Alma Mater Studiorum · Università di Bologna

Second Cycle Degree
Artificial Intelligence

Fundamentals of Artificial Intelligence and Knowledge Rapresentation

Student:
Matteo Canghiari

Academic Year 2025/2026

Chapter 1

Acting under uncertainty

1.1 Basic probability notation

Every agent based on **decision theory** needs a formal language to use and represent probabilistic informations. Typically AI needs a more suited and consistent approach than the traditional probability theory. This section includes all the necessary definitions and examples to understand the subsequent arguments in depth.

Definition

The set of all possible worlds is called the **sample space**, denoted Ω. Any subset $A \subseteq \Omega$ is an **event**. Any element $\omega \in \Omega$ is called **sample point**.

Definition

A **probability space** is a sample space with an assignment $P(\omega)$ for every $\omega \in \Omega$ where:

- $0 \leq P(\omega) \leq 1$
- $\sum P(w) = 1$ for every $\omega \in \Omega$

Definition

A **random variable** is a function from sample points to some range, e.g., the reals or Booleans.

e.g. *Odd(1) = true*

Definition

$\boldsymbol{P}$ induces a **probability distribution** for any random variable $\boldsymbol{X}$:

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

A **probability distribution** gives values for all possible assignment.

Definition

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

e.g. *P(Cavity = True) = 0.1*

Definition

The **Joint Probability Distribution** for a set of random variables gives the probability of every sample point on those random variables.

e.g. $\boldsymbol{P}$*(Weather, Cavity) = a* 2×4 *matrix of values:*

Weather =	*sunny*	*rain*	*cloudy*	*snow*
Cavity = True	0.144	0.02	0.016	0.02
Cavity = False	0.576	0.08	0.064	0.08

Table 1.1: Probability distribution of the Weather random variable

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

Definition

A function $p : R \to R$ is a **probability density function** (**pdf**) for X if it is a nonnegative integrable function s.t.

$$\int_{Val(X)} p(x)dx = 1$$

Definition

Conditional or **posterior probabilities** $P(X|Evidence)$ represent a more informed distribution in the light of new **evidence**.

e.g. $P(cavity|toothache) = 0.8$

It does not mean *"if I have toothache then there is* 80% *of chance that there is also a cavity"*, instead the evidence mean *"given toothache evidence is all I know"*.

The typically definition of conditional or posterior probability is:

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

Otherwise, numerator can be written by the **product rule**:

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

The product rule at the same time is applied to whole distributions, not only for single values as done previously.

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

1.2 Inference using full joint distribution

This paragraph describes a new method to retrieve informations from data, named **probabilistic inference**. It allows the computation of conditional probabilities for query propositions by given evidence. Starting from an example is defined the **full joint distribution** as the knowledge base from which answers to all questions.

Example

e.g. (*Toothache, Cavity, Catch*) is just a domain consisting of three Boolean variables. *Catch* condition occurs when the dentist's steel probe catches in the tooth. Based on the domain, the **full joint distribution** seems like this:

	toothache		$\neg$*toothache*	
	catch	$\neg$*catch*	*catch*	$\neg$*catch*
cavity	0.108	0.012	0.072	0.008
$\neg$*cavity*	0.016	0.064	0.144	0.576

Table 1.2: Full joint distribution of Toothache, Cavity and Catch

The equation

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

gives a direct way to calculate probabilities of any assertions, summing up all the possible worlds that satisfy the original proposition.

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

e.g. $P(cavity \vee toothache) = 0.108+0.012+0.072+0.008+0.016+0.064 = 0.28$
It's also possible compute conditional probabilities:
e.g. $P(\neg cavity|toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)} = \frac{0.016+0.064}{0.2} = 0.4$
Notice that in this calculation the term $P(toothache)$ remains constant, no matter which value of *Cavity* is computed. In fact, it can be viewed as a **normalization constant** (α) for the whole distribution $\mathbf{P}(Cavity|toothache)$, ensuring that the positive and negative case sum up to one, as the second probability axiom requires.

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

Definition

The first probability calculated $P(toothache)$ is called **marginalization**, or more simply **summing out**, because it sums up the probabilities for each possible value of the other variables.

Definition

The second one $P(\neg cavity|toothache)$ is named **conditioning**, a variant of marginalization that involves conditional probabilities instead of joint probabilities.

Definition

From the example, it's possible to extract a general inference procedure. Let $\mathbf{Y}$ be the query variables. Let $\mathbf{E}$ be the list of evidence variables, let $\mathbf{e}$ be the list of observed values for them, and let $\mathbf{H}$ be the unobserved variables. The **probability query** $\mathbf{P}(Y|\mathbf{e})$ defines the posterior joint distribution of a set of **query variables** $\mathbf{Y}$ given specific values $\mathbf{e}$ for some **evidence variables** $\mathbf{E}$:

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

The full joint distribution can answer probabilistic queries for discrete variables, but only for small domains. It does not scale well: for a domain described by n Boolean variables, it requires an input table of size $O(2^n)$ and takes $O(2^n)$ time to process a question. The full joint distribution in tabular form is just not a practical tool for building reasoning systems.

Chapter 2

Independence

If we expand the full joint distribution defined in Figure 1.2 by adding a new random variable, *Weather*, it becomes $\mathbf{P}(Weather, Toothache, Cavity, Catch)$, which has $2 \times 2 \times 2 \times 4 = 32$ entries. But, what is the relationship between these four random variables? For instance, are the $P(cloudy, toothache, cavity, catch)$ and $P(toothache, cavity, catch)$ related? This last question can be expressed in probabilistic terms as:

$$P(cloudy, toothache, catch, cavity) = \\ P(cloudy|tootchache, cavity, catch)P(toothache, cavity, catch)$$

At the same time, we can imagine that *Toothache, Cavity, Catch* should be independent from *Weather*. Therefore, the following assertion seems reasonable:

$$P(cloudy|toothache, catch, cavity) = P(cloudy)$$

From this, we can deduce:

$$P(cloudy, toothache, catch, cavity) = P(cloudy)P(toothache, catch, cavity)$$

Or generally:

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

Thus, the initial 32 entries table can be divided from one 8-entries table and one 4-entries table. The property used in the previously equation is called **independence**.

First of all are introduced some basic definitions and examples to understand the effectiveness of independence.

Definition

A and B are **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 $\mathbf{P}(A|B) = \mathbf{P}(A)\mathbf{P}(B)$

When they are available, independence assertions can help in reducing the size of the domain representation and the complexity of the inference problem. Unfortunately, clean separation of entire sets of variables by independence are quite rare. Moreover, even the independence subset can be quite large, for instance, dentistry might involve dozens of diseases and symptoms, all of which are associated. To handle such problems, we need more specific methods than the general concept of independence, one of them is named **conditional independence**. Let see an example of conditional independence.

Example

i.e. given $\mathbf{P}(Toothache, Cavity, Catch)$ has $2^3 - 1 = 7$ independent entries [a] . If I have a cavity, the probability that the probe catches in it does not depend on whether I have toothache:

$$P(catch|toothache, cavity) = P(catch|cavity)$$

The same independence hold if I haven't got a cavity:

$$P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$$

Catch is **conditional indipendent** of Toothache given Cavity [b] .

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

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

Using the chain rule, the full joint distribution becomes:

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

$2 + 2 + 1 = 5$ indipendent numbers, we have less entries than before.

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

[a] *Why 7 independent entries and not 8 as before? Simply, if we know 7 of them the 8th is automatically determined, must be the last value remaining.*

[b] *This introduces the meaning of the flow of influence.*

2.1 Bayes' rule

The Section 1.1 defined the **product rule**. It can be written in two forms:

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

Combining the right-hand side of each equation and dividing by $P(a)$, we get the **Bayes' rule**.

Bayes' theorem

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

or in distribution form:

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

This turns out to be very useful for assessing **diagnostic** probability from **causal** probability.

2.2 Bayes' rule: a simple case study

On the surface, Bayes' rule does not seem very useful. It allows to compute the single term $P(b|a)$ in terms of three items: $P(a|b)$, $P(b)$ and $P(a)$. But the Bayes' rule is useful in practice because there are many cases where we have probabilities for these three items and need to compute the fourth. Often, we perceive as evidence the **effect** of some unknown **cause** and we would like to solve for that cause. In that case, the Bayes' rules becomes:

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

The conditional probability $P(effect|cause)$ defines the relationship in the **causal** direction, while $P(cause|effect)$ describes the **diagnostic** direction. Let see an example.

Example

Say 1 individual in 50.000 suffers from meningitis, 1% from a stiff neck, and 70% of the times meningitis causes a stiff neck. *What is the probability that an individual with a stiff neck has meningitis?*

$P(s|m) = 0.7$
$P(m) = 1/50.000$
$P(s) = 0.01$

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.7 \times (1/50.000)}{0.01} = 0.0014$$

We have seen that the Bayes' rule seems useful for answering probabilistic queries conditionated on one piece of evidence. But, what happens when we have two or more pieces of evidence? For instance, what a dentist conclude if her steel probe cathes in the tooth of a patient?

Example

i.e. If we know the full joint distribution 1.2, we can define the answer as:

$\mathbf{P}(Cavity|toothache \wedge catch) = \alpha\langle 0.108, 0.016\rangle = \langle 0.871, 0.129\rangle$

However, this approach does not scale up to larger number of variables. We can try using the Bayes' rule to reformulate the problem:

$\mathbf{P}(Cavity|toothache \wedge catch) = \alpha\mathbf{P}(toothache \wedge catch|Cavity)\mathbf{P}(Cavity)$

For this reformulation, we must know the conditional probabilities of the conjuction for each value of Cavity. That might be simple for just two variables, but again it does not scale up. Thus, we need to find some assertions about the domain that will enable us to simplify the expressions.

The notion of **independence** provides a clue. It would be nice if Toothache and Catch were independent, but they aren't: if the probe catches in the tooth, then it is likely that the tooth has a cavity and that cavity causes the tootchache. By this last assertion, we can allude that these variables are independent, given the presence or the abscence of a cavity. Each effects is directly caused by the cavity, but neither has a direct effect on the other. Mathematically, this property is written as follows:

$$\mathbf{P}(toothache \wedge catch|Cavity) = \mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)$$

This equation introduces the meaning of **conditional independence**: *toothache* is conditionally independent from *catch* given *Cavity*. Now the information requirements are the same as for inference, using each piece of evidence separately: the prior probability $\mathbf{P}(Cavity)$ for the query variable and the conditional probability for each effect, given its cause.

$$\mathbf{P}(toothache \wedge catch|Cavity) =$$
$$\alpha\mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)\mathbf{P}(Cavity)$$

Definition

The **conditional independence** of two variables X and Y, given a third variable Z, is:

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

In this way, the original table is decomposed into three small tables. The table 1.2 has seven independent entries. The smaller tables contain five indipendent numbers, 2 for the conditional probability distributions and 1 for the prior distribution $\mathbf{P}(Cavity)$. Right now the size of the representation grows as $O(n)$ instead of $O(2^n)$, it grows by a **linear** pace not anymore by a **exponential** pace. Finally, we can say that conditional independence and absolute independence can allow probabilistic systems to scale up.

Example

i.e. Conceptually, Cavity **separates** Toothache and Catch because it is a direct cause of both of them.

Definition

The full joint distribution can be written as:

$\mathbf{P}(Cause, Effect_1, Effect_2, ..., Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i|Cause)$

This probability distribution is called **Naive Bayes** [a] .

[a] The naive Bayes model is the most common way to solve labeling tasks, such as classification. The total number of parameters grows **linearly**.

Chapter 3

Bayesian networks

The previous chapter noted the importance of absolute and conditional independence relationships in simplifying probabilistic representation. This section introduces a systematic way to represent such relationships in the form of **Bayesian networks**. We define the syntax and semantics of these networks and show how they can be used to capture uncertain knowledge.

A Bayesian network is a simple graphical notation for conditional independence assertions and hence for a compact specification of full joint distribution. The Bayesian network's syntax is composed by:

1. Each node corresponds to a random variable.
2. A set of directed links or arrows connects pairs of nodes.
3. Each node X_i has a conditional probability distribution $\mathbf{P}(X_i|Parents(X_i))$, that quantifies the effect of the parents on the node.

Example

i.e. Topology of network encodes conditional independence assertions:

- Weather is independent of the other variables [a] .
- Toothache and Catch are conditionally independent given Cavity [b] .

[a] Formally, the absolute or conditional independence is indicated by the absence of a link between nodes.

[b] The intuitive meaning of an arrow is typically that X has a direct influnce on Y, which suggests that causes should be parents of effects.

Example

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

The random variables are: *Burglar, Earthquake, Alarm, MaryCalls, JohnCalls.*
[a] [b] [c]

[a]The network topology reflects **causal** knowledge, from the causes nodes we define the effects nodes.

[b]For each node the conditional distribution are shown as a **conditional probability table**, or simply CPT.

[c]Let's take a look at the tables. In this network we are talking about joint distribution, not full joint distribution. Simply, the full joint distribution about boolean random variables can be computed by $1 - P(a)$.

3.1 Reasoning patterns

We begin the discussion with a simple toy example, the *student network*.

Example

i.e. A student's grade depends on intelligence and on the difficulty of the course. SAT scores are correlated with intelligence. A professor writes recommendation letters by only looking at grades.

In this case, our probability space is composed by five relevant random variables *Difficulty (D)*, *Intelligence (I)*, *SAT score (S)*, *Grade (G)* and *Letter (L)*.

Consider a particular student, George, that he would like to reason using the student network. We might ask how likely George is to get a strong recommandation from his professor in Analysis. Knowing nothing else about George and his grade, this probability is around the 50 percent.

We now find out that George is not so intelligent. The probability that he gets a strong letter from the professor goes down to 39. We now further discover that Analysis is an easy class. The probability that George receive a strong letter is now about 51 percent.

Queries and answers such as these, where we predict the behavior of numerous factors, is called **causal reasoning** or **prediction** [a] .

[a]It reflects the causal direction, from the parent nodes we are just defining which is the influence on their children.