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

Introduction to uncertainty and probabilistic reasoning

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

P induces a **probability distribution** for any random variable ***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. ***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

Bayesian network representation

2.1 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:

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

$2 + 2 + 1 = 5$ 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.*

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.

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**.

2.2 Bayesian network representation

The previous paragraph 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 $1 - P(a)$.

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 recommendation 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 this, where we predict the behaviors from causes to effects, are called **causal reasoning** or **prediction** [a] .

Now, assume a recruiter, trying to hire George based on our previous model. By a prior probability, the recruiter believes that George is 30 percent likely to be intelligent. He obtains George's grade record for the class Analysis and sees that George received a low score in that class. His probability that George has high intelligence goes down, about 7.9 percent. We note that the probability that the class is a difficult one also goes up, from 40 percent to 62.9 percent. Now, consider that the recruiter lost George's trascript of records, and has only the recommendation letter from George's professor in Analysis, which is weak. The probability that George has high intelligence still goes down, but only to 14 percent. Note that if the recruiter has both the grade and the letter, we have the same probability as if he had only the grade.

Queries such as this, where we reason from effects to causes, are named **evidential reasoning** or **explanation** [b] .

Finally, George submits his high SAT score to the recruiter. The probability that George has high intelligence goes up dramatically, from 7.9 percent to 57.8 percent. Intuitively, the reason that the SAT score outweights the poor grade is that a student with low intelligence are unlikely to get good scores on their SAT, whereas students with high intelligence can still get C grades in hard class. Indeed, we see that the probability that Analysis is a difficult one goes up from 62.9 percent to 76 percent.

This last pattern is a interesting one. The information about SAT score told us other informations about the student's intelligence, which, in conjuction with the student's grade, gave us some clues about the difficulty of the course. Let examine this pattern in probabilistic terms.

We are saying

$P_s(i^1|g^3) = 0.079$

On the other hand, if we consider that Analysis is a hard class, we have

$P_s(i^1|d^1, g^3) = 0.11$

Here, we are partially explaining why George has got a poor grade. By the way, taking a more tricky example, for instance George got a middle grade in Analysis, we have that

$P_s(i^1|g^2) = 0.175$

Also if Analysis is a hard class, we get

$P_s(i^1|d^1, g^2) = 0.34$

In effect, we have justified the poor grade via the difficulty of the class. Explaining away is an instance of a general pattern called **intercasual reasoning**, where different causes of the same effect, so they are parents of the effect node, can interact.

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

[b] In this case, we are not reasoning from top to bottom, but instead from bottom to top.

Compactness

A Bayesian network can often be more *compact* than the full joint distribution. The **compactness** of a Bayesian network is a property that makes easy to handle domain with many random variables.

At the same time, a Bayesian network grows *linearly*, instead of an *exponential* growth by full joint distribution. Assuming a domain composed by **n** Boolean variables, where each of them is associated with a CPT *(Conditional Probability Table)* [-9]. If each variable has no more then **k** parents, the complete network requires $O(n \times 2^k)$ numbers, as we already know to complete the full joint distribution are necessary $O(2^n)$.

Example

i.e. Comparison of parameters required from the previously example between Bayesian network and full joint distribution.

For the burglar network:

$1 + 1 + 4 + 2 + 2 = 10$ numbers required by the Bayesian network

$2^5 - 1 = 47$ numbers required by the full joint distribution

Global semantics

A Bayesian network is a directed acyclic graph with some numeric parameters attached to each node. One way to define what the network means is to define the way in which it rapresents a full joint distribution.

[-9] A CPT for a Boolean variable $\mathbf{X_i}$ with **k** parents, has $\mathbf{2^k}$ rows for the combinations of parents values and usually is defined only the True case; the negative case, however $X_i = False$, can be simply computed by the third probability axiom $\mathbf{1 - p}$.

Definition

Global semantics defines the full joint distribution as the product of the local conditional distributions:

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

Example

i.e.
$P(j, m, a, \neg b, \neg e)$
$= P(j|a)P(m|a)P(a|\neg b \wedge \neg e)P(\neg b)P(\neg e)$
$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 = 0.000628$

Flow of influence

Until now, we used the intuition that edges represent direct dependence. For instance, we said that the letter recommendation from the professor depends only on the student's grade; this state was encoded by the fact that there is an exit edge from G that arrive to L. This intuition is **not** always true.

The aim of this section is to understand when we can guarantee independence between random variables. First of all, we begin with a simple case analysis: we try to understand when a variable **X** can influence **Y** given **Z**.

Direct connection. This is the simple case, when X and Y are directly connected via an edge. If X and Y are directly connected, we can always get examples where they influence each other, regardless of **Z**.

Indirect connection. Now we are considering the more complicated case when X and Y are not directly connected, but there is a trail between them. There are four cases where X and Y are connected via Z.

The first two correspond to causal chains, the third to a common cause, and the last one to a common effect.

Causal trail. We have a chain like $X \rightarrow Z \rightarrow Y$. X cannot influence Y via Z if Z is observed.

Effect trail. Now we have the same trail but in the opposite direction, so $Y \rightarrow Z \rightarrow X$. Another time, Y cannot influence X via Z if Z is observed.

Common cause. This type of trail defines the same grade of independence as before. If Z is observed then neither X or Y can influence each other.

Common effect. In the previously cases, we see a common pattern: if Z is observed then neither X or Y can influence each other. By the way, this kind of trail, $X \rightarrow Z \leftarrow Y$, has a new behavior. If Z is not observed influence cannot flow along the trail. So if Z is observed X and Y are independent. In the student example we analyzed this case, which we called *intercausal reasoning*; we showed that the probability that student has high intelligence goes down when we observe that his grade is a poor score, but then goes up when we observe that the class is a hard one. Let us consider a variant of the same case. Assume that we do not observe the student's grade, but we observed that he received an awful recommendation letter. Intuitively, the same phenomenon happens. The weak letter told us that he received a low grade, and it is sufficient to correlate Intelligence and Difficulty.

Definition

If influence can flow from X to Y via Z, the trail $X \longleftrightarrow Z \longleftrightarrow Y$ is **active**.

If we consider a longer trail $X_1 \rightarrow \cdots \rightarrow X_n$, the first variable X_1 can influence the last variable X_n, if influence can flow through every single node of the trail. This will be true if and only if every two-edge trail $X_{i-1} \longleftrightarrow X_i \longleftrightarrow X_{i+1}$ along the trail allows influence to flow.

Definition

Let **Z** be a subset of observed variables. The trail $X_{i-1} \longleftrightarrow X_i \longleftrightarrow X_{i+1}$ is **active given** Z if

- $\forall X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, X_i or one of its descendants are in Z [a]
- no other node along the trail is in Z

[a] The trail reported is the common effect case.

However, inside the Bayesian network literacy there is another important notion, named **d-separation**.

Definition

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

It's possible to summarize few steps to follow to determine if X and Y are independent given Z:

1. Mark all nodes in Z or having descendants in Z.

2. Traverse the graph from X to Y, stopping if we get to a **blocked** node [-9].
3. If we can't reach Y, then X and Y are independent.

Another aspects about independence are introduces by the meaning of **local semantics** and **Markov blanket**.

Definition

Local semantics define that each node is conditionally independent of its **non-descendants** given its parents [a] .

[a] Here, we are considering the parents of the first node, not the parents of non-descendants.

Definition

Each node is conditionlly independent of all the others nodes given its **Markov blanket**: so its parents, children and children's parents.

[-9] A node is blocked if that node is the middle of an unmarked v-structure *(common effect case)*, or belongs to Z *(cannot be both)*.

Chapter 3

Constructing Bayesian networks