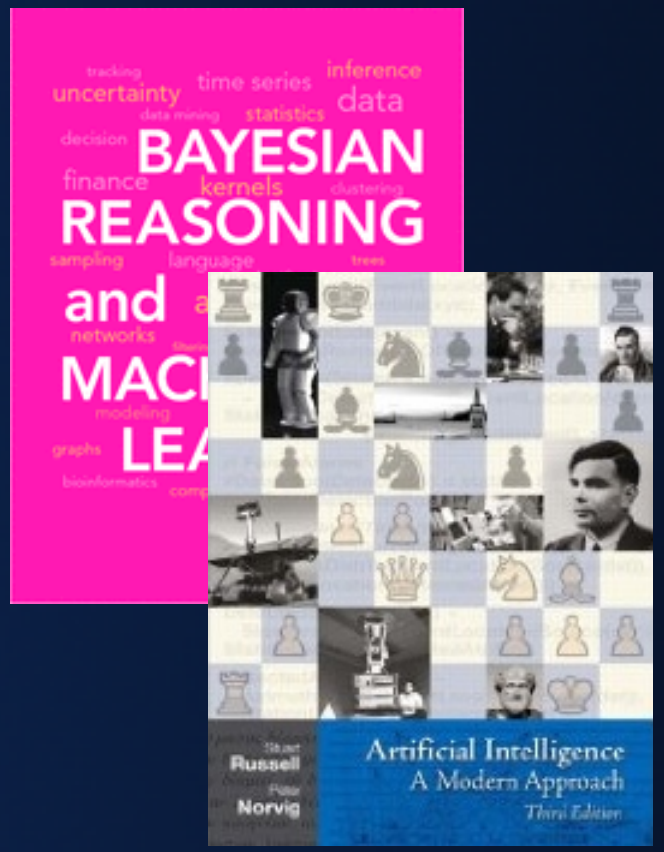


AI Fundamentals: Uncertain reasoning

Maria Simi

IN SUPREMÆ DIGNITATIS
1343

Quantifying uncertainty (AIMA chapter 13)

LESSON 1: INTRODUCTION – BASIC PROBABILITY NOTATION – INFERENCE WITH FULL JOINT DISTRIBUTION – BAYES RULE – INDEPENDENCE

Acting under uncertainty

Agents are inevitably forced to reason and make decisions based on incomplete information. They need a way to handle uncertainty deriving from:

1. partial observability (uncertainty in sensors)
2. nondeterministic actions (uncertainty in actions)

A partial answer would be to consider, instead of a single world, **a set of possible worlds** (those that the agent considers possible – *a belief set*) but planning by anticipating all the possible contingencies can be really complex.

Moreover, if no plan guarantees the achievement of the goal, but still the agents needs to act, how can he evaluate the relative merits of alternative plans?

Probability theory offers a clean way to **quantify uncertainty** (common sense reduced to calculus).

Motivating example 1

Suppose the goal for a taxi-driver agent is "*delivering a passenger to the airport on time for the flight*"

Consider action A_t = leave for airport t minutes before flight.
How can we be sure that A_{90} will succeed?

There are many sources of uncertainty:

1. partial observability or noisy sensors: road state, other drivers' plans, police control, inaccurate traffic reports etc.
2. uncertainty in action outcomes (flat tire, car problems, bad weather etc.)

With a logic approach it is difficult to anticipate **everything that can go wrong** (*qualification problem*). A_{90} may be the **most rational action**, given that the airport is 5 miles away and you want to avoid long waits at the airport and still catch the flight.

Decision theory

The rational decision depends on both the **relative importance** of various goals and the **likelihood** that, and degree to which, they will be achieved.

When there are conflicting goals the agent may express **preferences** among them by means of a **utility function**.

Utilities are combined with probabilities in the general theory of rational decisions called **decision theory**:

Decision theory = probability theory + utility theory

An agent is rational if and only if it chooses *the action that yields the maximum expected utility, averaged over all the possible outcomes of the action*.

This is called the principle of **Maximum Expected Utility** (**MEU**).

Motivating example 2

A medical diagnosis example. Given the symptoms (*toothache*) infer the cause (*cavity*).

How to encode this relation in logic?

- $Toothache \Rightarrow Cavity$ (diagnostic rule)
  $Toothache \Rightarrow Cavity \vee GumProblem \vee Abscess \ldots$ there are many possible causes
- $Cavity \Rightarrow Toothache$ (causal rule)
- $Cavity \wedge C_1 \ldots \wedge C_k \Rightarrow Toothache$ but maybe not always …

Problems in specifying the correct logical rules:

- *Complexity*: too much work to list the complete set of antecedents or consequents
- *Theoretical ignorance*: no complete theory for the domain
- *Practical ignorance*: no complete knowledge of the patient

Probability provides a way of summarizing the uncertainty that comes from complexity and ignorance, thereby solving the qualification problem.

Probability theory

Logic theory and probability theory both talk about a world make of propositions which are *true* or *false*. They share the **ontological commitment**.

What is different is the **epistemological commitment**: a logical agent believes each sentence to be true or false or has no opinion, whereas a probabilistic agent may have a numerical **degree of belief** between 0 (for sentences that are certainly false) and 1 (certainly true).

Example: The patient who has a toothache has a cavity with 0.8 probability.

The uncertainty is not in the world, but in the beliefs of the agent (**state of knowledge**).

If the knowledge about the world changes (we learn more information about the patient) the probability changes, but there is no contradiction.

Probabilities: a very gentle AI-*sh* introduction

Probabilistic assertions are assertions about **possible worlds** stating how **probable** a world is.

The set of possible worlds, also called the **sample space,** Ω (Omega)

The possible worlds are **mutually exclusive** and **exhaustive.**

Example: the 36 possible outcomes of rolling two dices.

A fully specified **probability model** associates a probability P(a real number between 0 and 1) to each possible world w in Ω.

Basic axiom of probability

$$0 \leq P(w) \leq 1 \text{ for every } w \text{ and } \sum_{w \in \Omega} P(w) = 1 \qquad (1)$$

Events

Usually we deal with subsets of possible worlds (**events**) which can be described by an expression (a **proposition**) in a formal language.

Events are the possible worlds where the proposition holds.

Def $P(\Phi)$: For any event Φ, $P(\Phi) = \sum_{w \in \Phi} P(w)$ (2)

Example 1: when rolling two fair dice the probability of the event "*total is 11*", is the probability of all the possible worlds where the sum of the dice is 11.

$P(Total{=}11) = P(Dice_1{=}5, Dice_2{=}6) + P(Dice_1{=}6, Dice_2{=}5) = 1/36 + 1/36 = 1/18$

Example 2: *Double* is the proposition for the event of both dice giving the same number

$P(doubles) = \ldots$

$P(Total{=}11)$ and $P(double)$ are called **unconditional** or **prior** probabilities or **priors**.

They refer to degrees of belief in propositions **in the absence of any other information.**

Conditional probabilities

Most often we are given some **evidence** restricting the number of possible worlds and *conditioning* the probability of an event.

Example 1: we can talk of the probability of a *double* given that we know that $Dice_1 = 5$.

$$P(Doubles \mid Dice_1 = 5)$$

Example 2: $P(Cavity) = 0.2$ compared to $P(Cavity \mid Toothache) = 0.6$

These probabilities are called **conditional** or **posterior** probabilities.

Definition of **conditional probability**:

$$P(a \mid b) = \frac{P(a, b)}{P(b)} \qquad \text{with } P(b) > 0 \qquad \text{(}\textit{Conditional probability}\text{)}$$

Note: observing *b* restricts the number of possible worlds to those where *b* is true.
Very often, the definition is used in the following equivalent form:

$$P(a, b) = P(a \mid b)\ P(b) \qquad \text{(}\textit{Product rule}\text{)}$$

Basic probability notation

We will assume that a world is represented by a set of variable/value pairs (a **factored representation** as in CSP). Includes the propositional case.

X: a [random] variable (uppercase)

$dom(X)$: domain of a variable, the values X can take $\{v_1, v_2 \ldots v_k\}$ (values are lowercase)

$P(X=v)$: the probability that $X=v$ where $v \in dom(X)$

$P(v)$: the probability that $X=v$ when there is no ambiguity

e.g. $P(Weather = sunny) = P(sunny)$

If A is a boolean variable, $dom(A) = \{true, false\}$, we can also write

$P(A=true) = P(a)$ and $P(A=false) = P(\neg a)$

e.g. $P(Doubles = true) = P(doubles)$

With ordered domains we can also use comparison operators: e.g. $Age > 10$

The language for propositions

To express complex propositions we can use the connectives of classical propositional logic:

$$P(X = a \wedge Y = b) = P(X = a, Y = b) \quad \textit{joint probability}$$

$$P(X = a \vee Y = b)$$

$$P(\neg(X = a))$$

Examples: *Cavity, Toothache* and *Teen* are three Boolean variables

- $P(cavity \mid \neg toothache \wedge teen) =$

$$= P(cavity \mid (Toothache = false) \wedge (Teen = true)) =$$

$$= P(cavity \mid (Toothache = false, \ Teen = true))$$

Probability distribution: discrete

A full specification of the probability for all values of X is a **probability distribution**

Example 1: if C (coin) is a random variable with values $\{head, tails\}$

- $P(C = head) = 0.5$ $P(C = tails) = 0.5$ is a probability distribution

Example 2: $dom(Weather) = \{sunny, rain, cloudy, snow\}$. Probability distribution:

- $P(Weather = sunny) = 0.6$
- $P(Weather = rain) = 0.1$
- $P(Weather = cloudy) = 0.29$
- $P(Weather = snow) = 0.01$

We can use a vector **P** to represent the distribution of random variables:

- $\mathbf{P}(C) = \langle 0.5, 0.5 \rangle$ $\mathbf{P}(Weather) = \langle 0.6, 0.1, 0.29, 0.01 \rangle$ $\mathbf{P}(sunny) = \langle 0.6 \rangle \Leftrightarrow P(sunny) = 0.6$
- Similarly $\mathbf{P}(X \mid Y)$ is the table of values $P(X = x_i \mid Y = y_j)$, one value for each pair i, j

The vector notation assumes that the values are ordered.

Probability distribution: continuous

For continuous variables we can define the probability that a random variable takes some value x as a function of x, called **probability density function** or **pdf.**

Example: we can assert that the temperature at noon is **distributed uniformly** between 18 and 26 degrees Celsius:

$$f(NoonTemp = x) = Uniform_{[18C,\ 26C]}(x) = 1/8\text{C if } 18\text{C} < x < 26\text{C; 0 otherwise}$$

Then

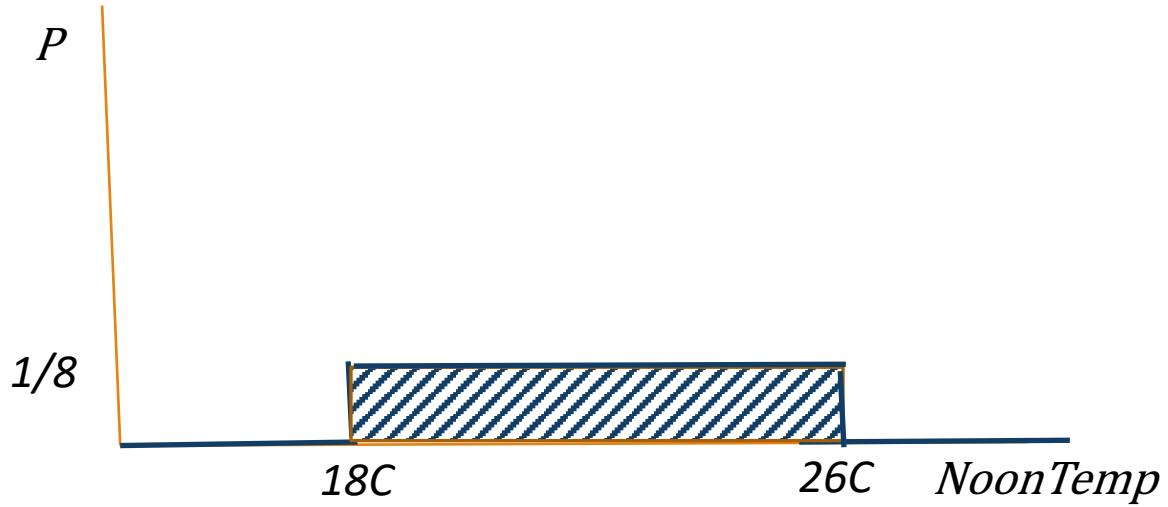
$$P(18 \leq NoonTemp \leq 19) = 1/8$$

$$P(18 \leq NoonTemp \leq 22) = 1/2$$

In general, if f is the density function

$$P(a \leq x \leq b) = \int_a^b f(x)dx$$

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1 \qquad \text{for } f(x) \geq 0$$



Joint probability distribution

A joint probability distribution is a distribution over a set of variables.

P(*Weather* , *Cavity*) denotes the probabilities of all combinations of the values for *Weather* and *Cavity,* A 4 x 2 table of probabilities (|*dom*(*Weather*)| x |*dom*(*Cavity*)|)

P(*sunny* , *Cavity*) where *sunny* is a value, denotes a 2-elements vector $\langle P(sunny, cavity), P(sunny, \neg cavity)\rangle$

The product rule in compact form using variables:

P(*Weather*, *Cavity*) = **P**(*Weather* | *Cavity*) **P**(*Cavity*) is the same as

$$P(W=sunny \wedge C=true) = P(W=sunny \mid C=true)\ P(C=true)$$
$$P(W=rain \wedge C=true) = P(W=rain \mid C=true)\ P(C=true)$$
$$P(W=snow \wedge C=true) = P(W=snow \mid C=true)\ P(C=true)$$
$$P(W=cloudy \wedge C=true) = P(W=cloudy \mid C=true)\ P(C=true)$$

... 8 combinations

Semantics

A **possible world** is defined to be an assignment of values to all the random variables under consideration. It follows that possible worlds are mutually exclusive and exhaustive.

The truth of a complex proposition in a world can be determined using the same recursive definition of truth used for formulas in propositional logic.

If we have a joint probability distribution of **all** the variables (a **full joint probability distribution**), we can perform any inference.

Given that:

1. A proposition identifies a set of possible worlds
2. An entry in the table gives the probability of a possible world
3. For any event Φ, $P(\Phi) = \sum_{w \in \Phi} P(w)$

we can compute the probability of any proposition by taking the sum of the probabilities of the relevant possible worlds in the distribution.

Probability properties (for discrete variables)

Given the basic properties of probabilities:

1. $0 \leq P(w) \leq 1$ for every w and $\sum_{w \in \Omega} P(w) = 1$ (1)
   The summation of the probabilities over all possible worlds is 1

2. For any proposition Φ, $P(\Phi) = \sum_{w \in \Phi} P(w)$ (2)
   The P of a proposition is the sum of the P's of all the worlds satisfying the proposition

Other properties follow:

3. $P(\neg a) = 1 - P(a)$

$$P(\neg a) = \sum_{w \in \neg a} P(w) = \left(\sum_{w \in \neg a} P(w) + \sum_{w \in a} P(w)\right) - \sum_{w \in a} P(w)$$

$$= \sum_{w \in \Omega} P(w) - P(a) = 1 - P(a)$$

4. $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$ (***inclusion-exclusion principle***)

The properties 1-4 are called **Kolmogorov's axioms**. A different axiomatization in FCA

Discussion

Plausibility. Suppose a state of beliefs violating Kolmogorov's axioms:

$$P(a) = 0.4 \quad P(a \wedge b) = P(a, b) = 0.0 \qquad P(b) = 0.3 \qquad P(a \vee b) = 0.8$$

De Finetti proved that if an agent holds an inconsistent set of beliefs, then if he bets according to this set of beliefs against another agent, then the other agent can devise a strategy to make him always loose money.

De Finetti's theorem implies that no rational agent can have beliefs that violate the axioms of probability.

Origin of probability. What is the nature and source of probability numbers?

- *Frequentist view*: the numbers come from experiments
- *Objectivist view:* probabilities are real aspects of the way objects behave in the world
- *Subjectivist view*: probabilities as a way of characterizing an agent's beliefs ascribing values (**Bayesian probability** or subjective probability)

Parallel with classical logical inference (Barber)

In propositional logic "*All apples are fruit*" ($A \Rightarrow F$) and "*All fruits grow on trees*" ($F \Rightarrow T$) lead to the conclusion that "*All apples grow on trees*" ($A \Rightarrow T$), by transitivity of ⇒.

Using Bayesian reasoning.

1. $P(fruit \mid apple) = 1$ "*All apples are fruit*"
2. $P(tree \mid fruit) = 1$ "*All fruit grows on trees*"

We then want to show that 1-2 imply:

$P(tree \mid apple) = 1$ which is the same as proving $P(\neg tree \mid apple) = 0$

$P(\neg tree,\ apple) = 0$ assuming $P(apple) > 0$, by definition of conditional probability

Given $P(\neg tree,\ apple) = P(\neg tree, apple, fruit) + P(\neg tree, apple, \neg fruit)$
we can show that both terms on the right are zero.

- ✓ $P(\neg tree, apple, fruit) \leq P(\neg tree, fruit) = [1 - P(tree \mid fruit)]\ P(fruit) = 0$ (*Product rule, 2*)
- ✓ $P(\neg tree, apple, \neg fruit) \leq P(\neg fruit, apple) = [1 - P(fruit \mid apple)]\ P(apple) = 0$ (*Product rule, 1*)

Probabilistic inference

COMPUTING WITH FULL JOINT DISTRIBUTION -

Probabilistic inference

We will know assume to have a full joint distribution and show how several inferences can be done, by using the axioms of probability.

We assume to use the **full joint distribution** as the “knowledge base”.

1. Compute the prior probability of a single variable (*marginalization*)
2. Compute probability of complex propositions (and, or, not);
3. Compute the *posterior probability* for a proposition given observed evidence.

Computing *marginals*

How to compute the probability of a variable from the full joint distribution.

Given a joint distribution $\mathbf{P}(X, Y)$ over variables X and Y, the distribution of the single variable X is given by:

$$\mathbf{P}(X) = \sum_{y \in dom(Y)} \mathbf{P}(X, y) = \sum_{y} \mathbf{P}(X, y)$$

This operation is also called **marginalization** or **summing out**.

In general, if **Y** and **Z** are sets of variables, such that **Y** ∪ **Z** are all the variables in the full joint distribution

$$\mathbf{P}(\mathbf{Y}) = \sum_{Z \in \mathbf{Z}} \mathbf{P}(\mathbf{Y}, z)$$

where z are all possible tuples of variables in **Z**

Marginal probability, *marginalization*

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

Boolean variables: *Toothache*, *Cavity*, and *Catch* (the dentist's nasty steel probe catching in the tooth). Three Booleans. The full joint distribution is a 2 x 2 x 2 entry table.

Unconditional or **marginal** probability of *cavity*:

- $\boldsymbol{P}(cavity) = \boldsymbol{P}(Cavity = True) = \sum_{\mathbf{z} \in \{Catch,\ Toothache\}} \boldsymbol{P}(cavity, \mathbf{z}) =$

  $= 0.108 + 0.012 + 0.072 + 0.008 = 0.2$ *summing out*
- $\mathbf{P}(Cavity) = \langle 0.2, 0.8 \rangle$

Conditioning

Y and **Z** are sets of variables:

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{Z} \in \mathbf{Z}} \mathbf{P}(\mathbf{Y}, \mathbf{z}) \qquad (\textit{marginalization})$$

A variant of the marginalization rule, called **conditioning**, involves conditional probabilities instead of joint probabilities

It can be obtained from the previous using the *product rule*:

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{Z} \in \mathbf{Z}} \mathbf{P}(\mathbf{Y} \mid \mathbf{z}) \, P(\mathrm{z}) \qquad (\textit{conditioning})$$

Inference with full 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

For a proposition with '∧' we sum the numbers of the entries satisfying both conjuncts:

$P(cavity \wedge toothache) = P(cavity, toothache) = (0.108 + 0.012) = 0.12$

$P(cavity \wedge catch) = (0.108 + 0.072) = 0.18$

Inference with full 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

$$P(cavity \vee toothache) = (0.108 + 0.012 + 0.072 + 0.008) + (0.108 + 0.012 + 0.016 + 0.064) - (0.108 + 0.012) = 0.28$$

$$P(\neg cavity) = (0.016 + 0.064 + 0.144 + 0.576) = 0.8$$

Conditional probability and normalization

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

Computing conditionals:

$P(cavity \mid toothache) = P(cavity, toothache) \,/\, P(toothache) =$

$(0.108 + 0.012) \,/(0.108 + 0.012 + 0.016 + 0.064) = 0.12/0.2 = 0.6$

$P(\neg cavity \mid toothache) = P(\neg cavity, toothache)/P(toothache) =$

$(0.016 + 0.064) \,/(0.108 + 0.012 + 0.016 + 0.064) = 0.08/0.2 = 0.4$

The term $1/0.2 = \alpha$ is a **normalization costant** that doesn't need to be computed.

$\mathbf{P}(Cavity \mid toothache) = \alpha\, \mathbf{P}(Cavity, toothache) =$ ***P*(*toothache*) not necessary**

$\alpha\, [\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}\,(Cavity, toothache, \neg catch)] =$

$\alpha\, [\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] = \alpha\, \langle 0.12, 0.08\,\rangle = \langle 0.6, 0.4\rangle$ dividing by 0.2

A general inference procedure

If the query involves a single variable X (i.e. *Cavity*), **e** is the list of the observed values, the evidence (i.e. *Toothache*), and **Y** the rest of unobserved variables (i.e. *Catch*):

$$\mathbf{P}(X|\mathbf{e}) = \alpha\, \mathbf{P}(X, \mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y})$$

However the complexity of the joint distribution table is intractable: if n is the number of boolean variables, it requires an input table of size $O(2^n)$ and takes $O(2^n)$ time to process,

Next. we will introduce more practical reasoning mechanisms:

- ✓ leveraging on the notion of **independence**
- ✓ using Bayes theorem

Adding an independent variable

Let's add the variable *Weather* with 4 values {*cloudy, sunny, …*}

P(*Toothache, Catch, Cavity, Weather*) is the new full distribution with 32 entries.

$P(toothache, catch, cavity, cloudy) =$ (*product rule*)

$P(cloudy \mid toothache, catch, cavity)\ P(toothache, catch, cavity)$

Since "cloudiness" (and *Weather* in general) has nothing to do with dental problems:

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

Therefore:

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

This property is called **independence**.

Independence

Independence of propositions a and b:

$$P(a \mid b) = P(a) \qquad P(b \mid a) = P(b) \qquad P(a, b) = P(a)\ P(b)$$

Independence of variables X and Y:

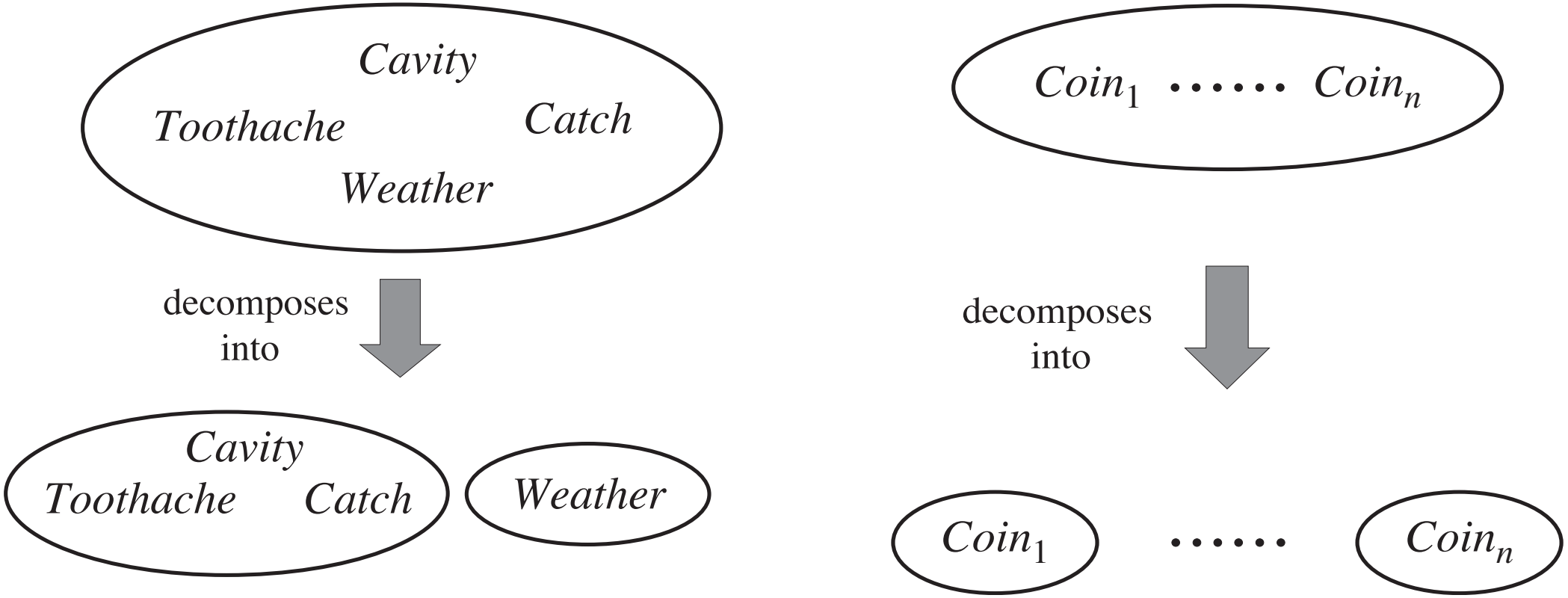
$$\mathbf{P}(X \mid Y) = \mathbf{P}(X) \qquad \mathbf{P}(Y \mid X) = \mathbf{P}(Y) \qquad \mathbf{P}(X, Y) = \mathbf{P}(X)\ \mathbf{P}(Y)$$

Independence assumptions can reduce the size of the representation and the complexity of the inference problem.

For the $Cavity + Weather$ problem the full distribution is actually made of two tables (an 8-entry table and a 4-entry table instead of a 32-entry table)

A notation for independence: $X \perp\!\!\!\perp Y$

Examples


Cavity
Toothache
Catch
Weather
decomposes into
Cavity
Toothache
Catch
Weather
$Coin_1$ ..... $Coin_n$
decomposes into
$Coin_1$
.....
$Coin_n$

Bayes rule

Given the product rule:

1. $P(a, b) = P(a|b)\ P(b)$ and $P(a, b) = P(b|a)\ P(a)$ (*product rule*)
2. $P(a|b)\ P(b) = P(b|a)\ P(a)$ (*equating the right-hand sides*)
3. $P(b|a) = \frac{P(a|b)\ P(b)}{P(a)}$ (***Bayes theorem**/rule/law*)

A more general form, corresponding to a set of equations:

$$\mathbf{P}(Y|X) = \frac{\mathbf{P}(X|Y)\ \mathbf{P}(Y)}{\mathbf{P}(X)} \quad (\textit{Bayes theorem})$$

Bayes' rule tells us how to update the agent's belief in hypothesis $\boldsymbol{h}$ as new evidence $\boldsymbol{e}$ arrives, given the background knowledge $\boldsymbol{k}$.

$$\mathbf{P}(h|\boldsymbol{e}, \boldsymbol{k}) = \frac{\mathbf{P}(e|h, k)\ \mathbf{P}(h|k)}{\mathbf{P}(e|k)} \quad (\textit{Bayes theorem})$$

Use of Bayes rule in a practical case

Why is it useful? Let's give a "diagnostic meaning" to the rule.

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

$P(cause \mid effect)$ goes from effect to cause, *i.e. diagnosys*

$P(effect \mid cause)$ goes from cause to effect, an expert doctor is more likely to have **causal knowledge,** i.e. $P(symptoms \mid desease)$, by knowing how things work and statistics from experience.

Example: $s = stiff\text{-}neck\ ;\ m = meningitis$

$P(s \mid m) = 0.7$

$P(m) = 1/50000$

$P(s) = 0.01$

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.7 \times 1/50000}{0.01} = 0.0014$$

$\mathbf{P}(M \mid s) = \alpha \langle P(s \mid m) \times P(m),\ P(s \mid \neg m) \times P(\neg m) \rangle$

Actually you do not need the evidence $P(s)$

Using Bayes rule: combining evidence

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

Figure 13.3 A full joint distribution for the *Toothache*, *Cavity*, *Catch* world.

P(*Cavity* | *toothache*, *catch*) = α ⟨0.108, 0.016⟩ ≈ ⟨0.871, 0.129⟩

Using Bayes's rule:

P(*Cavity* | *toothache*, *catch*) = α **P**(*toothache*, *catch* | *Cavity*) **P**(*Cavity*) … still complex to compute

Toothache and *Catch* are not independent since they both depend on the presence of a cavity, but they are independent given the presence or the absence of a cavity.

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

Using Bayes rule: conditional independence

We need a refinement of the independence property, called **conditional independence**:

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

With this condition we have:

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

Conditional independence properties can allow probabilistic systems to scale up; moreover, they are much more commonly available than absolute independence.

In general, X and Y are *conditionally independent given* Z, when

$$\mathbf{P}(X, Y \mid Z) = \mathbf{P}(X \mid Z)\ \mathbf{P}(Y \mid Z) \qquad \text{(conditional independence)}$$

Alternative formulation

Given: $\mathbf{P}(X, Y \mid Z) = \mathbf{P}(X \mid Z)\ \mathbf{P}(Y \mid Z)$ *conditional independence*

We can derive the alternative formulation as follows:

$\mathbf{P}(X \mid Y, Z) = \mathbf{P}(X, Y, Z) / \mathbf{P}(Y, Z)$ *def. of conditional probability*

$= \mathbf{P}(Z)\ \mathbf{P}(X, Y \mid Z) / \mathbf{P}(Y, Z)$ *product rule*

$= \mathbf{P}(Z)\ \mathbf{P}(X, Y \mid Z) / \mathbf{P}(Z)\ \mathbf{P}(Y \mid Z)$ *product rule*

$= \mathbf{P}(Z)\ \mathbf{P}(X \mid Z)\ \mathbf{P}(Y \mid Z) / \mathbf{P}(Z)\ \mathbf{P}(Y \mid Z)$ *conditional independence*

$= \mathbf{P}(X \mid Z)$ *simplification*

X and Y are *conditionally independent given* Z:

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

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

Naïve Bayes model

The previous example corresponds to a commonly occurring pattern in which a single cause (*Cavity*) directly influences a number of effects (*Toothache, Catch, …*), all of which are *conditionally independent*, given the cause.

Given this simplifying assumption, the full joint distribution can be computed as:

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

This is called the **Naïve Bayes** model, used in Naïve Bayes classifiers.

Naive Bayes distributions can be learned from observations.

Conclusions

- ✓ We reviewed the basics of probability calculus.
- ✓ Probabilistic inference as a way to compute queries using a full joint distribution table as a KB to be queried.
- ✓ Independence assumptions lead to smaller tables and more efficient computation.
- ✓ Next **belief networks**: a way to encode the representation of a domain to make use of these simplifying assumptions.

References

Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach* (3rd edition). Pearson Education 2010 [Cap 13 – Quantifying uncertainty]

David Barber, *Bayesian Reasoning and Machine Learning,* Online version February 2017 (Ch. 1)

For an alternative, higher level, introduction:

David L. Poole, Alan K. Mackworth. *Artificial Intelligence: foundations of computational agents* (2nd edition), Cambridge University Press, 2017– Computers. http://artint.info/2e/html/ArtInt2e.html