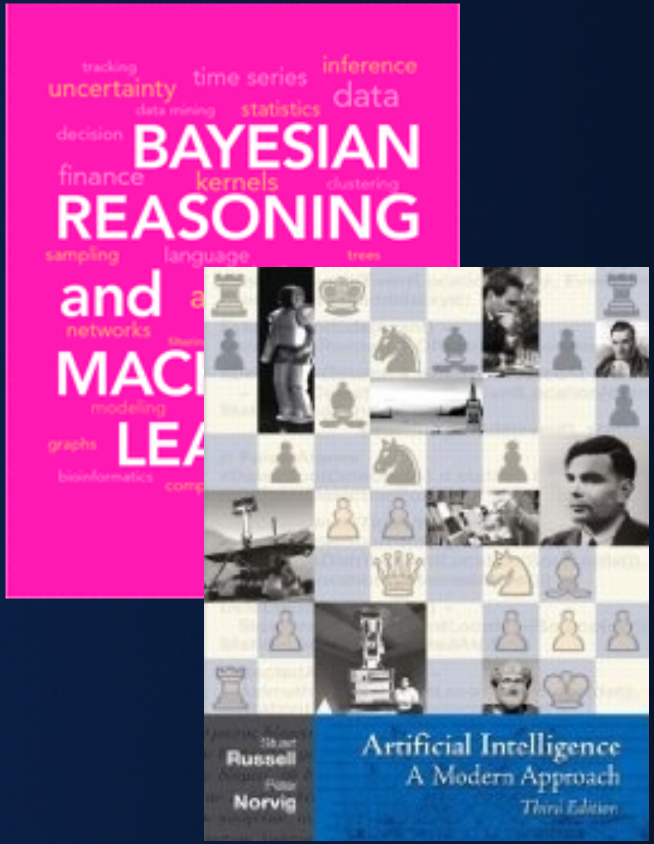


AI Fundamentals: Uncertain reasoning

Maria Simi

Quantifying uncertainty
(AIMA chapter 13)

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

Acting under uncertainty

In **problem solving** we make restrictive assumptions: the world is completely known (**accessible**) and the actions have a predictable effect (**deterministic**).

Agents need a way to handle uncertainty deriving from:

1. partial observability
2. nondeterminism

A partial answer is 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.

Probability theory offers a clean way to quantify uncertainty.

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? You can't.

There are too many sources of uncertainty:

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

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

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

Motivating example 2

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

How to encode the relation in logic?

- $Toothache \Rightarrow Cavity$ (diagnostic rule)
  $Toothache \Rightarrow Cavity \lor GumProblem \lor Abscess \ldots$ there are many possible causes
- $Cavity \Rightarrow Toothache$ (causal rule)
- $Cavity \land C_1 \ldots \land C_k \Rightarrow Toothache$ not always!

Problems in specifying the correct logical rules:

- *Laziness*: 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 our laziness and ignorance, thereby solving the qualification problem.

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

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

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

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

Events

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

Events are the possible worlds where the proposition holds.

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

$P(double) = ?$

$P(Total{=}11)$ 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 have 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 $Dice_1 = 5$.

$$P(Double \,|\, Dice_1 = 5)$$

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

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

Definition of **conditional probability**:

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

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:

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

Basic probability notation

We will assume that a world is represented by a set of variable/value pair (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 $\{v_1, v_2 \dots v_k\}$ (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. $P(Double = \text{true}) = P(double)$

The language for propositions

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

$P(X = a \land Y = b) = P(X = a, Y = b)$ *joint probability*

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

$P(\neg X = a)$

Examples:

- $P(cavity \mid \neg Toothache \land Teen) =$

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

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

Probability distribution: discrete

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

Example 1: if C is a random variable with values $\{head, tail\}$

- $P(C = head) = 0.5$ $P(C = tails) = 0.5$ 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 use a vector **P** to denote the distribution of the random variables C (coin) and $Weather$:

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

The vectorial 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:

$$f(NoonTemp = x) = Uniform_{[18C,\ 26C](}x)$$

Then

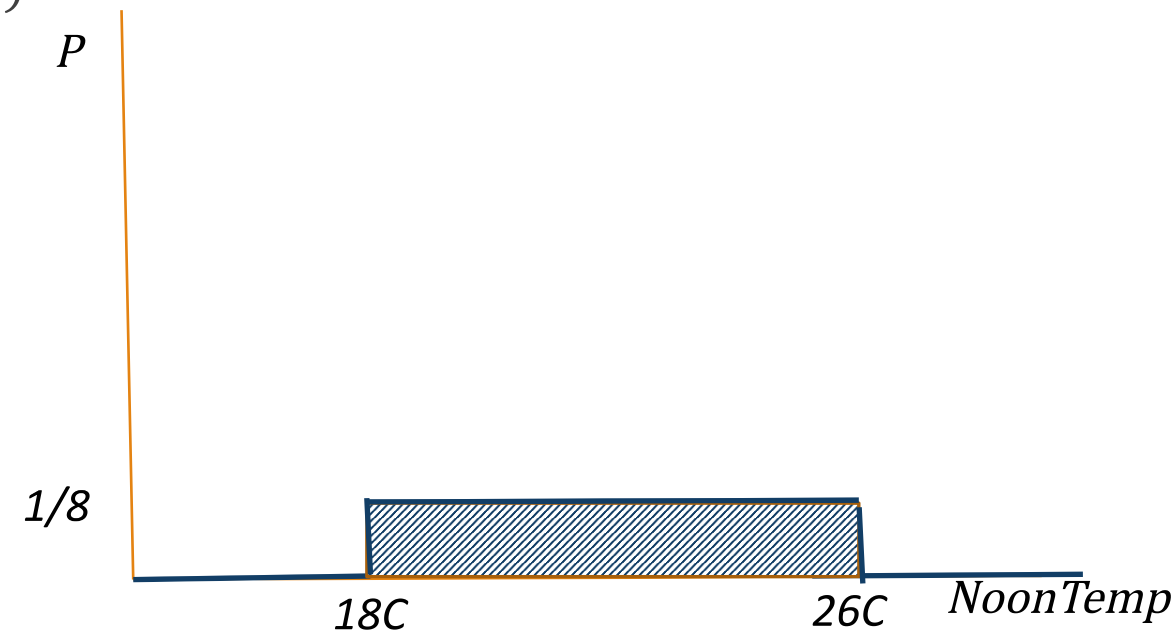
$$P(18 \leq NoonTemp \leq 19) = 1/8$$

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

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. With

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

we denote the probabilities of all combinations of the values for $Weather$ and $Cavity$: A 4 x 2 table of probabilities in this case ($|dom(Weather)|$ x $|dom(Cavity)|$)

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

1. Any proposition identifies a set of possible worlds
2. Any 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.

Computing *marginals, conditioning*

Given a joint distribution $\mathbf{P}(x, y)$ the distribution of a single variable 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**.

$$\mathbf{P}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = \sum_{x_i} \mathbf{P}(x_1, \dots, x_n)$$

In general, if **Y** and **Z** are sets of variables:

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

A variant of this rule, called **conditioning**, involves conditional probabilities instead of joint probabilities, using the *product rule*:

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

Probability axioms (for discrete variables)

Given the basic axioms:

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. $P(\neg a) = 1 - P(a)$
   $P(\neg a) = \sum_{w \in \neg a} P(w) = (\sum_{w \in \neg a} P(w) + \sum_{w \in a} P(w)) - \sum_{w \in a} P(w)$
   $= \sum_{w \in \Omega} P(w) - P(a) = 1 - P(a)$
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**

Example: Classical logical inference (Barber)

In 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 $\Rightarrow$.

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:

$P(\text{Tree} \mid Apple) = 1$ $\quad P(\neg\text{Tree} \mid Apple) = 0$

$P(\neg Tree,\ Apple) = 0$ $\quad$ assuming $P(Apple) > 0$, by definition of conditional probability

Given $P(\neg\text{Tree},\ Apple) = P(\neg\text{Tree}, Apple, Fruit) + P(\neg\text{Tree}, Apple, \neg Fruit)$ $\quad (Marginalization)$
we can show that both terms on the right are zero.

1. $P(\neg\text{Tree}, Apple, Fruit) \leq P(\neg\text{Tree}, Fruit) = [1 - P(\text{Tree}|Fruit)]\, P(Fruit) = 0$ $\;(Product\ rule)$
2. $P(\neg\text{Tree}, Apple, \neg Fruit) \leq P(\neg Fruit, Apple) = [1 - P(Fruit|Apple)]\, P(Apple) = 0$

Discussion

Plausibility. State of beliefs, apparently plausible, violating Kolmogorov's axioms:

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

De Finetti proved that if an agent holds an inconsistent set of beliefs, than if he bets according to this set of beliefs against another agent, then he will 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

Probabilistic inference

Probabilistic inference

We will know assume to have a full joint distribution and show how several inferences can be done.

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

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

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). The full joint distribution is a 2 x 2 x 2 entry table.

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

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

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

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

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

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

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.

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

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

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

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*):

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

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

Next lesson will introduce more practical reasoning mechanisms, leveraging on the notion of **independence**.

Adding an independent variable

Let’s add the variable $Weather$ with 4 values $\{cloudy, sunny, \ldots\}$

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

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

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

Since “cloudiness” (and $Weather$ in general) has nothing to do with dental problems:

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

Therefore:

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

This property is called **independence**.

Independence

Independence of propositions a and b:

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

Independence of variables X and Y:

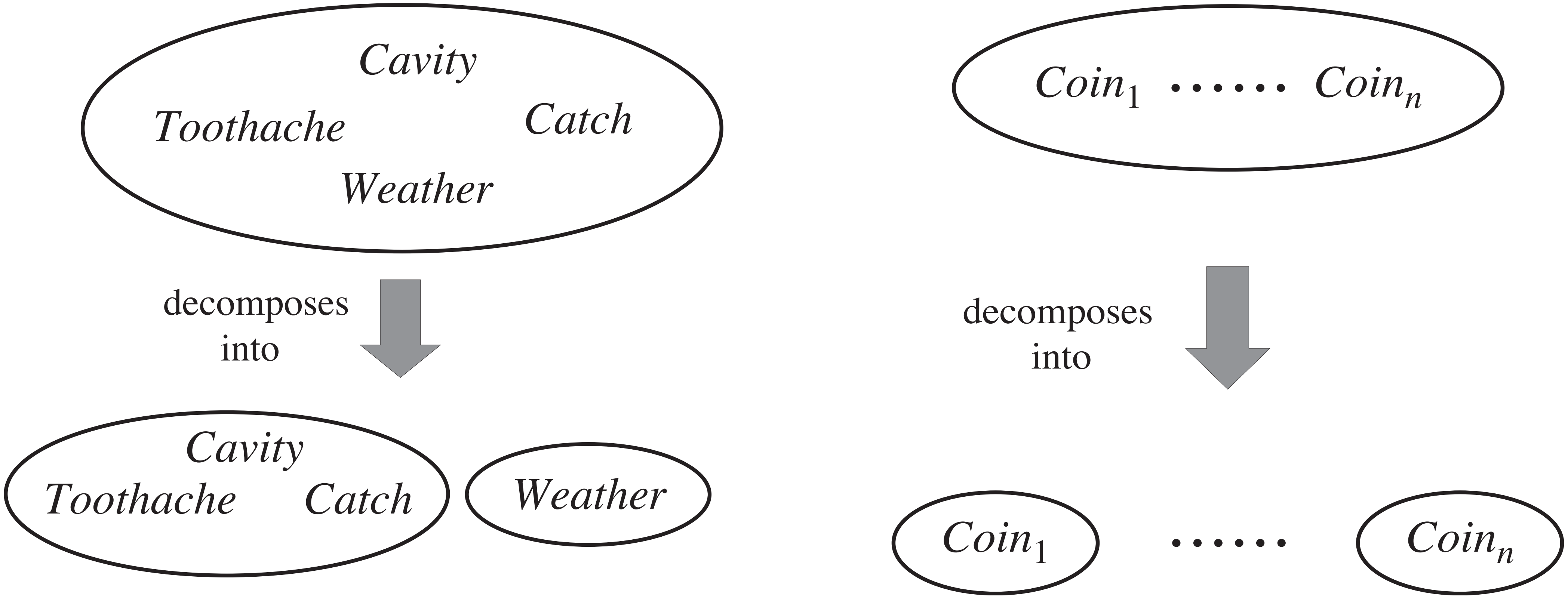
$\mathbf{P}(X\,|\,Y) = \mathbf{P}(X) \qquad \mathbf{P}(Y\,|\,X) = \mathbf{P}(Y) \qquad \mathbf{P}(X, Y) = \mathbf{P}(X)\,\mathbf{P}(Y)$

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. $P(a, b) = P(a \mid b)\ P(b) \qquad P(a, b) = P(b \mid a)\ P(a) \qquad (\textit{product rule})$
2. $P(a \mid b)\ P(b) = P(b \mid a)\ P(a) \qquad (\textit{equating the right-hand sides})$
3. $P(b \mid a) = \dfrac{P(a \mid b)\ P(b)}{P(a)} \qquad (\textit{Bayes theorem/rule/law})$

Other more general forms:

$$\mathbf{P}(Y \mid X) = \frac{\mathbf{P}(X \mid Y)\ \mathbf{P}(Y)}{\mathbf{P}(X)} \qquad (\textit{Bayes theorem})$$

$$\mathbf{P}(Y \mid X, \mathbf{e}) = \frac{\mathbf{P}(X \mid Y, \mathbf{e})\ \mathbf{P}(Y \mid \mathbf{e})}{\mathbf{P}(X \mid \mathbf{e})} \qquad (\textit{Bayes theorem})$$

where ***e*** is a set of evidence variables.

Use of Bayes rule: the simple case

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

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

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

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

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

$P(s \mid m) = 0.7$

$P(m) = 1/50000$

$P(s) = 0.01$

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

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

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*)

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:

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

In general, given the variables X, Y and Z

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

Alternative formulation:

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

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

Naïve Bayes model

This 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:

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

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

Naive Bayes distributions can be learnied from observations.

Conclusions

- ✓ We reviewed the basics of propositional 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 in the representation of a domain these simplifying assumptions.

Your turn

- If you don't feel confident with these computations, do some of the exercises at the end of Barber (Chapter 1).
- *Are basic axioms of probability reasonable*?
  - Discuss the meaning of the basic axioms of "classical probability" and their implications. Start with AIMA, 13.2.3
- *Where do probabilities come from*?
  - Present the different views about the source of probabilities: frequentist, objectivist, subjectivist … Start with AIMA (3rd edition), box page 491

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)