Fundamentals of Artificial Intelligence and Knowledge Representation

Module 3

Matteo Donati

August 27, 2021

Contents

Chapter 1

Introduction

In general, intelligent agents may need to handle uncertainty due to multiple factors (e.g. partial observability, non-determinism, etc.). Modelling an uncertain world with standard logic notation will lead either to false statements or to conclusions that are too weak for decision making. Some of the methods used for handling uncertainty rely on the following:

- Default or non-monotonic logic, in which reasonable assumptions are made in order to model the specific world.
- Rule-based systems with fudge factors, which introduce operators with quantified probability (e.g. $A_{25} \mapsto_{0.3} AtAirportOnTime$).
  └ action: leave for airport 25 minutes before flight
  Sometimes these fudge factors (e.g. 0.3) are not enough for "see" the evidence
- Probabilistic reasoning.
  └ most common way of handling uncertainty

1.1 Probability

Probabilistic assertions summarize uncertainty. In particular, probabilities relate propositions to one's own state of knowledge (e.g. $P(A_{25}|\text{no reported accident}) = 0.06$). Moreover, probabilities of propositions change with new evidence. Given a probabilistic model of the world one needs a decision theory in order to make decisions. └ $P(A_{25} \mid \text{no reported accident, 5 a.m.}) = 0.06$)

In order to make rational decisions (i.e. maximize utility) one needs a decision theory

1.1.1 Basic Probability Notation

possible worlds

Considering the assertions about possible worlds, **logical assertions** state which worlds are ruled out, while **probabilistic assertions** state how probable these worlds are. In particular:

- The set of all possible worlds is called the **sample space** (Ω).
- Any subset $A \subseteq \Omega$ is an **event**. — combination of some possible worlds
- Any element $\omega \in \Omega$ is a **sample point / possible world / atomic event**.
- A **probability space** or **probability model** is a sample space with an assignment $P(\omega)\ \forall \omega \in \Omega$ such that $0 \leq P(\omega \leq 1)$ and $\sum_\omega P(\omega) = 1$. Accordingly, $P(A) = \sum_{\omega \in A} P(\omega)$.

* Prior probability:
Prob. of propositions correspond to belief prior to arrival of any new evidence.
└ e.g. P(Weather = sunny) = 0.72

e.g. odd(1) = true
↓
boolean

- A **random variable** is a function from sample points to some range (e.g. $X : \Omega \to \mathbb{R}$ or $X : \Omega \to$ Booleans). Moreover, random variables can be either discrete, when the codomain has a discrete number of elements, or continuous, when the codomain has an infinite number of elements.

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

$$P(X = x_i) = \sum_{\omega : X(\omega) = x_i} P(\omega) \tag{1.1}$$

Moreover, a probability distribution gives values for all possible assignments. For example, considering the random variable $Weather$ with possible values $\{sunny, rainy, cloudy, snowy\}$, $\mathbf{P}(Weather) = \{0.72, 0.1, 0.08, 0.1\}$ (which is normalized, i.e. sums to 1).

- The **joint probability distribution** for a set of random variables gives the probability of every atomic event on those random variables (i.e. every sample point). For example $\mathbf{P}(Weather, Cavity)$ is a 4×2 matrix of values:

/ discrete r.v | boolean r.v.

$Weather =$	$sunny$	$rainy$	$cloudy$	$snowy$
$Cavity = true$	0.144	0.02	0.016	0.02
$Cavity = false$	0.576	0.08	0.064	0.08

- The probability distribution of a continuous random variable is called **probability density function**. In particular, a function $p : \mathbb{R} \to \mathbb{R}$ is a probability density function (pdf) for X if it is a non-negative integrable function such that $\int_{Val(X)} p(x)dx = 1$.

- With respect to prior probabilities (belief prior to arrival of any new evidence) $P(X)$, **conditional probabilities** $P(X|Evidence)$ represent a more informed distribution in the light of the new $Evidence$. In particular:

\ new evidence

Bayes theorem: $P(a|b) = \frac{P(b|a)P(a)}{P(b)}$ ← product rule: $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$$P(a|b) = \frac{P(a \wedge b)}{P(b)} \text{ if } P(b) \neq 0 \quad *^2 \tag{1.2}$$

| given all I know, b. the probability of a is ... (implements beliefs)

1.1.2 Inference Using Full Joint Distributions

Full joint distribution allow the extrapolation of information. In particular, for any proposition ϕ:

$$P(\phi) = \sum_{\omega : \omega \models \phi} P(\omega) \tag{1.3}$$

For example, considering the following full join distribution:

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
Cavity = true — $cavity$	0.108	0.012	0.072	0.008
Cavity = false — $\neg cavity$	0.016	0.064	0.144	0.576

- Upper case (e.g. Cavity): r variable
- Lower case (e.g. cavity): value
¬cavity

*²: Chain rule:
$P(X_1, \ldots, X_m) = \prod_{i=1}^{m} P(X_i | X_1, \ldots, X_{i-1})$
$= P(x_1) P(x_2|x_1) P(x_3|x_1, x_2) \cdots P(x_m | x_1, \ldots, x_{m-1})$

$* = \alpha [P(Cavity, toothache, catch) + P(Cavity, toothache, \neg catch)]$
↳ hidden variable
$= \alpha [\langle 0.108, 0.016 \rangle + \langle 0.012, 0.064 \rangle]$
$= \alpha \langle 0.12, 0.08 \rangle$, $\alpha = 0.12 + 0.08 = 0.2$ (normalization factor to obtain probability mass of 1)
$= \langle 0.6, 0.4 \rangle$
↳ posterior distribution

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

Moreover, one can compute distributions such as:

(not the specific event, but the distribution, this is why there exists a denominator, α)
$$\mathbf{P}(Cavity | toothache) = \alpha \mathbf{P}(Cavity \wedge toothache) \quad *$$

where α is used as a normalization factor. This type of inference is called **inference by enumeration**. The general idea is to compute the distribution on a query variable by fixing evidence variables and summing over hidden variables. In particular, a **probability query** $\mathbf{P}(\mathbf{Y}|\mathbf{e})$ defines the posterior joint distribution of a set of query variables $\mathbf{Y}$ given specific values $\mathbf{e}$ for some evidence variables. Given a set of of query variables $\mathbf{Y}$, a set of evidence variables $\mathbf{E}$ and a set of hidden variables $\mathbf{H}$ one could answer a given query by summing out:

$$\mathbf{P}(\mathbf{Y}|\mathbf{E} = \mathbf{e}) = \alpha \mathbf{P}(\mathbf{Y}, \mathbf{E} = \mathbf{e}) = \alpha \sum_{\mathbf{h}} \mathbf{P}(\mathbf{Y}, \mathbf{E} = \mathbf{e}, \mathbf{H} = \mathbf{h}) \tag{1.4}$$

Time and space complexity of this approach are $\mathcal{O}(d^n)$, where d is the largest arity.

exponential, non-tractable

Chapter 2

Bayesian Network Representation

Given two random variables, A and B:

- These variables are said to be **marginal independent**, denoted $\mathbf{P} \models (A \perp B)$, if and only if $\mathbf{P}(A|B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(A, B) = \mathbf{P}(A)\mathbf{P}(B)$.
- These variables are said to be **conditionally independent**, denoted $\mathbf{P} \models (A \perp B|C)$, if and only if, considering a third variable C, $\mathbf{P}(A|B, C) = \mathbf{P}(A|C)$.

2.1 Bayes Rule

- Given two events, a and b:

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)} \tag{2.1}$$

Alternatively, in distribution form:

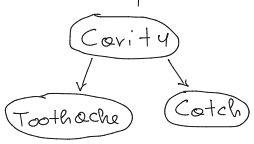
$$\mathbf{P}(A|B) = \frac{\mathbf{P}(B|A)\mathbf{P}(A)}{\mathbf{P}(B)} \tag{2.2}$$

This is especially useful for assessing diagnostic probability from causal probability:

$$P(cause|effect) = \frac{P(effect|cause)P(cause)}{P(effect)}$$

- The Bayes rule and the concept of conditional independence are related. Considering α as a normalization factor:

$$\begin{aligned}\mathbf{P}(Cavity|toothache \wedge catch) &= \alpha\mathbf{P}(toothache \wedge catch|Cavity)\mathbf{P}(Cavity) \\ &= \alpha\mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)\mathbf{P}(Cavity)\end{aligned}$$



5

- Naive Bayes model are used to solve classification tasks. In particular:

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i | Cause) \tag{2.3}$$

2.2 Bayesian Networks

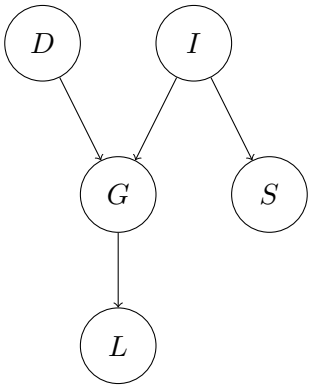
Bayesian networks are a simple, graphical notation (acyclic and directed graph) for conditional independence assertions and for compact specification of full joint distributions. In particular:

- There exist a ~~set of nodes per variable~~ node per variable.
- Connections between nodes means "directly influences".
- There exist a conditional distribution for each node given its parents, i.e. $\mathbf{P}(X_i | Parents(X_i))$. This distribution is represented as a conditional probability table (CPT). In particular, a CPT for boolean X_i with k boolean parents requires $\mathcal{O}(n \cdot 2^k)$ numbers instead of the $\mathcal{O}(2^n)$ for the full joint distribution.

Given a Bayesian network:

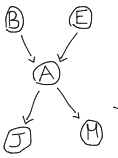
- The main reasoning patterns are the following:
  - **Causal reasoning**, used to make prediction.
  - **Evidential reasoning**, used to give explanation.
  - **Intercausal reasoning**, used to infer reasons behind a conclusion.

  For example, if a student's grade depends on intelligence and on the difficulty of the course, SAT scores are correlated with intelligence and the professor writes recommendation letters by only looking at grades, the corresponding Bayesian network would be the following:



P(Letter = strong), which is computed based on evidence

In this case an example of causal reasoning would be the query "will a student get a strong reference letter?", an example of evidential reasoning would be the query "is a students a good potential recruit?" and an example of intercausal reasoning would be "why did a student score low/high?".



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

- The **global semantics** of the network defines the full joint distribution as the product of the local conditional distributions:

product rule — this can be applied thanks to independence

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | \text{ parents}(X_i)) \tag{2.4}$$

- It is possible to define a **flow of probabilistic influence**. In particular, given the variables X, Y, Z:
  - $X \to Y$ is said to be a **direct cause**.
  - $X \leftarrow Y$ is said to be a **direct effect**.
  - $X \to Z \to Y$ is said to be a **causal trail**.
  - $X \leftarrow Z \leftarrow Y$ is said to be a **evidential trail**.
  - $X \leftarrow Z \to Y$ is said to be a **common cause**.
  - $X \to Z \leftarrow Y$ is said to be a **common effect**.

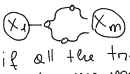
active two-edge trail

Moreover, if the influence can flow from X to Y via Z, the trail $X \leftrightharpoons Z \leftrightharpoons Y$ is said to be **active**. More specifically, let $X_1 \leftrightharpoons \cdots \leftrightharpoons X_n$ be a trail and $\mathbf{Z}$ be a subset of observed variables (i.e. evidence). The trail $X_{i-1} \leftrightharpoons X_i \leftrightharpoons X_{i+1}$ is active trail given $\mathbf{Z}$ if:

- $\forall X_{i-1} \to X_i \leftarrow X_{i+1}$, X_i or one of its descendants are in $\mathbf{Z}$.
- No other node along the trail is in $\mathbf{Z}$.

evidence

trails



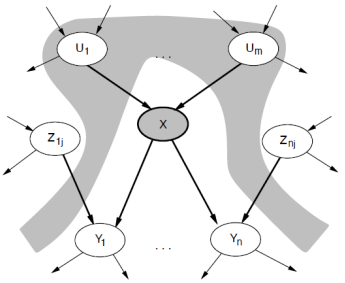
e.g. if all the trails from X_1 to X_m are not active, then X_1 is independent from X_m

Moreover, two sets of nodes $\mathbf{X}$, $\mathbf{Y}$ are **d-separated** given $\mathbf{Z}$ if there is no active trail between any $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ given $\mathbf{Z}$.

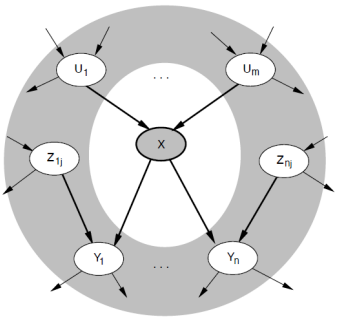
- It is possible to define also a **local semantics** which allows to infer that each node is conditionally independent of its non-descendants given its parent (i.e. ~~it~~ if its parent are evidence):

Theorem: local semantics ⇕ global semantics

$\left(P(x_1, \ldots, x_m) = \prod_{i=1}^{m} P(x_i | \text{Parents}(X_i))\right)$



- Each node is conditionally independent of all the others given its **Markov blanket**, namely its parents, its children and its children's parents:



In order to construct a Bayesian network one needs a method such that a series of locally testable assertions of conditional independence guarantees the required global semantics:

1. Choose an ordering of variables $X_1, \ldots, X_n$. This ordering is usually chosen by following a causality intuition.
2. For i to n:
   (a) Add X_i to the network.
   (b) Select the parents from $X_1, \ldots, X_{i-1}$ such that $\mathbf{P}(X_i|\,\text{parents}(X_i)) = \mathbf{P}(X_i|X_1, \ldots, X_{i-1})$.

   This choice of parents guarantees the global semantics.

In order to maintain a small number of network's parameters one must:

- Follow a causal ordering for the variables of the network (for each ordering there exists a different Bayesian network).
- Introduce hidden states which ~~reduce sparsity (the number of CPT's entries grows exponentially with the number of parents)~~.
- Use canonical distributions. An example of such distributions are the **Noisy-OR** distributions, which model multiple non-interacting causes. In particular, given the parents $U_1, \ldots, U_k$, which include all causes, and making the assumption that the failure probability for each cause alone is q_i, then:

$$P(X|\overbrace{U_1, \ldots, U_j}^{\text{causes}}, \neg U_{j+1}, \ldots, \neg U_k) = 1 - \prod_{i=1}^{j} q_i \tag{2.5}$$

For example, considering as parents the boolean variables $Cold$, Flu and $Malaria$, starting from some medical knowledge it is possible to populate the following CPT:

$Cold$	Flu	$Malaria$	$P(Fever)$	$P(\neg Fever)$
F	F	F	0.0	
F	F	T		0.1
F	T	F		0.2
F	T	T		
T	F	F		0.6
T	F	T		
T	T	F		
T	T	T		

-> in 10% of the cases, malaria does not cause fever

Every other entry of the CPT is obtained by the given medical knowledge:

$Cold$	Flu	$Malaria$	$P(Fever)$	$P(\neg Fever)$
F	F	F	0.0	1.0
F	F	T	0.9	0.1
F	T	F	0.8	0.2
F	T	T	0.98	$0.02 = 0.2 \cdot 0.1$
T	F	F	0.4	0.6
T	F	T	0.94	$0.06 = 0.6 \cdot 0.1$
T	T	F	0.88	$0.12 = 0.6 \cdot 0.2$
T	T	T	0.988	$0.012 = 0.6 \cdot 0.2 \cdot 0.1$

By using this particular distribution, the CPT needs only to be populated with k entries, where k is the number of causes (i.e. parents).

Chapter 3

Exact Inference

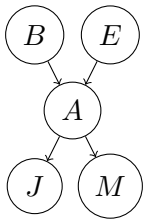
The possible inference tasks carried out using Bayesian networks are the following:

- Simple queries, e.g. $\mathbf{P}(X_i|\mathbf{E} = \mathbf{e})$. In the rest of this chapter only this type of query will be considered.
- Conjunctive queries, e.g. $\mathbf{P}(X_i, X_j|\mathbf{E} = \mathbf{e}) = \mathbf{P}(X_i|\mathbf{E} = \mathbf{e})\mathbf{P}(X_j|\mathbf{E} = \mathbf{e})$.
- Taking optimal decisions (decision making).
- Selecting which evidence to seek next (value of information).
- Selecting the most critical probabilities (sensitivity analysis).
- Explaining something (explanation).
- Finding the most probable values for a variable given the evidence (map-query).

In particular, an exact inference allows to obtain a specific value as the result of a specific query.

3.1 Inference by Enumeration

Considering the following Bayesian network:

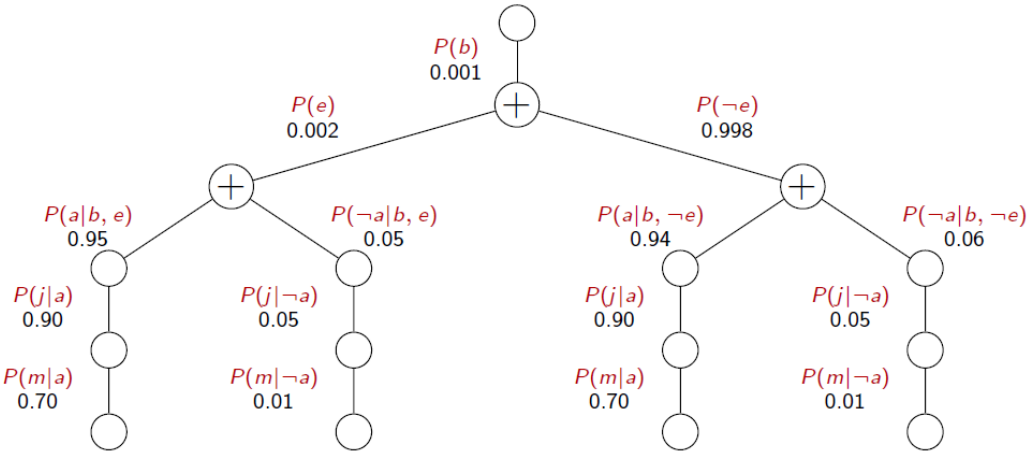


then a simple query would be $\mathbf{P}(B|j, m)$:

$$
\begin{aligned}
\mathbf{P}(B|j,m) &= \frac{\mathbf{P}(B,j,m)}{P(j,m)} \\
&= \alpha \mathbf{P}(B,j,m) \\
&= \alpha \sum_e \sum_a \mathbf{P}(B,e,a,j,m) \\
&= \alpha \sum_e \sum_a \mathbf{P}(B)P(e)\mathbf{P}(a|B,e)P(j|a)P(m|a) \\
&= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e)P(j|a)P(m|a)
\end{aligned}
$$

sum over hidden variables

This process produces an evaluation tree which contains repeated computation and, thus, redundancy:



3.2 Inference by Variable Elimination

dynamic programming (storing partial results)

In order to avoid repeated computation, one can carry out summations right-to-left, storing intermediate results (**factors**). In this case, the factors are stored into tables. For example, given the following query:

$$
\mathbf{P}(B|j,m) = \alpha \underbrace{\mathbf{P}(B)}_{B} \sum_e \underbrace{P(e)}_{E} \sum_a \underbrace{\mathbf{P}(a|B,e)}_{A} \underbrace{P(j|a)}_{J} \underbrace{P(m|a)}_{M}
$$

→ factors

and given the following factors:

$\mathbf{P}(B)$
0.001

$\mathbf{P}(E)$
0.002

P(A|B,E)

B	E	$\mathbf{P}(B)$
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	$\mathbf{P}(J\|A)$
T	0.90
F	0.05

A	$\mathbf{P}(M\|A)$
T	0.70
F	0.01

then it possible to re-write the previous expression by applying factor products and by summing out variables:

$$\begin{aligned}
\mathbf{P}(B|j,m) &= \alpha \underbrace{\mathbf{P}(B)}_{B} \sum_e \underbrace{P(e)}_{E} \sum_a \underbrace{\mathbf{P}(a|B,e)}_{A} \underbrace{P(j|a)}_{J} \underbrace{P(m|a)}_{M} \\
&= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e) P(j|a) f_M(a) \qquad \hookrightarrow \text{this values are found by using the stored partial results} \\
&= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e) f_J(a) f_M(a) \\
&= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a f_A(a,b,e) f_J(a) f_M(a) \quad - \text{(sum out A)} \\
&= \alpha \mathbf{P}(B) \sum_e P(e) f_{\bar{A}JM}(b,e) \quad - \text{(sum out E)} \\
&= \alpha \mathbf{P}(B) f_{\bar{E}\bar{A}JM}(b) \\
&= \alpha f_B(b) \times f_{\bar{E}\bar{A}JM}(b)
\end{aligned}$$

A has been eliminated

In order to reimmate $\sum_a$ and $\sum_e$ one updates the previously stored tables

3.3 Irrelevant Variables

Given a Bayesian network and a specific query, a variable Y is said to be **irrelevant** unless $Y \in$ ancestors($\{\mathrm{X}\} \cup \mathbf{E}$), where $\mathbf{E}$ are the evidences. In other words, Y is irrelevant if it is d-separated from X by $\mathbf{E}$. For example, considering the previous Bayesian network and the query $\mathbf{P}(J|b)$:

$$\mathbf{P}(J|b) = \alpha P(b) \sum_e P(e) \sum_a P(a|b,e) \mathbf{P}(J|a) \sum_m P(m|a)$$

it is possible to notice that $\sum_m P(m|a) = 1$ by definition. In this case, the variable M is irrelevant to the query.

* Approximate inference, differently from exact inference, is based on sampling.

Chapter 4

usually used when dealing with Bayesian networks

* Approximate Inference

especially when there exist multiple paths between variables

Approximate inference can be applied whenever the given network is too large to apply methods such as variable elimination.

4.1 Inference by Stochastic Simulation

The basic idea of this method, starting from an empty network, is the following:

1. Draw N samples from a sampling distribution S. - usually a continuous distribution between zero and one
2. Compute an approximate posterior probability $\hat{P}$. For example, if $\hat{N}$ is the number of samples related to some specific event, the probability of this event will be $\hat{P} = \hat{N}/N$.
3. Show that $\hat{P}$ converges to the true probability P if $N \to \infty$.

On the other hand, if one has prior knowledge (i.e. evidence), $\hat{\mathbf{P}}(X|\mathbf{e})$ is estimated only from samples agreeing with the given evidence $\mathbf{e}$, and not from all the N samples.

contrary from the empty network

rejection sampling

- Likelihood weighting
- Markov chain Monte Carlo