}, \text{rain}, \text{cloudy}, \text{snow} \rangle$

Probability distribution function (PDF)

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

$$P(Weather = \text{sunny}) = 0.7$$

$$P(Weather = \text{rain}) = 0.2$$

$$P(Weather = \text{cloudy}) = 0.08$$

$$P(Weather = \text{snow}) = 0.02$$

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

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

	<i>Toothache</i> = <i>t</i>		<i>Toothache</i> = <i>f</i>	
	<i>Catch</i> = <i>t</i>	<i>Catch</i> = <i>f</i>	<i>Catch</i> = <i>t</i>	<i>Catch</i> = <i>f</i>
<i>Cavity</i> = <i>t</i>	0.108	0.012	0.072	0.008
<i>Cavity</i> = <i>f</i>	0.016	0.064	0.144	0.576

Theorem:

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

Events

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

$$Cavity = true \vee Toothache = false$$

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

$$Cavity = true \wedge Toothache = false \wedge Catch = true$$

Atomic events are **mutually exclusive** and **exhaustive**

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

$$\begin{aligned} Cavity = true \vee Toothache = false \equiv \\ (Cavity = true \wedge Toothache = true \wedge Catch = true) \vee \dots \end{aligned}$$

Probabilistic inference

Computing the probability of an event of interest

A dentist may be interested in

- ▶ $P(\textit{Cavity} = \text{true})$
- ▶ $P(\textit{Cavity} = \text{true} \wedge \textit{Toothache} = \text{true})$

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

Probabilistic inference: marginalisation (sum rule)

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

	<i>Toothache</i> = <i>t</i>		<i>Toothache</i> = <i>f</i>	
	<i>Catch</i> = <i>t</i>	<i>Catch</i> = <i>f</i>	<i>Catch</i> = <i>t</i>	<i>Catch</i> = <i>f</i>
<i>Cavity</i> = <i>t</i>	0.108	0.012	0.072	0.008
<i>Cavity</i> = <i>f</i>	0.016	0.064	0.144	0.576

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

		<i>Toothache</i>	
		true	false
<i>Cavity</i>	true	0.120	0.080
	false	0.080	0.720

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

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

Prior and posterior (conditional) probability

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

$$\blacktriangleright P(\text{Cavity} = \text{true} | \text{Toothache} = \text{true}) \triangleq \frac{P(\text{Cavity}=\text{true}, \text{Toothache}=\text{true})}{P(\text{Toothache}=\text{true})}$$

$$\blacktriangleright P(\text{Cavity} = \text{true} | \text{Catch} = \text{true}) \triangleq \frac{P(\text{Cavity}=\text{true}, \text{Catch}=\text{true})}{P(\text{Catch}=\text{true})}$$

$$\blacktriangleright P(\text{Cavity} = \text{true} | \text{Toothache} = \text{true}, \text{Catch} = \text{false}) \triangleq \frac{P(\text{Cavity}=\text{true}, \text{Toothache}=\text{true}, \text{Catch}=\text{false})}{P(\text{Toothache}=\text{true}, \text{Catch}=\text{false})}$$

$$\blacktriangleright P(\text{Cavity} = \text{true}, \text{Catch} = \text{false} | \text{Toothache} = \text{true}) \triangleq \frac{P(\text{Cavity}=\text{true}, \text{Catch}=\text{false}, \text{Toothache}=\text{true})}{P(\text{Toothache}=\text{true})}$$

Conditional PDF

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

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

Also the values of a conditional PDF sum to 1

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

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

Probabilistic inference

Also **conditional** probabilities can be computed from the **full joint** PDF

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

Full joint PDF:

	<i>Toothache</i> = t		<i>Toothache</i> = f	
	<i>Catch</i> = t	<i>Catch</i> = f	<i>Catch</i> = t	<i>Catch</i> = f
<i>Cavity</i> = t	0.108	0.012	0.072	0.008
<i>Cavity</i> = f	0.016	0.064	0.144	0.576

Sum rule:

$$P(\text{Cavity} = \text{t}, \text{Toothache} = \text{t}) = 0.108 + 0.012 = 0.120$$

$$P(\text{Toothache} = \text{t}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.200$$

Probabilistic inference: summary

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

- ▶ computing a **marginal** (prior) PDF: **sum rule**

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

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

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

Probabilistic inference using the full joint PDF: issues

How to compute or estimate the full joint PDF?

- ▶ $Coin \in \langle \text{heads}, \text{tails} \rangle$: $\mathbf{P}(Coin) = \langle ?, ? \rangle$
- ▶ $Dice \in \langle \square, \begin{smallmatrix} \cdot \\ \cdot \end{smallmatrix}, \begin{smallmatrix} \cdot & \cdot \\ \cdot & \cdot \end{smallmatrix}, \begin{smallmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{smallmatrix}, \begin{smallmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{smallmatrix}, \begin{smallmatrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{smallmatrix} \rangle$: $\mathbf{P}(Dice) = \langle ?, ?, ?, ?, ?, ? \rangle$

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

- ▶ $Coin \in \langle \text{heads}, \text{tails} \rangle$: $\mathbf{P}(Coin) = \langle \frac{1}{2}, \frac{1}{2} \rangle$
- ▶ $Dice \in \langle \square, \begin{smallmatrix} \cdot \\ \cdot \end{smallmatrix}, \begin{smallmatrix} \cdot & \cdot \\ \cdot & \cdot \end{smallmatrix}, \begin{smallmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{smallmatrix}, \begin{smallmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{smallmatrix}, \begin{smallmatrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{smallmatrix} \rangle$: $\mathbf{P}(Dice) = \langle \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6} \rangle$

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

- ▶ getting an even face up ($\begin{smallmatrix} \cdot & \cdot \\ \cdot & \cdot \end{smallmatrix}$ or $\begin{smallmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{smallmatrix}$ or $\begin{smallmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{smallmatrix}$)
- ▶ getting a sum of the faces up equal to 6 after throwing two dice ($\begin{smallmatrix} \cdot & \cdot \\ \cdot & \cdot \end{smallmatrix}$, $\begin{smallmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{smallmatrix}$ or $\begin{smallmatrix} \cdot & \cdot \\ \cdot & \cdot \end{smallmatrix}$, $\begin{smallmatrix} \cdot & \cdot \\ \cdot & \cdot \end{smallmatrix}$, ...)

Probabilistic inference using the full joint PDF: issues

How to compute or estimate the full joint PDF?

	<i>Toothache</i> = t		<i>Toothache</i> = f	
	<i>Catch</i> = t	<i>Catch</i> = f	<i>Catch</i> = t	<i>Catch</i> = f
<i>Cavity</i> = t	?	?	?	?
<i>Cavity</i> = 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 CECAL 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

	<i>Toothache</i> = t		<i>Toothache</i> = f	
	<i>Catch</i> = t	<i>Catch</i> = f	<i>Catch</i> = t	<i>Catch</i> = f
<i>Cavity</i> = t	?	?	?	?
<i>Cavity</i> = 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:

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

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

Simplifying probabilistic modelling and inference

A useful tool: the **product rule**

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

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

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

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* $\in \langle \text{sunny, rain, cloudy, snow} \rangle$

How many probability values must be estimated to specify the full joint PDF $\mathbf{P}(\textit{Toothache}, \textit{Catch}, \textit{Cavity}, \textit{Weather})$?

$$(2 \times 2 \times 2 \times 4) - 1 = \textcolor{red}{31} \dots$$

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**:

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

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:

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

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**):

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

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

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

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

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

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

Why is independence useful?

Example: n Boolean variables

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

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

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

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 $P(\text{effects}|\text{cause})$:

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

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

Simplifying probabilistic modelling and inference

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

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

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

$$\mathbf{P}(\textit{Toothache}|\textit{Catch} = \text{true}) \neq \mathbf{P}(\textit{Toothache}|\textit{Catch} = \text{false})$$

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

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

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:

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

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

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:

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

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

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

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

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

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

Usefulness of conditional independence

From the above **conditional independence** assumption it follows that:

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

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

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

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

To specify the other distributions one needs to estimate

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

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

Usefulness of conditional independence

Example: n Boolean variables

The conditional PDF

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

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

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

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

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 $P(\text{effects}|\text{cause})$: distinct effects of a common cause may be **conditionally independent**, **given** the cause, e.g.:

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

In practice, **diagnostic** inference $P(\text{cause}|\text{effects})$ 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**:

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

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

it follows that:

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

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:

$$\begin{aligned} & \mathbf{P}(Cavity|Toothache, Catch) \\ &= \frac{\mathbf{P}(Toothache, Catch|Cavity) \mathbf{P}(Cavity)}{\mathbf{P}(Toothache, Catch)} \\ &= \frac{\mathbf{P}(Toothache, Catch|Cavity) \mathbf{P}(Cavity)}{\sum_{c \in \{t, f\}} \mathbf{P}(Toothache, Catch|Cavity = c) \mathbf{P}(Cavity = c)} \end{aligned}$$

using the conditional independence assumption previously made:

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

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 $P(\text{effects}|\text{cause})$ (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, $P(\text{effects}|\text{cause})$, 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.:

end effect $\underleftarrow{\text{Catch, Toothache, Cavity}}$ 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):

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

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

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

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:

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

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, \dots, X_n :

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

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

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

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** representation 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 (BNs): **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

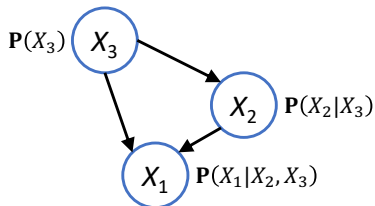
$$\begin{aligned}\mathbf{P}(X_1, X_2, \dots, X_n) &= \prod_{k=1}^n \mathbf{P}(X_k | X_{k-1}, \dots, X_1) \\ &= \prod_{k=1}^n \mathbf{P}(X_k | pa(X_k))\end{aligned}$$

- ▶ **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 $X_p \in pa(X_q)$

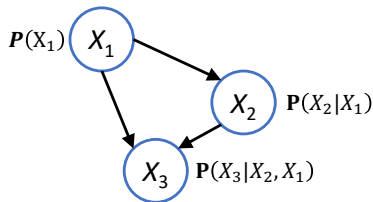
Bayesian network structure

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

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



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

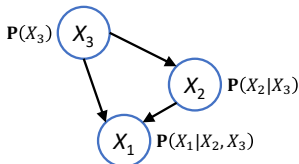


Bayesian network structure

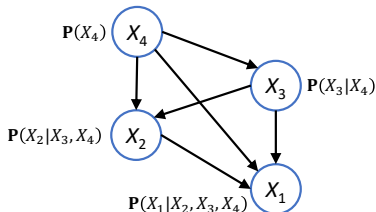
By definition a BN is a **fully connected** DAG, i.e., there is an edge (regardless of the direction) between **every** pair of nodes

Examples:

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



$$\begin{aligned} P(X_1, X_2, X_3, X_4) = \\ P(X_1|X_2, X_3, X_4)P(X_2|X_3, X_4) \times \\ P(X_3|X_4)P(X_4) \end{aligned}$$



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:

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

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.,

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

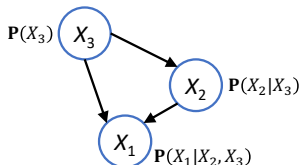
to estimate $\mathbf{P}(X_1|X_3)$ one needs to specify $2 \times (2^1 - 1) = 2$ probability values instead of $2^2 \times (2^1 - 1) = 4$

Conditional independence relations in Bayesian networks

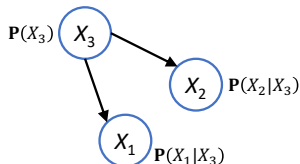
In a BN conditional independence relations are expressed by the **absence** of the corresponding edges: the resulting DAG is **no more** fully connected

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

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



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

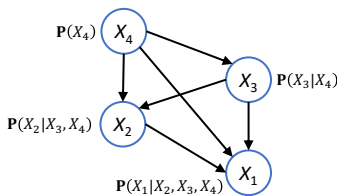


Conditional independence relations in Bayesian networks

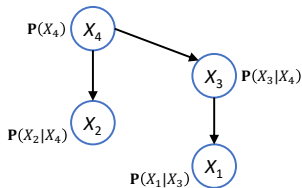
Example: left: a possible, general expression of the joint PDF $\mathbf{P}(X_1, X_2, X_3, X_4)$; right: the one obtained by the following conditional independence assumptions, with the corresponding BNs

- ▶ 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 probability 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 $\mathbf{P}(X_1, X_2, X_3, X_4)$:

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

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

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*:

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

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

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

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

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

Conditional independence relations in Bayesian networks

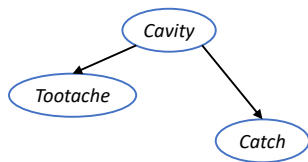
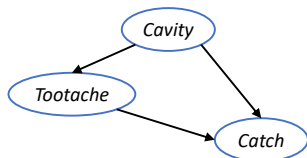
BNs corresponding to the joint PDFs of the previous example:

general expression:

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

simpler expression:

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



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?

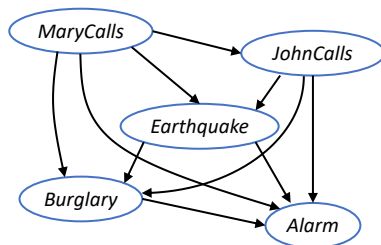
A possible choice is the following set of Boolean variables, describing the occurrence of certain events, e.g., over a given day

- ▶ *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*:



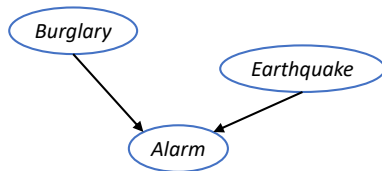
The next node to add is *Earthquake*, with **no** incoming edge from *Burglary* for the same reason above:



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*:

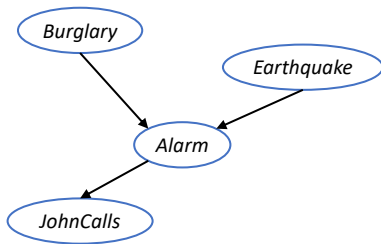


Defining Bayesian networks: an example

What about *JohnCalls* and *MaryCalls*?

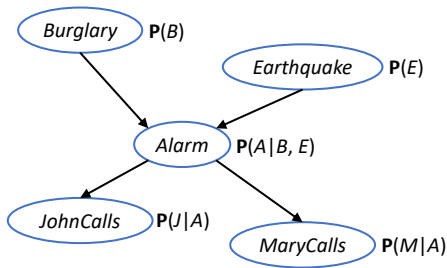
- ▶ both are **directly** influenced by the state of the alarm, but **not** by burglaries and earthquakes
- ▶ we can assume John and Mary do not communicate with each other, so neither variable **directly** influences the other

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, $P(\text{effect}|\text{causes})$, 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., $P(M|B) \neq P(M)$): 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*:

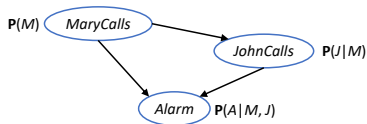


- ▶ adding *JohnCalls*: if Mary calls, it is likely the alarm has sounded, which makes it more likely that John calls: *JohnCalls* needs *MaryCalls* as a parent

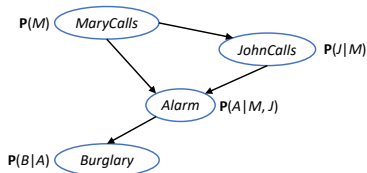


Defining Bayesian networks: an example

- 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

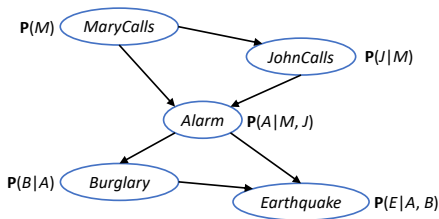


- adding *Burglary*: if we know the alarm is sounding, a call from John or Mary does not provide additional information about burglary: only *Alarm* is needed as parent



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 $P(M = t | A = f)$ 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

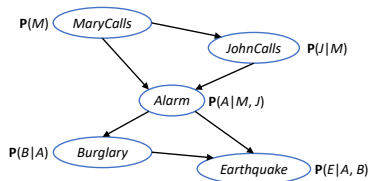
On the other hand, the conditional PDFs associated to a BN **summarise** a **potentially infinite** set of circumstances in which the considered events can happen or not, e.g.:

- ▶ $P(A|E, B)$ 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.)
- ▶ $P(J|A)$ and $P(M|A)$ 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

1,4	2,4	3,4	4,4
1,3	2,3	3,3	4,3
1,2 B OK	2,2	3,2	4,2
1,1 OK	2,1 B OK	3,1	4,1

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?