Fundamentals of Artificial Intelligence and Knowledge Representation (Module 3)

Last update: 27 October 2023

Academic Year 2023 – 2024
Alma Mater Studiorum · University of Bologna

Contents

1 Introduction

1.1 Uncertainty

Uncertainty A task is uncertain if we have: Uncertainty

- Partial observations
- Noisy or wrong information
- Uncertain action outcomes
- Complex models

A purely logic approach leads to:

- Risks falsehood: unreasonable conclusion when applied in practice.
- Weak decisions: too many conditions required to make a conclusion.

1.1.1 Handling uncertainty

Default/nonmonotonic logic Works on assumptions. An assumption can be contradicted by an evidence. Default/nonmonotonic logic

Rule-based systems with fudge factors Formulated as premise $\rightarrow_{\text{prob.}}$ effect. Have the following issues: Rule-based systems with fudge factors

- Locality: how can the probability account all the evidence.
- Combination: chaining of unrelated concepts.

Probability Assign a probability given the available known evidence. Probability

Note: fuzzy logic handles the degree of truth and not the uncertainty.

Decision theory Defined as: Decision theory

$$\text{Decision theory} = \text{Utility theory} + \text{Probability theory}$$

where the utility theory depends on one's preferences.

2 Probability

Sample space Set Ω of all possible worlds. Sample space

Event Subset $A \subseteq \Omega$. Event

Sample point/Possible world/Atomic event Element $\omega \in \Omega$. Sample point

Probability space A probability space/model is a function $\mathcal{P}(\cdot) : \Omega \to [0, 1]$ assigned to a sample space such that: Probability space

- $0 \leq \mathcal{P}(\omega) \leq 1$
- $\sum_{\omega \in \Omega} \mathcal{P}(\omega) = 1$
- $\mathcal{P}(A) = \sum_{\omega \in A} \mathcal{P}(\omega)$

Random variable A function from an event to some range (e.g. reals, booleans, ...). Random variable

Probability distribution For any random variable X: Probability distribution

$$\mathcal{P}(X = x_i) = \sum_{\omega \text{ st } X(\omega) = x_i} \mathcal{P}(\omega)$$

Proposition Event where a random variable has a certain value. Proposition

$$a = \{\omega \mid A(\omega) = \texttt{true}\}$$

$$\neg a = \{\omega \mid A(\omega) = \texttt{false}\}$$

$$(\texttt{Weather} = \texttt{rain}) = \{\omega \mid B(\omega) = \texttt{rain}\}$$

Prior probability Prior/unconditional probability of a proposition based on known evidence. Prior probability

Probability distribution (all) Gives all the probabilities of a random variable. Probability distribution (all)

$$\mathbf{P}(A) = \langle \mathcal{P}(A = a_1), \dots, \mathcal{P}(A = a_n) \rangle$$

Joint probability distribution The joint probability distribution of a set of random variables gives the probability of all the different combinations of their atomic events. Joint probability distribution

Note: Every question on a domain can, in theory, be answered using the joint distribution. In practice, it is hard to apply.

Example. $\mathbf{P}(\texttt{Weather}, \texttt{Cavity}) =$

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

Probability density function The probability density function (PDF) of a random variable X is a function $p : \mathbb{R} \to \mathbb{R}$ such that: Probability density function

$$\int_{\mathcal{T}_X} p(x) \, dx = 1$$

Uniform distribution Uniform distribution

$$p(x) = \text{Unif}[a,b](x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Gaussian (normal) distribution Gaussian (normal) distribution

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$\mathcal{N}(0,1)$ is the standard Gaussian.

Conditional probability Probability of a prior knowledge with new evidence: Conditional probability

$$\mathcal{P}(a|b) = \frac{\mathcal{P}(a \wedge b)}{\mathcal{P}(b)}$$

The product rule gives an alternative formulation:

$$\mathcal{P}(a \wedge b) = \mathcal{P}(a|b)\mathcal{P}(b) = \mathcal{P}(b|a)\mathcal{P}(a)$$

Chain rule Successive application of the product rule: Chain rule

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

Independence Two random variables A and B are independent ($A \perp B$) iff: Independence

$$\mathbf{P}(A|B) = \mathbf{P}(A) \;\text{ or }\; \mathbf{P}(B|A) = \mathbf{P}(B) \;\text{ or }\; \mathbf{P}(A,B) = \mathbf{P}(A)\mathbf{P}(B)$$

Conditional independence Two random variables A and B are conditionally independent iff: Conditional independence

$$\mathbf{P}(A \,|\, C, B) = \mathbf{P}(A \,|\, C)$$

2.1 Inference with full joint distributions

Given a joint distribution, the probability of any proposition ϕ can be computed as the sum of the atomic events where ϕ is true:

$$\mathcal{P}(\phi) = \sum_{\omega:\, \omega \models \phi} \mathcal{P}(\omega)$$

Example. Given the following joint distribution:

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

We have that:

- $\mathcal{P}(\texttt{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$
- $\mathcal{P}(\texttt{cavity} \vee \texttt{toothache}) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$
- $\mathcal{P}(\neg\texttt{cavity} \,|\, \texttt{toothache}) = \frac{\mathcal{P}(\neg\texttt{cavity}\wedge\texttt{toothache})}{\mathcal{P}(\texttt{toothache})} = \frac{0.016+0.064}{0.2} = 0.4$

Marginalization The probability that a random variable assumes a specific value is given by the sum off all the joint probabilities where that random variable assumes the given value.

Marginalization

Example. Given the joint distribution:

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

We have that $\mathcal{P}(\texttt{Weather} = \texttt{sunny}) = 0.144 + 0.576$

Conditioning Adding a condition to a probability (reduction and renormalization).

Conditioning

Normalization Given a conditional probability distribution $\mathbf{P}(A|B)$, it can be formulated as:

Normalization

$$\mathbf{P}(A|B) = \alpha\mathbf{P}(A, B)$$

where α is a normalization constant. In fact, fixed the evidence B, the denominator to compute the conditional probability is the same for each probability.

Example. Given the joint distribution:

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

We have that:

$$\mathbf{P}(\texttt{Cavity}|\texttt{toothache}) = \langle\frac{\mathcal{P}(\texttt{cavity}, \texttt{toothache}, \texttt{catch})}{\mathcal{P}(\texttt{toothache})}, \frac{\mathcal{P}(\neg\texttt{cavity}, \texttt{toothache}, \neg\texttt{catch})}{\mathcal{P}(\texttt{toothache})}\rangle$$

Probability query Given a set of query variables $\boldsymbol{Y}$, the evidence variables $\mathbf{e}$ and the other hidden variables $\boldsymbol{H}$, the probability of the query can be computed as:

Probability query

$$\mathbf{P}(\boldsymbol{Y}|\boldsymbol{E} = \mathbf{e}) = \alpha\mathbf{P}(\boldsymbol{Y}, \boldsymbol{E} = \mathbf{e}) = \alpha\sum_{\mathbf{h}}\mathbf{P}(\boldsymbol{Y}, \boldsymbol{E} = \mathbf{e}, \boldsymbol{H} = \mathbf{h})$$

The problem of this approach is that it has exponential time and space complexity that makes it not applicable in practice.

To reduce the size of the variables, conditional independence can be exploited.

Example. Knowing that $\mathbf{P} \models (\texttt{Catch} \perp \texttt{Toothache}|\texttt{Cavity})$, we can compute the distribution $\mathbf{P}(\texttt{Toothache}, \texttt{Catch}, \texttt{Cavity})$ as follows:

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

$\mathbf{P}(\texttt{Toothache}, \texttt{Catch}, \texttt{Cavity})$ has 7 independent values that grows exponentially ($2 \cdot 2 \cdot 2 = 8$ values, but one of them can be omitted as a probability always sums up to 1).

$\mathbf{P}(\texttt{Toothache} \,|\, \texttt{Cavity})\mathbf{P}(\texttt{Catch} \,|\, \texttt{Cavity})\mathbf{P}(\texttt{Cavity})$ has 5 independent values that grows linearly ($4 + 4 + 2 = 10$, but a value of $\mathbf{P}(\texttt{Cavity})$ can be omitted. The conditional probabilities require two tables (one for each prior) each with 2 values, but for each table a value can be omitted, therefore requiring 2 independent values per conditional probability instead of 4).

2.2 Bayesian networks

Bayes' rule Bayes' rule

$$\mathcal{P}(a \,|\, b) = \frac{\mathcal{P}(b \,|\, a)\mathcal{P}(a)}{\mathcal{P}(b)}$$

Bayes' rule and conditional independence Given the random variables $\texttt{Cause}$ and $\texttt{Effect}_1, \dots, \texttt{Effect}_n$, with $\texttt{Effect}_i$ independent from each other, we can compute $\mathbf{P}(\texttt{Cause}, \texttt{Effect}_1, \dots, \texttt{Effect}_n)$ as follows:

$$\mathbf{P}(\texttt{Cause}, \texttt{Effect}_1, \dots, \texttt{Effect}_n) = \left(\prod_i \mathbf{P}(\texttt{Effect}_i \,|\, \texttt{Cause}) \right) \mathbf{P}(\texttt{Cause})$$

The number of parameters is linear.

Example. Knowing that $\mathbf{P} \models (\texttt{Catch} \perp \texttt{Toothache} | \texttt{Cavity})$:

$$\begin{aligned}
&\mathbf{P}(\texttt{Cavity} \,|\, \texttt{toothache} \land \texttt{catch}) \\
&\quad = \alpha \mathbf{P}(\texttt{toothache} \land \texttt{catch} \,|\, \texttt{Cavity})\mathbf{P}(\texttt{Cavity}) \\
&\quad = \alpha \mathbf{P}(\texttt{toothache} \,|\, \texttt{Cavity})\mathbf{P}(\texttt{catch} \,|\, \texttt{Cavity})\mathbf{P}(\texttt{Cavity})
\end{aligned}$$

Bayesian network Graph for conditional independence assertions and a compact specification of full joint distributions. Bayesian network

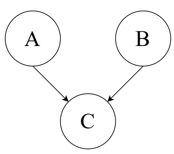
- Directed acyclic graph.
- Nodes represent variables.
- The conditional distribution of a node is given by its parents

$$\mathbf{P}(X_i \,|\, \texttt{parents}(X_i))$$

In other words, if there is an edge from A to B, then A (cause) influences B (effect).

Conditional probability table (CPT) In the case of boolean variables, the conditional distribution of a node can be represented using a table by considering all the combinations of the parents. Conditional probability table (CPT)

Example. Given the boolean variables A, B and C, with C depending on A and B, we have that:



A	B	$\mathcal{P}(c \vert A, B)$	$\mathcal{P}(\neg c \vert A, B)$
a	b	α	$1 - \alpha$
¬a	b	β	$1 - \beta$
a	¬b	γ	$1 - \gamma$
¬a	¬b	δ	$1 - \delta$

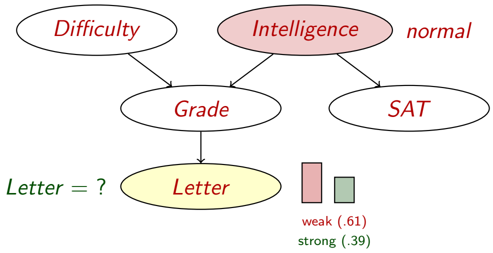
Reasoning patterns Given a Bayesian network, the following reasoning patterns can be used:

Reasoning patterns

Causal To make a prediction. From the cause, derive the effect.

Causal reasoning

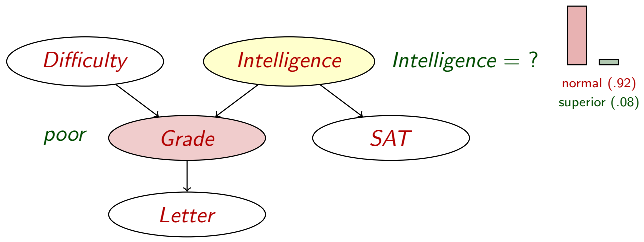
Example. Knowing `Intelligence`, it is possible to make a prediction of `Letter`.



Evidential To find an explanation. From the effect, derive the cause.

Evidential reasoning

Example. Knowing `Grade`, it is possible to explain it by estimating `Intelligence`.

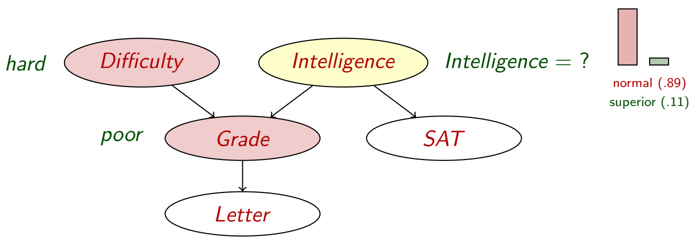


Explain away Observation obtained ”passing through” other observations.

Explain away reasoning

Example. Knowing `Difficulty` and `Grade`, it is possible to estimate `Intelligence`.

Note that if `Grade` was not known, `Difficulty` and `Intelligence` would be independent.

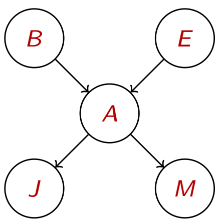


Global semantics Given a Bayesian network, the full joint distribution can be defined as the product of the local conditional distributions:

Global semantics

$$\mathcal{P}(x_1, \dots, x_n) = \prod_{i=1}^{n} \mathcal{P}(x_i \,|\, \texttt{parents}(X_i))$$

Example. Given the following Bayesian network:

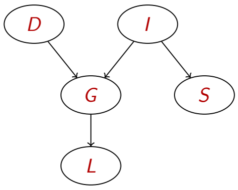


$$\mathcal{P}(j \land m \land a \land \neg b \land \neg e)$$
$$= \mathcal{P}(\neg b)\mathcal{P}(\neg e)\mathcal{P}(a \,|\, \neg b, \neg e)\mathcal{P}(j \,|\, a)\mathcal{P}(m \,|\, a)$$

Independence Intuitively, an effect is independent from a cause, if there is another cause in the middle whose value is already known.

Bayesian network independence

Example.



$$\mathbf{P} \models (\mathtt{L} \perp \mathtt{D}, \mathtt{I}, \mathtt{S} \mid \mathtt{G})$$

$$\mathbf{P} \models (\mathtt{S} \perp \mathtt{L} \mid \mathtt{G})$$

$$\mathbf{P} \models (\mathtt{S} \perp \mathtt{D}) \text{ but } \mathbf{P} \models (\mathtt{S} \not\perp \mathtt{D} \mid \mathtt{G}) \text{ (explain away)}$$