Artificial Intelligence

A Modern Approach

Second Edition

Stuart J. Russell and Peter Norvig

Contributing writers:
John F. Canny
Douglas D. Edwards
Jitendra M. Malik
Sebastian Thrun

Pearson Education, Inc., Upper Saddle River, New Jersey 07458

13 UNCERTAINTY

In which we see what an agent should do when not all is crystal clear.

13.1 ACTING UNDER UNCERTAINTY

UNCERTAINTY

The logical agents described in Parts III and IV make the epistemological commitment that propositions are true, false, or unknown. When an agent knows enough facts about its environment, the logical approach enables it to derive plans that are guaranteed to work. This is a good thing. Unfortunately, *agents almost never have access to the whole truth about their environment.* Agents must, therefore, act under **uncertainty**. For example, an agent in the wumpus world of Chapter 7 has sensors that report only local information; most of the world is not immediately observable. A wumpus agent often will find itself unable to discover which of two squares contains a pit. If those squares are *en route* to the gold, then the agent might have to take a chance and enter one of the two squares.

The real world is far more complex than the wumpus world. For a logical agent, it might be impossible to construct a complete and correct description of how its actions will work. Suppose, for example, that the agent wants to drive someone to the airport to catch a flight and is considering a plan, A_{90}, that involves leaving home 90 minutes before the flight departs and driving at a reasonable speed. Even though the airport is only about 15 miles away, the agent will not be conclude with certainty that "Plan A_{90} will get us to the airport in time." Instead, it reaches the weaker conclusion "Plan A_{90} will get us to the airport in time, as long as my car doesn't break down or run out of gas, and I don't get into an accident, and there are no accidents on the bridge, and the plane doesn't leave early, and . . . ." None of these conditions can be deduced, so the plan's success cannot be inferred. This is an example of the **qualification problem** mentioned in Chapter 10.

If a logical agent cannot conclude that any particular course of action achieves its goal, then it will be unable to act. Conditional planning can overcome uncertainty to some extent, but only if the agent's sensing actions can obtain the required information and only if there are not too many different contingencies. Another possible solution would be to endow the agent with a simple but incorrect theory of the world that *does* enable it to derive a plan;

presumably, such plans will work *most* of the time, but problems arise when events contradict the agent's theory. Moreover, handling the tradeoff between the accuracy and usefulness of the agent's theory seems itself to require reasoning about uncertainty. In sum, no purely logical agent will be able to conclude that plan A_{90} is the right thing to do.

Nonetheless, let us suppose that A_{90} *is* in fact the right thing to do. What do we mean by saying this? As we discussed in Chapter 2, we mean that out of all the plans that could be executed, A_{90} is expected to maximize the agent's performance measure, given the information it has about the environment. The performance measure includes getting to the airport in time for the flight, avoiding a long, unproductive wait at the airport, and avoiding speeding tickets along the way. The information the agent has cannot guarantee any of these outcomes for A_{90}, but it can provide some degree of belief that they will be achieved. Other plans, such as A_{120}, might increase the agent's belief that it will get to the airport on time, but also increase the likelihood of a long wait. *The right thing to do—the* ***rational decision****—therefore depends on both the relative importance of various goals and the likelihood that, and degree to which, they will be achieved.* The remainder of this section hones these ideas, in preparation for the development of the general theories of uncertain reasoning and rational decisions that we present in this and subsequent chapters.

Handling uncertain knowledge

In this section, we look more closely at the nature of uncertain knowledge. We will use a simple diagnosis example to illustrate the concepts involved. Diagnosis—whether for medicine, automobile repair, or whatever—is a task that almost always involves uncertainty. Let us try to write rules for dental diagnosis using first-order logic, so that we can see how the logical approach breaks down. Consider the following rule:

$$\forall p \;\; Symptom(p, Toothache) \Rightarrow Disease(p, Cavity) \; .$$

The problem is that this rule is wrong. Not all patients with toothaches have cavities; some of them have gum disease, an abscess, or one of several other problems:

$$\forall p \;\; Symptom(p, Toothache) \Rightarrow \\ Disease(p, Cavity) \vee Disease(p, GumDisease) \vee Disease(p, Abscess) \ldots$$

Unfortunately, in order to make the rule true, we have to add an almost unlimited list of possible causes. We could try turning the rule into a causal rule:

$$\forall p \;\; Disease(p, Cavity) \Rightarrow Symptom(p, Toothache) \; .$$

But this rule is not right either; not all cavities cause pain. The only way to fix the rule is to make it logically exhaustive: to augment the left-hand side with all the qualifications required for a cavity to cause a toothache. Even then, for the purposes of diagnosis, one must also take into account the possibility that the patient might have a toothache and a cavity that are unconnected.

Trying to use first-order logic to cope with a domain like medical diagnosis thus fails for three main reasons:

LAZINESS

- ◇ **Laziness**: It is too much work to list the complete set of antecedents or consequents needed to ensure an exceptionless rule and too hard to use such rules.

THEORETICAL IGNORANCE

◇ **Theoretical ignorance**: Medical science has no complete theory for the domain.

PRACTICAL IGNORANCE

◇ **Practical ignorance**: Even if we know all the rules, we might be uncertain about a particular patient because not all the necessary tests have been or can be run.

The connection between toothaches and cavities is just not a logical consequence in either direction. This is typical of the medical domain, as well as most other judgmental domains: law, business, design, automobile repair, gardening, dating, and so on. The agent's knowledge can at best provide only a **degree of belief** in the relevant sentences. Our main tool for dealing with degrees of belief will be **probability theory**, which assigns to each sentence a numerical degree of belief between 0 and 1. (Some alternative methods for uncertain reasoning are covered in Section 14.7.)

DEGREE OF BELIEF

PROBABILITY THEORY

Probability provides a way of **summarizing** *the uncertainty that comes from our laziness and ignorance.* We might not know for sure what afflicts a particular patient, but we believe that there is, say, an 80% chance—that is, a probability of 0.8—that the patient has a cavity if he or she has a toothache. That is, we expect that out of all the situations that are indistinguishable from the current situation as far as the agent's knowledge goes, the patient will have a cavity in 80% of them. This belief could be derived from statistical data—80% of the toothache patients seen so far have had cavities—or from some general rules, or from a combination of evidence sources. The 80% summarizes those cases in which all the factors needed for a cavity to cause a toothache are present and other cases in which the patient has both toothache and cavity but the two are unconnected. The missing 20% summarizes all the other possible causes of toothache that we are too lazy or ignorant to confirm or deny.

Assigning probability of 0 to a given sentence corresponds to an unequivocal belief that the sentence is false, while assigning a probability of 1 corresponds to an unequivocal belief that the sentence is true. Probabilities between 0 and 1 correspond to intermediate degrees of belief in the truth of the sentence. The sentence itself is *in fact* either true or false. It is important to note that a degree of belief is different from a degree of truth. A probability of 0.8 does not mean "80% true" but rather an 80% degree of belief—that is, a fairly strong expectation. Thus, probability theory makes the same ontological commitment as logic—namely, that facts either do or do not hold in the world. Degree of truth, as opposed to degree of belief, is the subject of **fuzzy logic**, which is covered in Section 14.7.

In logic, a sentence such as "The patient has a cavity" is true or false depending on the interpretation and the world; it is true just when the fact it refers to is the case. In probability theory, a sentence such as "The probability that the patient has a cavity is 0.8" is about the agent's beliefs, not directly about the world. These beliefs depend on the percepts that the agent has received to date. These percepts constitute the **evidence** on which probability assertions are based. For example, suppose that the agent has drawn a card from a shuffled pack. Before looking at the card, the agent might assign a probability of $1/52$ to its being the ace of spades. After looking at the card, an appropriate probability for the same proposition would be 0 or 1. Thus, an assignment of probability to a proposition is analogous to saying whether a given logical sentence (or its negation) is entailed by the knowledge base, rather than whether or not it is true. Just as entailment status can change when more sentences are

EVIDENCE

added to the knowledge base, probabilities can change when more evidence is acquired.[1]

All probability statements must therefore indicate the evidence with respect to which the probability is being assessed. As the agent receives new percepts, its probability assessments are updated to reflect the new evidence. Before the evidence is obtained, we talk about **prior** or **unconditional** probability; after the evidence is obtained, we talk about **posterior** or **conditional** probability. In most cases, an agent will have some evidence from its percepts and will be interested in computing the posterior probabilities of the outcomes it cares about.

Uncertainty and rational decisions

The presence of uncertainty radically changes the way an agent makes decisions. A logical agent typically has a goal and executes any plan that is guaranteed to achieve it. An action can be selected or rejected on the basis of whether it achieves the goal, regardless of what other actions might achieve. When uncertainty enters the picture, this is no longer the case. Consider again the A_{90} plan for getting to the airport. Suppose it has a 95% chance of succeeding. Does this mean it is a rational choice? Not necessarily: There might be other plans, such as A_{120}, with higher probabilities of success. If it is vital not to miss the flight, then it is worth risking the longer wait at the airport. What about A_{1440}, a plan that involves leaving home 24 hours in advance? In most circumstances, this is not a good choice, because, although it almost guarantees getting there on time, it involves an intolerable wait.

PREFERENCES
OUTCOMES
UTILITY THEORY

To make such choices, an agent must first have **preferences** between the different possible **outcomes** of the various plans. A particular outcome is a completely specified state, including such factors as whether the agent arrives on time and the length of the wait at the airport. We will be using **utility theory** to represent and reason with preferences. (The term **utility** is used here in the sense of "the quality of being useful," not in the sense of the electric company or water works.) Utility theory says that every state has a degree of usefulness, or utility, to an agent and that the agent will prefer states with higher utility.

The utility of a state is relative to the agent whose preferences the utility function is supposed to represent. For example, the payoff functions for games in Chapter 6 are utility functions. The utility of a state in which White has won a game of chess is obviously high for the agent playing White, but low for the agent playing Black. Or again, some players (including the authors) might be happy with a draw against the world champion, whereas other players (including the former world champion) might not. There is no accounting for taste or preferences: you might think that an agent who prefers jalapeño bubble-gum ice cream to chocolate chocolate chip is odd or even misguided, but you could not say the agent is irrational. A utility function can even account for altruistic behavior, simply by including the welfare of others as one of the factors contributing to the agent's own utility.

DECISION THEORY

Preferences, as expressed by utilities, are combined with probabilities in the general theory of rational decisions called **decision theory**:

Decision theory = *probability theory* + *utility theory* .

[1] This is quite different from a sentence's becoming true or false as the world changes. Handling a changing world via probabilities requires the same kinds of mechanisms—situations, intervals, and events—that we used in Chapter 10 for logical representations. These mechanisms are discussed in Chapter 15.

The fundamental idea of decision theory is that *an agent is rational if and only if it chooses the action that yields the highest expected utility, averaged over all the possible outcomes of the action.* This is called the principle of **Maximum Expected Utility** (MEU). We saw this principle in action in Chapter 6 when we touched briefly on optimal decisions in backgammon. We will see that it is in fact a completely general principle.

Design for a decision-theoretic agent

Figure 13.1 sketches the structure of an agent that uses decision theory to select actions. The agent is identical, at an abstract level, to the logical agent described in Chapter 7. The primary difference is that the decision-theoretic agent's knowledge of the current state is uncertain; the agent's **belief state** is a representation of the probabilities of all possible actual states of the world. As time passes, the agent accumulates more evidence and its belief state changes. Given the belief state, the agent can make probabilistic predictions of action outcomes and hence select the action with highest expected utility. This chapter and the next concentrate on the task of representing and computing with probabilistic information in general. Chapter 15 deals with methods for the specific tasks of representing and updating the belief state and predicting the environment. Chapter 16 covers utility theory in more depth, and Chapter 17 develops algorithms for making complex decisions.

BELIEF STATE

```
function DT-AGENT(percept) returns an action
  static: belief_state, probabilistic beliefs about the current state of the world
          action, the agent's action

  update belief_state based on action and percept
  calculate outcome probabilities for actions,
    given action descriptions and current belief_state
  select action with highest expected utility
    given probabilities of outcomes and utility information
  return action
```

Figure 13.1 A decision-theoretic agent that selects rational actions. The steps will be fleshed out in the next five chapters.

13.2 Basic Probability Notation

Now that we have set up the general framework for a rational agent, we will need a formal language for representing and reasoning with uncertain knowledge. Any notation for describing degrees of belief must be able to deal with two main issues: the nature of the sentences to which degrees of belief are assigned and the dependence of the degree of belief on the agent's experience. The version of probability theory we present uses an extension of propositional

logic for its sentences. The dependence on experience is reflected in the syntactic distinction between prior probability statements, which apply before any evidence is obtained, and conditional probability statements, which include the evidence explicitly.

Propositions

Degrees of belief are always applied to **propositions**—assertions that such-and-such is the case. So far we have seen two formal languages—propositional logic and first-order logic—for stating propositions. Probability theory typically uses a language that is slightly more expressive than propositional logic. This section describes that language. (Section 14.6 discusses ways to ascribe degrees of belief to assertions in first-order logic.)

RANDOM VARIABLE

The basic element of the language is the **random variable**, which can be thought of as referring to a "part" of the world whose "status" is initially unknown. For example, $Cavity$ might refer to whether my lower left wisdom tooth has a cavity. Random variables play a role similar to that of CSP variables in constraint satisfaction problems and that of proposition symbols in propositional logic. We will always capitalize the names of random variables. (However, we still use lowercase, single-letter names to represent an unknown random variable, for example: $P(a) = 1 - P(\neg a)$.)

DOMAIN

Each random variable has a **domain** of values that it can take on. For example, the domain of $Cavity$ might be $\langle true, false \rangle$.[2] (We will use lowercase for the names of values.) The simplest kind of proposition asserts that a random variable has a particular value drawn from its domain. For example, $Cavity = true$ might represent the proposition that I do in fact have a cavity in my lower left wisdom tooth.

As with CSP variables, random variables are typically divided into three kinds, depending on the type of the domain:

BOOLEAN RANDOM VARIABLES

- **Boolean random variables**, such as $Cavity$, have the domain $\langle true, false \rangle$. We will often abbreviate a proposition such as $Cavity = true$ simply by the lowercase name $cavity$. Similarly, $Cavity = false$ would be abbreviated by $\neg cavity$.

DISCRETE RANDOM VARIABLES

- **Discrete random variables**, which include Boolean random variables as a special case, take on values from a *countable* domain. For example, the domain of $Weather$ might be $\langle sunny, rainy, cloudy, snow \rangle$. The values in the domain must be mutually exclusive and exhaustive. Where no confusion arises, we will use, for example, $snow$ an an abbreviation for $Weather = snow$.

CONTINUOUS RANDOM VARIABLES

- **Continuous random variables** take on values from the real numbers. The domain can be either the entire real line or some subset such as the interval [0,1]. For example, the proposition $X = 4.02$ asserts that the random variable X has the exact value 4.02. Propositions concerning continuous random variables can also be inequalities, such as $X \leq 4.02$

With some exceptions, we will be concentrating on the discrete case.

Elementary, propositions such as $Cavity = true$ and $Toothache = false$, can be combined to form complex propositions using all the standard logical connectives. For example,

[2] One might expect the domain to be written as a set: $\{true, false\}$. We write it as a tuple because it will be convenient later to impose an ordering on the values.

$Cavity = true \wedge Toothache = false$ is a proposition to which one may ascribe a degree of (dis)belief. As explained in the previous paragraph, this proposition may also be written as $cavity \wedge \neg toothache$.

Atomic events

ATOMIC EVENT

The notion of an **atomic event** is useful in understanding the foundations of probability theory. An atomic event is a *complete* specification of the state of the world about which the agent is uncertain. It can be thought of as an assignment of particular values to all the variables of which the world is composed. For example, if my world consists of only the Boolean variables $Cavity$ and $Toothache$, then there are just four distinct atomic events; the proposition $Cavity = false \wedge Toothache = true$ is one such event.[3]

Atomic events have some important properties:

- They are *mutually exclusive*—at most one can actually be the case. For example, $cavity \wedge toothache$ and $cavity \wedge \neg toothache$ cannot both be the case.
- The set of all possible atomic events is *exhaustive*—at least one must be the case. That is, the disjunction of all atomic events is logically equivalent to $true$.
- Any particular atomic event entails the truth or falsehood of every proposition, whether simple or complex. This can be seen by using the standard semantics for logical connectives (Chapter 7). For example, the atomic event $cavity \wedge \neg toothache$ entails the truth of $cavity$ and the falsehood of $cavity \Rightarrow toothache$.
- Any proposition is logically equivalent to the disjunction of all atomic events that entail the truth of the proposition. For example, the proposition $cavity$ is equivalent to disjunction of the atomic events $cavity \wedge toothache$ and $cavity \wedge \neg toothache$.

Exercise 13.4 asks you to prove some of these properties.

Prior probability

UNCONDITIONAL
PRIOR PROBABILITY

The **unconditional** or **prior probability** associated with a proposition a is the degree of belief accorded to it *in the absence of any other information*; it is written as $P(a)$. For example, if the prior probability that I have a cavity is 0.1, then we would write

$$P(Cavity = true) = 0.1 \quad \text{or} \quad P(cavity) = 0.1 \ .$$

It is important to remember that $P(a)$ can be used only when there is no other information. As soon as some new information is known, we must reason with the *conditional* probability of a given that new information. Conditional probabilities are covered in the next section.

Sometimes, we will want to talk about the probabilities of all the possible values of a random variable. In that case, we will use an expression such as $\mathbf{P}(Weather)$, which denotes a *vector* of values for the probabilities of each individual state of the weather. Thus, instead

[3] Many standard formulations of probability theory take atomic events, also known as **sample points**, as primitive and define a random variable as a function taking an atomic event as input and returning a value from the appropriate domain. Such an approach is perhaps more general, but also less intuitive.

of writing the four equations

$$P(Weather = sunny) = 0.7$$
$$P(Weather = rain) = 0.2$$
$$P(Weather = cloudy) = 0.08$$
$$P(Weather = snow) = 0.02 .$$

we may simply write

$$\mathbf{P}(Weather) = \langle 0.7, 0.2, 0.08, 0.02 \rangle .$$

PROBABILITY DISTRIBUTION

This statement defines a prior **probability distribution** for the random variable $Weather$.

We will also use expressions such as $\mathbf{P}(Weather, Cavity)$ to denote the probabilities of all combinations of the values of a set of random variables.[4] In that case, $\mathbf{P}(Weather, Cavity)$ can be represented by a 4×2 table of probabilities. This is called the **joint probability distribution** of $Weather$ and $Cavity$.

JOINT PROBABILITY DISTRIBUTION

Sometimes it will be useful to think about the complete set of random variables used to describe the world. A joint probability distribution that covers this complete set is called the **full joint probability distribution**. For example, if the world consists of just the variables $Cavity$, $Toothache$, and $Weather$, then the full joint distribution is given by

FULL JOINT PROBABILITY DISTRIBUTION

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

This joint distribution can be represented as a $2 \times 2 \times 4$ table with 16 entries. A full joint distribution specifies the probability of every atomic event and is therefore a complete specification of one's uncertainty about the world in question. We will see in Section 13.4 that any probabilistic query can be answered from the full joint distribution.

For continuous variables, it is not possible to write out the entire distribution as a table, because there are infinitely many values. Instead, one usually defines the probability that a random variable takes on some value x as a parameterized function of x. For example, let the random variable X denote tomorrow's maximum temperature in Berkeley. Then the sentence

$$P(X = x) = U[18, 26](x)$$

expresses the belief that X is distributed uniformly between 18 and 26 degrees Celsius. (Several useful continuous distributions are defined in Appendix A.) Probability distributions for continuous variables are called **probability density functions**. Density functions differ in meaning from discrete distributions. For example, using the temperature distribution given earlier, we find that $P(X = 20.5) = U[18, 26](20.5) = 0.125/C$. This does *not* mean that there's a 12.5% chance that the maximum temperature will be *exactly* 20.5 degrees tomorrow; the probability that this will happen is of course zero. The technical meaning is that the probability that the temperature is in a small region around 20.5 degrees is equal, in the limit, to 0.125 divided by the width of the region in degrees Celsius:

PROBABILITY DENSITY FUNCTIONS

$$\lim_{dx \to 0} P(20.5 \leq X \leq 20.5 + dx)/dx = 0.125/C .$$

[4] The general notational rule is that the distribution covers all values of the variables that are capitalized. Thus, the expression $\mathbf{P}(Weather, cavity)$ is a four-element vector of probabilities for the conjunction of each weather type with $Cavity = true$.

Some authors use different symbols for discrete distributions and density functions; we use P in both cases, since confusion seldom arises and the equations are usually identical. Note that probabilities are unitless numbers, whereas density functions are measured with a unit, in this case reciprocal degrees.

Conditional probability

CONDITIONAL PROBABILITY
POSTERIOR PROBABILITY

Once the agent has obtained some evidence concerning the previously unknown random variables making up the domain, prior probabilities are no longer applicable. Instead, we use **conditional** or **posterior** probabilities. The notation used is $P(a|b)$, where a and b are any propositions.[5] This is read as "the probability of a, given that *all we know* is b." For example,

$$P(cavity|toothache) = 0.8$$

indicates that if a patient is observed to have a toothache and no other information is yet available, then the probability of the patient's having a cavity will be 0.8. A prior probability, such as $P(cavity)$, can be thought of as a special case of the conditional probability $P(cavity|\ )$, where the probability is conditioned on no evidence.

Conditional probabilities can be defined in terms of unconditional probabilities. The defining equation is

$$P(a|b) = \frac{P(a \wedge b)}{P(b)} \tag{13.1}$$

which holds whenever $P(b) > 0$. This equation can also be written as

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

PRODUCT RULE

which is called the **product rule**. The product rule is perhaps easier to remember: it comes from the fact that, for a and b to be true, we need b to be true, and we also need a to be true given b. We can also have it the other way around:

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

In some cases, it is easier to reason in terms of prior probabilities of conjunctions, but for the most part, we will use conditional probabilities as our vehicle for probabilistic inference.

We can also use the **P** notation for conditional distributions. $\mathbf{P}(X|Y)$ gives the values of $P(X = x_i|Y = y_j)$ for each possible i, j. As an example of how this makes our notation more concise, consider applying the product rule to each case where the propositions a and b assert particular values of X and Y respectively. We obtain the following equations:

$$\begin{aligned}
P(X = x_1 \wedge Y = y_1) &= P(X = x_1|Y = y_1)P(Y = y_1)\ . \\
P(X = x_1 \wedge Y = y_2) &= P(X = x_1|Y = y_2)P(Y = y_2)\ . \\
&\vdots
\end{aligned}$$

We can combine all these into the single equation

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

Remember that this denotes a set of equations relating the corresponding individual entries in the tables, *not* a matrix multiplication of the tables.

[5] The "|" operator has the lowest possible precedence, so $P(a \wedge b|c \vee d)$ means $P((a \wedge b)|(c \vee d))$.

It is tempting, but wrong, to view conditional probabilities as if they were logical implications with uncertainty added. For example, the sentence $P(a|b) = 0.8$ *cannot* be interpreted to mean "whenever b holds, conclude that $P(a)$ is 0.8." Such an interpretation would be wrong on two counts: first, $P(a)$ always denotes the prior probability of a, not the posterior probability given some evidence; second, the statement $P(a|b) = 0.8$ is immediately relevant just when b is the *only* available evidence. When additional information c is available, the degree of belief in a is $P(a|b \wedge c)$, which may could little relation to $P(a|b)$. For example, c might tell us directly whether a is true or false. If we examine a patient who complains of toothache, and discover a cavity, then we have additional evidence $cavity$, and we conclude (trivially) that $P(cavity|toothache \wedge cavity) = 1.0$.

13.3 THE AXIOMS OF PROBABILITY

So far, we have defined a syntax for propositions and for prior and conditional probability statements about those propositions. Now we must provide some sort of semantics for probability statements. We begin with the basic axioms that serve to define the probability scale and its endpoints:

1. All probabilities are between 0 and 1. For any proposition a,

$$0 \leq P(a) \leq 1 .$$

2. Necessarily true (i.e., valid) propositions have probability 1, and necessarily false (i.e., unsatisfiable) propositions have probability 0.

$$P(true) = 1 \qquad P(false) = 0 .$$

Next, we need an axiom that connects the probabilities of logically related propositions. The simplest way to do this is to define the probability of a disjunction as follows:

3. The probability of a disjunction is given by

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b) .$$

This rule is easily remembered by noting that the cases where a holds, together with the cases where b holds, certainly cover all the cases where $a \vee b$ holds; but summing the two sets of cases counts their intersection twice, so we need to subtract $P(a \wedge b)$.

KOLMOGOROV'S AXIOMS

These three axioms are often called **Kolmogorov's axioms** in honor of the Russian mathematician Andrei Kolmogorov, who showed how to build up the rest of probability theory from this simple foundation. Notice that the axioms deal only with prior probabilities rather than conditional probabilities; this is because we have already defined the latter in terms of the former via Equation (13.1).

WHERE DO PROBABILITIES COME FROM?

There has been endless debate over the source and status of probability numbers. The **frequentist** position is that the numbers can come only from *experiments*: if we test 100 people and find that 10 of them have a cavity, then we can say that the probability of a cavity is approximately 0.1. In this view, the assertion "the probability of a cavity is 0.1" means that 0.1 is the fraction that would be observed in the limit of infinitely many samples. From any finite sample, we can estimate the true fraction and also calculate how accurate our estimate is likely to be.

The **objectivist** view is that probabilities are real aspects of the universe—propensities of objects to behave in certain ways—rather than being just descriptions of an observer's degree of belief. For example, that a fair coin comes up heads with probability 0.5 is a propensity of the coin itself. In this view, frequentist measurements are attempts to observe these propensities. Most physicists agree that quantum phenomena are objectively probabilistic, but uncertainty at the macroscopic scale—e.g., in coin tossing—usually arises from ignorance of initial conditions and does not seem consistent with the propensity view.

The **subjectivist** view describes probabilities as a way of characterizing an agent's beliefs, rather than as having any external physical significance. This allows the doctor or analyst to make the numbers up—to say, "In my opinion, I expect the probability of a cavity to be about 0.1." Several more reliable techniques, such as the betting systems described earlier, have also been developed for eliciting probability assessments from humans.

In the end, even a strict frequentist position involves subjective analysis, so the difference probably has little practical importance. The **reference class** problem illustrates the intrusion of subjectivity. Suppose that a frequentist doctor wants to know the chances that a patient has a particular disease. The doctor wants to consider other patients who are similar in important ways—age, symptoms, perhaps sex—and see what proportion of them had the disease. But if the doctor considered everything that is known about the patient—weight to the nearest gram, hair color, mother's maiden name, etc.—the result would be that there are no other patients who are exactly the same and thus no reference class from which to collect experimental data. This has been a vexing problem in the philosophy of science.

Laplace's **principle of indifference** (1816) states that propositions that are syntactically "symmetric" with respect to the evidence should be accorded equal probability. Various refinements have been proposed, culminating in the attempt by Carnap and others to develop a rigorous **inductive logic**, capable of computing the correct probability for any proposition from any collection of observations. Currently, it is believed that no unique inductive logic exists; rather, any such logic rests on a subjective prior probability distribution whose effect is diminished as more observations are collected.

Using the axioms of probability

We can derive a variety of useful facts from the basic axioms. For example, the familiar rule for negation follows by substituting $\neg a$ for b in axiom 3, giving us:

$$\begin{aligned} P(a \vee \neg a) &= P(a) + P(\neg a) - P(a \wedge \neg a) && \text{(by axiom 3 with } b = \neg a) \\ P(true) &= P(a) + P(\neg a) - P(false) && \text{(by logical equivalence)} \\ 1 &= P(a) + P(\neg a) && \text{(by axiom 2)} \\ P(\neg a) &= 1 - P(a) && \text{(by algebra).} \end{aligned}$$

The third line of this derivation is itself a useful fact and can be extended from the Boolean case to the general discrete case. Let the discrete variable D have the domain $\langle d_1, \ldots, d_n \rangle$. Then it is easy to show (Exercise 13.2) that

$$\sum_{i=1}^{n} P(D = d_i) = 1 \ .$$

That is, any probability distribution on a single variable must sum to 1.[6] It is also true that any *joint* probability distribution on any *set* of variables must sum to 1: this can be seen simply by creating a single megavariable whose domain is the cross product of the domains of the the original variables.

Recall that any proposition a is equivalent to the disjunction of all the atomic events in which a holds; call this set of events $\mathbf{e}(a)$. Recall also that atomic events are mutually exclusive, so the probability of any conjunction of atomic events is zero, by axiom 2. Hence, from axiom 3, we can derive the following simple relationship: *The probability of a proposition is equal to the sum of the probabilities of the atomic events in which it holds;* that is,

$$P(a) = \sum_{e_i \in \mathbf{e}(a)} P(e_i) \ . \tag{13.2}$$

This equation provides a simple method for computing the probability of any proposition, given a full joint distribution that specifies the probabilities of all atomic events. (See Section 13.4.) In subsequent sections we will derive additional rules for manipulating probabilities. First, however, we will examine the foundation for the axioms themselves.

Why the axioms of probability are reasonable

The axioms of probability can be seen as restricting the set of probabilistic beliefs that an agent can hold. This is somewhat analogous to the logical case, where a logical agent cannot simultaneously believe A, B, and $\neg(A \wedge B)$, for example. There is, however, an additional complication. In the logical case, the semantic definition of conjunction means that at least one of the three beliefs just mentioned *must be false in the world*, so it is unreasonable for an agent to believe all three. With probabilities, on the other hand, statements refer not to the world directly, but to the agent's own state of knowledge. Why, then, can an agent not hold the following set of beliefs, which clearly violates axiom 3?

$$\begin{aligned} P(a) &= 0.4 & \qquad P(a \wedge b) &= 0.0 \\ P(b) &= 0.3 & \qquad P(a \vee b) &= 0.8 \end{aligned} \tag{13.3}$$

[6] For continuous variables, the summation is replaced by an integral: $\int_{-\infty}^{\infty} P(X = x)\, dx = 1$.

This kind of question has been the subject of decades of intense debate between those who advocate the use of probabilities as the only legitimate form for degrees of belief and those who advocate alternative approaches. Here, we give one argument for the axioms of probability, first stated in 1931 by Bruno de Finetti.

The key to de Finetti's argument is the connection between degree of belief and actions. The idea is that if an agent has some degree of belief in a proposition a, then the agent should be able to state odds at which it is indifferent to a bet for or against a. Think of it as a game between two agents: Agent 1 states "my degree of belief in event a is 0.4." Agent 2 is then free to choose whether to bet for or against a, at stakes that are consistent with the stated degree of belief. That is, Agent 2 could choose to bet that a will occur, betting \$4 against Agent 1's \$6. Or Agent 2 could bet \$6 against \$4 that A will not occur.[7] If an agent's degrees of belief do not accurately reflect the world, then you would expect that it would tend to lose money over the long run to an opposing agent whose beliefs more accurately reflect the state of the world.

But de Finetti proved something much stronger: *If Agent 1 expresses a set of degrees of belief that violate the axioms of probability theory then there is a combination of bets Agent 2 that* **guarantees** *that Agent 1 will lose money* **every** *time.* So if you accept the idea that an agent should be willing to "put its money where its probabilities are," then you should accept that it is irrational to have beliefs that violate the axioms of probability.

One might think that this betting game is rather contrived. For example, what if one refuses to bet? Does that end the argument? The answer is that the betting game is an abstract model for the decision-making situation in which every agent is *unavoidably* involved at every moment. Every action (including inaction) is a kind of bet, and every outcome can be seen as a payoff of the bet. Refusing to bet is like refusing to allow time to pass.

We will not provide the proof of de Finetti's theorem, but we will show an example. Suppose that Agent 1 has the set of degrees of belief from Equation (13.3). Figure 13.2 shows that if Agent 2 chooses to bet \$4 on a, \$3 on b, and \$2 on $\neg(a \vee b)$, then Agent 1 always loses money, regardless of the outcomes for a and b.

Agent 1		Agent 2		Outcome for Agent 1			
Proposition	Belief	Bet	Stakes	$a \wedge b$	$a \wedge \neg b$	$\neg a \wedge b$	$\neg a \wedge \neg b$
a	0.4	a	4 to 6	–6	–6	4	4
b	0.3	b	3 to 7	–7	3	–7	3
$a \vee b$	0.8	$\neg(a \vee b)$	2 to 8	2	2	2	–8
				–11	–1	–1	–1

Figure 13.2 Because Agent 1 has inconsistent beliefs, Agent 2 is able to devise a set of bets that guarantees a loss for Agent 1, no matter what the outcome of a and b.

[7] One might argue that the agent's preferences for different bank balances are such that the possibility of losing \$1 is not counterbalanced by an equal possibility of winning \$1. One possible response is to make the bet amounts small enough to avoid this problem. Savage's analysis (1954) circumvents the issue altogether.

Other strong philosophical arguments have been put forward for the use of probabilities, most notably those of Cox (1946) and Carnap (1950). The world being the way it is, however, practical demonstrations sometimes speak louder than proofs. The success of reasoning systems based on probability theory has been much more effective in making converts. We now look at how the axioms can be deployed to make inferences.

13.4 INFERENCE USING FULL JOINT DISTRIBUTIONS

PROBABILISTIC INFERENCE

In this section we will describe a simple method for **probabilistic inference**—that is, the computation from observed evidence of posterior probabilities for query propositions. We will use the full joint distribution as the "knowledge base" from which answers to all questions may be derived. Along the way we will also introduce several useful techniques for manipulating equations involving probabilities.

We begin with a very simple example: a domain consisting of just the three Boolean variables *Toothache*, *Cavity*, and *Catch* (the dentist's nasty steel probe catches in my tooth). The full joint distribution is a $2 \times 2 \times 2$ table as shown in Figure 13.3.

	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

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

Notice that the probabilities in the joint distribution sum to 1, as required by the axioms of probability. Notice also that Equation (13.2) gives us a direct way to calculate the probability of any proposition, simple or complex: We simply identify those atomic events in which the proposition is true and add up their probabilities. For example, there are six atomic events in which $cavity \vee toothache$ holds:

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

One particularly common task is to extract the distribution over some subset of variables or a single variable. For example, adding the entries in the first row gives the unconditional or

MARGINAL PROBABILITY

marginal probability[8] of *cavity*:

$$P(cavity) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2 .$$

MARGINALIZATION

This process is called **marginalization**, or **summing out**—because the variables other than *Cavity* are summed out. We can write the following general marginalization rule for any sets of variables $\mathbf{Y}$ and $\mathbf{Z}$:

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y}, \mathbf{z}) . \tag{13.4}$$

[8] So called because of a common practice among actuaries of writing the sums of observed frequencies in the margins of insurance tables.

That is, a distribution over $\mathbf{Y}$ can be obtained by summing out all the other variables from any joint distribution containing $\mathbf{Y}$. A variant of this rule involves conditional probabilities instead of joint probabilities, using the product rule:

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y}|\mathbf{z})P(\mathbf{z}) \,. \tag{13.5}$$

CONDITIONING

This rule is called **conditioning**. Marginalization and conditioning will turn out to be useful rules for all kinds of derivations involving probability expressions.

In most cases, we will be interested in computing *conditional* probabilities of some variables, given evidence about others. Conditional probabilities can be found by first using Equation (13.1) to obtain an expression in terms of unconditional probabilities and then evaluating the expression from the full joint distribution. For example, we can compute the probability of a cavity, given evidence of a toothache, as follows:

$$\begin{aligned} P(cavity|toothache) &= \frac{P(cavity \wedge toothache)}{P(toothache)} \\ &= \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6 \,. \end{aligned}$$

Just to check, we can also compute the probability that there is no cavity, given a toothache:

$$\begin{aligned} P(\neg cavity|toothache) &= \frac{P(\neg cavity \wedge toothache)}{P(toothache)} \\ &= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4 \,. \end{aligned}$$

Notice that in these two calculations the term $1/P(toothache)$ remains constant, no matter which value of $Cavity$ we calculate. In fact, it can be viewed as a **normalization** constant for the distribution $\mathbf{P}(Cavity|toothache)$, ensuring that it adds up to 1. Throughout the chapters dealing with probability, we will use α to denote such constants. With this notation, we can write the two preceding equations in one:

NORMALIZATION

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

Normalization will turn out to be a useful shortcut in many probability calculations.

From the example, we can extract a general inference procedure. We will stick to the case in which the query involves a single variable. We will need some notation: let X be the query variable ($Cavity$ in the example), let $\mathbf{E}$ be the set of evidence variables (just $Toothache$ in the example), let $\mathbf{e}$ be the observed values for them, and let $\mathbf{Y}$ be the remaining unobserved variables (just $Catch$ in the example). The query is $\mathbf{P}(X|\mathbf{e})$ and can be evaluated as

$$\mathbf{P}(X|\mathbf{e}) = \alpha\, \mathbf{P}(X, \mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y}) \,, \tag{13.6}$$

where the summation is over all possible $\mathbf{y}$s (i.e., all possible combinations of values of the unobserved variables $\mathbf{Y}$). Notice that together the variables X, $\mathbf{E}$, and $\mathbf{Y}$ constitute the complete set of variables for the domain, so $\mathbf{P}(X, \mathbf{e}, \mathbf{y})$ is simply a subset of probabilities from the full joint distribution. The algorithm is shown in Figure 13.4. It loops over the values

```
function ENUMERATE-JOINT-ASK(X, e, P) returns a distribution over X
  inputs: X, the query variable
          e, observed values for variables E
          P, a joint distribution on variables {X} ∪ E ∪ Y   /* Y = hidden variables */

  Q(X) ← a distribution over X, initially empty
  for each value x_i of X do
    Q(x_i) ← ENUMERATE-JOINT(x_i, e, Y, [], P)
  return NORMALIZE(Q(X))

function ENUMERATE-JOINT(x, e, vars, values, P) returns a real number
  if EMPTY?(vars) then return P(x, e, values)
  Y ← FIRST(vars)
  return Σ_y ENUMERATE-JOINT(x, e, REST(vars), [y|values], P)
```

Figure 13.4 An algorithm for probabilistic inference by enumeration of the entries in a full joint distribution.

of X and the values of Y to enumerate all possible atomic events with **e** fixed, adds up their probabilities from the joint table, and normalizes the results.

Given the full joint distribution to work with, ENUMERATE-JOINT-ASK is a complete algorithm for answering probabilistic queries for discrete variables. It does not scale well, however: 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 the table. In a realistic problem, there might be hundreds or thousands of random variables to consider, not just three. It quickly becomes completely impractical to define the vast numbers of probabilities required—the experience needed in order to estimate each of the table entries separately simply cannot exist.

For these reasons, the full joint distribution in tabular form is not a practical tool for building reasoning systems (although the historical notes at the end of the chapter includes one real-world application of this method). Instead, it should be viewed as the theoretical foundation on which more effective approaches may be built. The remainder of this chapter introduces some of the basic ideas required in preparation for the development of realistic systems in Chapter 14.

13.5 INDEPENDENCE

Let us expand the full joint distribution in Figure 13.3 by adding a fourth variable, *Weather*. The full joint distribution then becomes **P**(*Toothache*, *Catch*, *Cavity*, *Weather*), which has 32 entries (because *Weather* has four values). It contains four "editions" of the table shown in Figure 13.3, one for each kind of weather. It seems natural to ask what relationship these editions have to each other and to the original three-variable table. For example, how are $P(toothache, catch, cavity, Weather = cloudy)$ and $P(toothache, catch, cavity)$ related?

One way to answer this question is to use the product rule:

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

Now, unless one is in the deity business, one should not imagine that one's dental problems influence the weather. Therefore, the following assertion seems reasonable:

$$P(Weather = cloudy | toothache, catch, cavity) = P(Weather = cloudy) . \quad (13.7)$$

From this, we can deduce

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

A similar equation exists for *every entry* in $\mathbf{P}(Toothache, Catch, Cavity, Weather)$. In fact, we can write the general equation

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

Thus, the 32-element table for four variables can be constructed from one 8-element table and one four-element table. This decomposition is illustrated schematically in Figure 13.5(a).

INDEPENDENCE

The property we used in writing Equation (13.7) is called **independence** (also **marginal independence** and **absolute independence**). In particular, the weather is independent of one's dental problems. Independence between propositions a and b can be written as

$$P(a|b) = P(a) \quad \text{or} \quad P(b|a) = P(b) \quad \text{or} \quad P(a \wedge b) = P(a)P(b) . \quad (13.8)$$

All these forms are equivalent (Exercise 13.7). Independence between variables X and Y can be written as follows (again, these are all equivalent):

$$\mathbf{P}(X|Y) = \mathbf{P}(X) \quad \text{or} \quad \mathbf{P}(Y|X) = \mathbf{P}(Y) \quad \text{or} \quad \mathbf{P}(X, Y) = \mathbf{P}(X)\mathbf{P}(Y) .$$

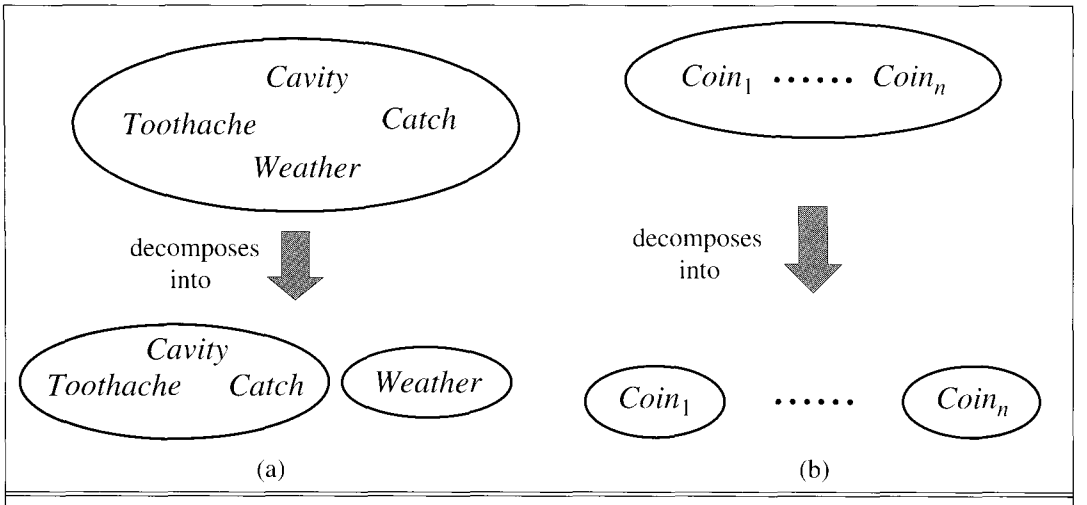

Figure 13.5 Two examples of factoring a large joint distribution into smaller distributions, using absolute independence. (a) Weather and dental problems are independent. (b) Coin flips are independent.

Independence assertions are usually based on knowledge of the domain. As we have seen, they can dramatically reduce the amount of information necessary to specify the full joint distribution. If the complete set of variables can be divided into independent subsets, then the full joint can be *factored* into separate joint distributions on those subsets. For example, the joint distribution on the outcome of n independent coin flips, $\mathbf{P}(C_1, \ldots, C_n)$, can be represented as the product of n single-variable distributions $\mathbf{P}(C_i)$. In a more practical vein, the independence of dentistry and meteorology is a good thing, because otherwise the practice of dentistry might require intimate knowledge of meteorology and *vice versa*.

When they are available, then, 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 is quite rare. Whenever a connection, however indirect, exists between two variables, independence will fail to hold. Moreover, even independent subsets can be quite large—for example, dentistry might involve dozens of diseases and hundreds of symptoms, all of which are interrelated. To handle such problems, we will need more subtle methods than the straightforward concept of independence.

13.6 Bayes' Rule and Its Use

On page 470, we defined the **product rule** and pointed out that it can be written in two forms because of the commutativity of conjunction:

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

Equating the two right-hand sides and dividing by $P(a)$, we get

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)} . \tag{13.9}$$

BAYES' RULE

This equation is known as **Bayes' rule** (also Bayes' law or Bayes' theorem).[9] This simple equation underlies all modern AI systems for probabilistic inference. The more general case of multivalued variables can be written in the **P** notation as

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

where again this is to be taken as representing a set of equations, each dealing with specific values of the variables. We will also have occasion to use a more general version conditionalized on some background evidence **e**:

$$\mathbf{P}(Y|X, \mathbf{e}) = \frac{\mathbf{P}(X|Y, \mathbf{e})\mathbf{P}(Y|\mathbf{e})}{\mathbf{P}(X|\mathbf{e})} . \tag{13.10}$$

[9] According to rule 1 on page 1 of Strunk and White's *The Elements of Style*, it should be Bayes's rather than Bayes'. The latter is, however, more commonly used.

Applying Bayes' rule: The simple case

On the surface, Bayes' rule does not seem very useful. It requires three terms—a conditional probability and two unconditional probabilities—just to compute one conditional probability.

Bayes' rule is useful in practice because there are many cases where we do have good probability estimates for these three numbers and need to compute the fourth. In a task such as medical diagnosis, we often have conditional probabilities on causal relationships and want to derive a diagnosis. A doctor knows that the disease meningitis causes the patient to have a stiff neck, say, 50% of the time. The doctor also knows some unconditional facts: the prior probability that a patient has meningitis is 1/50,000, and the prior probability that any patient has a stiff neck is 1/20. Letting s be the proposition that the patient has a stiff neck and m be the proposition that the patient has meningitis, we have

$$\begin{aligned} P(s|m) &= 0.5 \\ P(m) &= 1/50000 \\ P(s) &= 1/20 \\ P(m|s) &= \frac{P(s|m)P(m)}{P(s)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002 \ . \end{aligned}$$

That is, we expect only 1 in 5000 patients with a stiff neck to have meningitis. Notice that, even though a stiff neck is quite strongly indicated by meningitis (with probability 0.5), the probability of meningitis in the patient remains small. This is because the prior probability on stiff necks is much higher than that on meningitis.

Section 13.4 illustrated a process by which one can avoid assessing the probability of the evidence (here, $P(s)$) by instead computing a posterior probability for each value of the query variable (here, m and $\neg m$) and then normalizing the results. The same process can be applied when using Bayes' rule. We have

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

Thus, in order to use this approach we need to estimate $P(s|\neg m)$ instead of $P(s)$. There is no free lunch—sometimes this is easier, sometimes it is harder. The general form of Bayes' rule with normalization is

$$\mathbf{P}(Y|X) = \alpha \mathbf{P}(X|Y)\mathbf{P}(Y) \ , \tag{13.11}$$

where α is the normalization constant needed to make the entries in $\mathbf{P}(Y|X)$ sum to 1.

One obvious question to ask about Bayes' rule is why one might have available the conditional probability in one direction, but not the other. In the meningitis domain, perhaps the doctor knows that a stiff neck implies meningitis in 1 out of 5000 cases; that is, the doctor has quantitative information in the **diagnostic** direction from symptoms to causes. Such a doctor has no need to use Bayes' rule. Unfortunately, *diagnostic knowledge is often more fragile than causal knowledge.* If there is a sudden epidemic of meningitis, the unconditional probability of meningitis, $P(m)$, will go up. The doctor who derived the diagnostic probability $P(m|s)$ directly from statistical observation of patients before the epidemic will have no idea how to update the value, but the doctor who computes $P(m|s)$ from the other three values will see that $P(m|s)$ should go up proportionately with $P(m)$. Most importantly, the

causal information $P(s|m)$ is *unaffected* by the epidemic, because it simply reflects the way meningitis works. The use of this kind of direct causal or model-based knowledge provides the crucial robustness needed to make probabilistic systems feasible in the real world.

Using Bayes' rule: Combining evidence

We have seen that Bayes' rule can be useful for answering probabilistic queries conditioned on one piece of evidence—for example, the stiff neck. In particular, we have argued that probabilistic information is often available in the form $P(\mathit{effect}|\mathit{cause})$. What happens when we have two or more pieces of evidence? For example, what can a dentist conclude if her nasty steel probe catches in the aching tooth of a patient? If we know the full joint distribution (Figure 13.3), one can read off the answer:

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

We know, however, that such an approach will not scale up to larger numbers of variables.

We can try using use Bayes' rule to reformulate the problem:

$$\mathbf{P}(\mathit{Cavity}|\mathit{toothache} \wedge \mathit{catch}) = \alpha \mathbf{P}(\mathit{toothache} \wedge \mathit{catch}|\mathit{Cavity})\mathbf{P}(\mathit{Cavity}) . \quad (13.12)$$

For this reformulation to work, we need to know the conditional probabilities of the conjunction $\mathit{toothache} \wedge \mathit{catch}$ for each value of Cavity. That might be feasible for just two evidence variables, but again it will not scale up. If there are n possible evidence variables (X rays, diet, oral hygiene, etc.), then there are 2^n possible combinations of observed values for which we would need to know conditional probabilities. We might as well go back to using the full joint distribution. This is what first led researchers away from probability theory toward approximate methods for evidence combination that, while giving incorrect answers, require fewer numbers to give any answer at all.

Rather than taking this route, we need to find some additional assertions about the domain that will enable us to simplify the expressions. The notion of **independence** in Section 13.5 provides a clue, but needs refining. It would be nice if $\mathit{Toothache}$ and Catch were independent, but they are not: if the probe catches in the tooth, it probably has a cavity and that probably causes a toothache. These variables *are* independent, however, *given the presence or the absence of a cavity*. Each is directly caused by the cavity, but neither has a direct effect on the other: toothache depends on the state of the nerves in the tooth, whereas the probe's accuracy depends on the dentist's skill, to which the toothache is irrelevant.[10] Mathematically, this property is written as

$$\mathbf{P}(\mathit{toothache} \wedge \mathit{catch}|\mathit{Cavity}) = \mathbf{P}(\mathit{toothache}|\mathit{Cavity})\mathbf{P}(\mathit{catch}|\mathit{Cavity}) . \quad (13.13)$$

CONDITIONAL INDEPENDENCE

This equation expresses the **conditional independence** of $\mathit{toothache}$ and catch given Cavity. We can plug it into Equation (13.12) to obtain the probability of a cavity:

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

Now the information requirements are the same as for inference using each piece of evidence separately: the prior probability $\mathbf{P}(\mathit{Cavity})$ for the query variable and the conditional probability of each effect, given its cause.

[10] We assume that the patient and dentist are distinct individuals.

The general definition of 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 the dentist domain, for example, it seems reasonable to assert conditional independence of the variables *Toothache* and *Catch*, given *Cavity*:

$$\mathbf{P}(Toothache, Catch|Cavity) = \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity) . \quad (13.14)$$

Notice that this assertion is somewhat stronger than Equation (13.13), which asserts independence only for specific values of *Toothache* and *Catch*. As with absolute independence in Equation (13.8), the equivalent forms

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

can also be used.

Section 13.5 showed that absolute independence assertions allow a decomposition of the full joint distribution into much smaller pieces. It turns out that the same is true for conditional independence assertions. For example, given the assertion in Equation (13.14), we can derive a decomposition as follows:

$$\begin{aligned}
&\mathbf{P}(Toothache, Catch, Cavity) \\
&= \mathbf{P}(Toothache, Catch|Cavity)\mathbf{P}(Cavity) \quad \text{(product rule)} \\
&= \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity) \quad \text{[using (13.14)]}.
\end{aligned}$$

In this way, the original large table is decomposed into three smaller tables. The original table has seven independent numbers ($2^3 - 1$, because the numbers must sum to 1). The smaller tables contain five independent numbers ($2 \times (2^1 - 1)$ for each conditional probability distribution and $2^1 - 1$ for the prior on *Cavity*). This might not seem to be a major triumph, but the point is that, for n symptoms that are all conditionally independent given *Cavity*, the size of the representation grows as $O(n)$ instead of $O(2^n)$. Thus, *conditional independence assertions can allow probabilistic systems to scale up; moreover, they are much more commonly available than absolute independence assertions.* Conceptually, *Cavity* **separates** *Toothache* and *Catch* because it is a direct cause of both of them. The decomposition of large probabilistic domains into weakly connected subsets via conditional independence is one of the most important developments in the recent history of AI.

SEPARATION

The dentistry example illustrates a commonly occurring pattern in which a single cause directly influences a number of effects, all of which are conditionally independent, given the cause. The full joint distribution can be written as

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

NAIVE BAYES

Such a probability distribution is called a **naive Bayes** model—"naive" because it is often used (as a simplifying assumption) in cases where the "effect" variables are *not* conditionally independent given the cause variable. (The naive Bayes model is sometimes called a **Bayesian classifier**, a somewhat careless usage that has prompted true Bayesians to call it the **idiot Bayes** model.) In practice, naive Bayes systems can work surprisingly well, even when the independence assumption is not true. Chapter 20 describes methods for learning naive Bayes distributions from observations.

IDIOT BAYES

13.7 THE WUMPUS WORLD REVISITED

We can combine many of the ideas in this chapter to solve probabilistic reasoning problems in the wumpus world. (See Chapter 7 for a complete description of the wumpus world.) Uncertainty arises in the wumpus world because the agent's sensors give only partial, local information about the world. For example, Figure 13.6 shows a situation in which each of the three reachable squares—[1,3], [2,2], and [3,1]—might contain a pit. Pure logical inference can conclude nothing about which square is most likely to be safe, so a logical agent might be forced to choose randomly. We will see that a probabilistic agent can do much better than the logical agent.

Our aim will be to calculate the probability that each of the three squares contains a pit. (For the purposes of this example, we will ignore the wumpus and the gold.) The relevant properties of the wumpus world are that (1) a pit causes breezes in all neighboring squares, and (2) each square other than [1,1] contains a pit with probability 0.2. The first step is to identify the set of random variables we need:

- As in the propositional logic case, we want one Boolean variable P_{ij} for each square, which is true iff square $[i, j]$ actually contains a pit.
- We also have Boolean variables B_{ij} that are true iff square $[i, j]$ is breezy; we include these variables only for the observed squares—in this case, [1,1], [1,2], and [2,1].

The next step is to specify the full joint distribution, $\mathbf{P}(P_{1,1}, \ldots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1})$. Applying the product rule, we have

$$\mathbf{P}(P_{1,1}, \ldots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1}) = \mathbf{P}(B_{1,1}, B_{1,2}, B_{2,1} \mid P_{1,1}, \ldots, P_{4,4})\mathbf{P}(P_{1,1}, \ldots, P_{4,4}) .$$

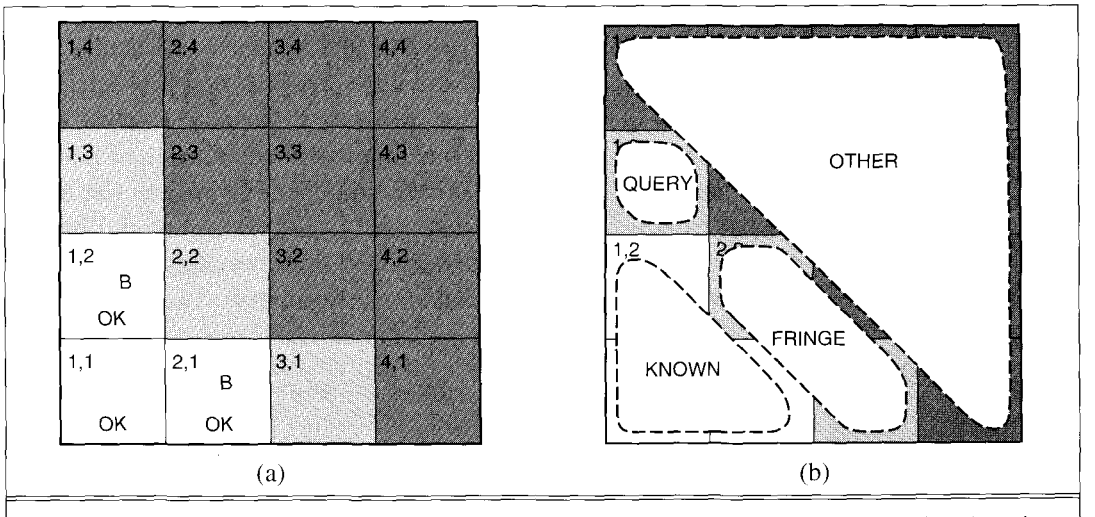

Figure 13.6 (a) After finding a breeze in both [1,2] and [2,1], the agent is stuck—there is no safe place to explore. (b) Division of the squares into *Known*, *Fringe*, and *Other*, for a query about [1,3].

This decomposition makes it very easy to see what the joint probability values should be. The first term is the conditional probability of a breeze configuration, given a pit configuration; this is 1 if the breezes are adjacent to the pits and 0 otherwise. The second term is the prior probability of a pit configuration. Each square contains a pit with probability 0.2, independently of the other squares; hence,

$$\mathbf{P}(P_{1,1},\ldots,P_{4,4}) = \prod_{i,j=1,1}^{4,4} \mathbf{P}(P_{i,j}) \; . \tag{13.15}$$

For a configuration with n pits, this is just $0.2^n \times 0.8^{16-n}$.

In the situation in Figure 13.6(a), the evidence consists of the observed breeze (or its absence) in each square that is visited, combined with the fact that each such square contains no pit. We'll abbreviate these facts as $b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$ and $known = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$. We are interested in answering queries such as $\mathbf{P}(P_{1,3}|known, b)$: how likely is it that [1,3] contains a pit, given the observations so far?

To answer this query, we can follow the standard approach suggested by Equation (13.6) and implemented in the ENUMERATE-JOINT-ASK, namely, summing over entries from the full joint distribution. Let $Unknown$ be a composite variable consisting of the $P_{i,j}$ variables for squares other than the $Known$ squares and the query square [1,3]. Then, by Equation (13.6), we have

$$\mathbf{P}(P_{1,3}|known, b) = \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b) \; .$$

The full joint probabilities have already been specified, so we are done—that is, unless we care about computation. There are 12 unknown squares; hence the summation contains $2^{12} = 4096$ terms. In general, the summation grows exponentially with the number of squares.

Intuition suggests that we are missing something here. Surely, one might ask, aren't the other squares irrelevant? The contents of [4,4] don't affect whether [1,3] has a pit! Indeed, this intuition is correct. Let $Fringe$ be the variables (other than the query variable) that are adjacent to visited squares, in this case just [2,2] and [3,1]. Also, let $Other$ be the variables for the other unknown squares; in this case, there are 10 other squares, as shown in Figure 13.6(b). The key insight is that the observed breezes are *conditionally independent* of the other variables, given the known, fringe, and query variables. The rest is, as they say, a small matter of algebra.

To use the insight, we manipulate the query formula into a form in which the breezes are conditioned on all the other variables, and then we simplify using conditional independence:

$$\begin{aligned}
&\mathbf{P}(P_{1,3}|known, b) \\
&= \alpha \sum_{unknown} \mathbf{P}(b|P_{1,3}, known, unknown)\mathbf{P}(P_{1,3}, known, unknown) \\
&\qquad\qquad \text{(by the product rule)} \\
&= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe, other)\mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe)\mathbf{P}(P_{1,3}, known, fringe, other) \; ,
\end{aligned}$$

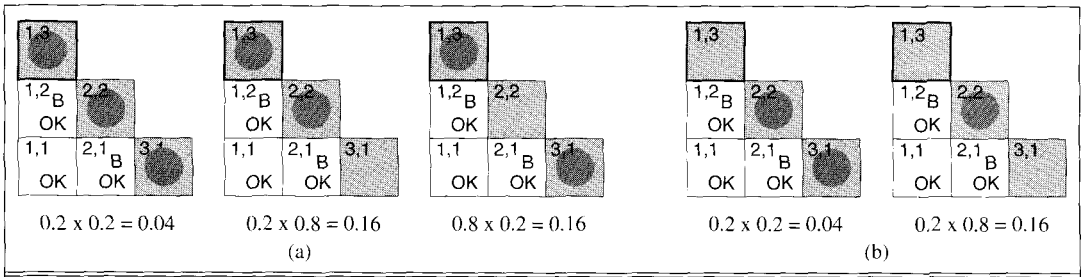

Figure 13.7 Consistent models for the fringe variables $P_{2,2}$ and $P_{3,1}$, showing $P(fringe)$ for each model: (a) three models with $P_{1,3} = true$ showing two or three pits, and (b) two models with $P_{1,3} = false$ showing one or two pits.

where the final step uses conditional independence. Now, the first term in this expression does not depend on the other variables, so we can move the summation inwards:

$$\begin{aligned}&\mathbf{P}(P_{1,3}|known, b)\\ &= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3}, known, fringe, other) \ .\end{aligned}$$

By independence, as in Equation (13.15), the prior term can be factored, and then the terms can be reordered:

$$\begin{aligned}&\mathbf{P}(P_{1,3}|known, b)\\ &= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3})P(known)P(fringe)P(other)\\ &= \alpha P(known)\mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe)P(fringe) \sum_{other} P(other)\\ &= \alpha' \mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe)P(fringe) \ ,\end{aligned}$$

where the last step folds $P(known)$ into the normalizing constant and uses the fact that $\sum_{other} P(other)$ equals 1.

Now, there are just four terms in the summation over the fringe variables $P_{2,2}$ and $P_{3,1}$. The use of independence and conditional independence has completely eliminated the other squares from consideration. Notice that the expression $\mathbf{P}(b|known, P_{1,3}, fringe)$ is 1 when the fringe is consistent with the breeze observations and 0 otherwise. Thus, for each value of $P_{1,3}$, we sum over the *logical models* for the fringe variables that are consistent with the known facts. (Compare with the enumeration over models in Figure 7.5.) The models and their associated prior probabilities—$P(fringe)$—are shown in Figure 13.7. We have

$$\mathbf{P}(P_{1,3}|known, b) = \alpha' \langle 0.2(0.04 + 0.16 + 0.16), \ 0.8(0.04 + 0.16)\rangle \approx \langle 0.31, 0.69\rangle \ .$$

That is, [1,3] (and [3,1] by symmetry) contains a pit with roughly 31% probability. A similar calculation, which the reader might wish to perform, shows that [2,2] contains a pit with roughly 86% probability. The wumpus agent should definitely avoid [2,2]!

What this section has shown is that even seemingly complicated problems can be formulated precisely in probability theory and solved using simple algorithms. To get *efficient*

solutions, independence and conditional independence relationships can be used to simplify the summations required. These relationships often correspond to our natural understanding of how the problem should be decomposed. In the next chapter, we will develop formal representations for such relationships as well as algorithms that operate on those representations to perform probabilistic inference efficiently.

13.8 SUMMARY

This chapter has argued that probability is the right way to reason about uncertainty.

- Uncertainty arises because of both laziness and ignorance. It is inescapable in complex, dynamic, or inaccessible worlds.
- Uncertainty means that many of the simplifications that are possible with deductive inference are no longer valid.
- Probabilities express the agent's inability to reach a definite decision regarding the truth of a sentence. Probabilities summarize the agent's beliefs.
- Basic probability statements include **prior probabilities** and **conditional probabilities** over simple and complex propositions.
- The **full joint probability distribution** specifies the probability of each complete assignment of values to random variables. It is usually too large to create or use in its explicit form.
- The axioms of probability constrain the possible assignments of probabilities to propositions. An agent that violates the axioms will behave irrationally in some circumstances.
- When the full joint distribution is available, it can be used to answer queries simply by adding up entries for the atomic events corresponding to the query propositions.
- **Absolute independence** between subsets of random variables might allow the full joint distribution to be factored into smaller joint distributions. This could greatly reduce complexity, but seldom occurs in practice.
- **Bayes' rule** allows unknown probabilities to be computed from known conditional probabilities, usually in the causal direction. Applying Bayes' rule with many pieces of evidence will in general run into the same scaling problems as does the full joint distribution.
- **Conditional independence** brought about by direct causal relationships in the domain might allow the full joint distribution to be factored into smaller, conditional distributions. The **naive Bayes** model assumes the conditional independence of all effect variables, given a single cause variable, and grows linearly with the number of effects.
- A wumpus-world agent can calculate probabilities for unobserved aspects of the world and use them to make better decisions than a purely logical agent makes.

BIBLIOGRAPHICAL AND HISTORICAL NOTES

Although games of chance date back at least to around 300 B.C., the mathematical analysis of odds and probability appears to be much more recent. Some work done by Mahaviracarya in India is dated to roughly the ninth century A.D. In Europe, the first attempts date only to the Italian Renaissance, beginning around 1500 A.D. The first significant systematic analyses were produced by Girolamo Cardano around 1565, but they remained unpublished until 1663. By that time, the discovery by Blaise Pascal (in correspondence with Pierre Fermat in 1654) of a systematic way of calculating probabilities had for the first time established probability as a mathematical discipline. The first published textbook on probability was *De Ratiociniis in Ludo Aleae* (Huygens, 1657). Pascal also introduced conditional probability, which is covered in Huygens's textbook. The Rev. Thomas Bayes (1702–1761) introduced the rule for reasoning about conditional probabilities that was named after him. It was published posthumously (Bayes, 1763). Kolmogorov (1950, first published in German in 1933) presented probability theory in a rigorously axiomatic framework for the first time. Rényi (1970) later gave an axiomatic presentation that took conditional probability, rather than absolute probability, as primitive.

Pascal used probability in ways that required both the objective interpretation, as a property of the world based on symmetry or relative frequency, and the subjective interpretation, based on degree of belief—the former in his analyses of probabilities in games of chance, the latter in the famous "Pascal's wager" argument about the possible existence of God. However, Pascal did not clearly realize the distinction between these two interpretations. The distinction was first drawn clearly by James Bernoulli (1654–1705).

Leibniz introduced the "classical" notion of probability as a proportion of enumerated, equally probable cases, which was also used by Bernoulli, although it was brought to prominence by Laplace (1749–1827). This notion is ambiguous between the frequency interpretation and the subjective interpretation. The cases can be thought to be equally probable either because of a natural, physical symmetry between them, or simply because we do not have any knowledge that would lead us to consider one more probable than another. The use of this latter, subjective consideration to justify assigning equal probabilities is known as the *principle of indifference* (Keynes, 1921).

The debate between objectivists and subjectivists became sharper in the 20th century. Kolmogorov (1963), R. A. Fisher (1922), and Richard von Mises (1928) were advocates of the relative frequency interpretation. Karl Popper's (1959, first published in German in 1934) "propensity" interpretation traces relative frequencies to an underlying physical symmetry. Frank Ramsey (1931), Bruno de Finetti (1937), R. T. Cox (1946), Leonard Savage (1954), and Richard Jeffrey (1983) interpreted probabilities as the degrees of belief of specific individuals. Their analyses of degree of belief were closely tied to utilities and to behavior—specifically, to the willingness to place bets. Rudolf Carnap, following Leibniz and Laplace, offered a different kind of subjective interpretation of probability—not as any actual individual's degree of belief, but as the degree of belief that an idealized individual *should* have in a particular proposition a, given a particular body of evidence $\mathbf{e}$. Carnap attempted to go further

CONFIRMATION

INDUCTIVE LOGIC

than Leibniz or Laplace by making this notion of degree of **confirmation** mathematically precise, as a logical relation between a and **e**. The study of this relation was intended to constitute a mathematical discipline called **inductive logic**, analogous to ordinary deductive logic (Carnap, 1948, 1950). Carnap was not able to extend his inductive logic much beyond the propositional case, and Putnam (1963) showed that some fundamental difficulties would prevent a strict extension to languages capable of expressing arithmetic.

The question of reference classes is closely tied to the attempt to find an inductive logic. The approach of choosing the "most specific" reference class of sufficient size was formally proposed by Reichenbach (1949). Various attempts have been made, notably by Henry Kyburg (1977, 1983), to formulate more sophisticated policies in order to avoid some obvious fallacies that arise with Reichenbach's rule, but such approaches remain somewhat *ad hoc*. More recent work by Bacchus, Grove, Halpern, and Koller (1992) extends Carnap's methods to first-order theories, thereby avoiding many of the difficulties associated with the straightfoward reference-class method..

Bayesian probabilistic reasoning has been used in AI since the 1960s, especially in medical diagnosis. It was used not only to make a diagnosis from available evidence, but also to select further questions and tests using the theory of information value (Section 16.6) when available evidence was inconclusive (Gorry, 1968; Gorry *et al.*, 1973). One system outperformed human experts in the diagnosis of acute abdominal illnesses (de Dombal *et al.*, 1974). These early Bayesian systems suffered from a number of problems, however. Because they lacked any theoretical model of the conditions they were diagnosing, they were vulnerable to unrepresentative data occurring in situations for which only a small sample was available (de Dombal *et al.*, 1981). Even more fundamentally, because they lacked a concise formalism (such as the one to be described in Chapter 14) for representing and using conditional independence information, they depended on the acquisition, storage, and processing of enormous tables of probabilistic data. Because of these difficulties, probabilistic methods for coping with uncertainty fell out of favor in AI from the 1970s to the mid-1980s. Developments since the late 1980s are described in the next chapter.

The naive Bayes representation for joint distributions has been studied extensively in the pattern recognition literature since the 1950s (Duda and Hart, 1973). It has also been used, often unwittingly, in text retrieval, beginning with the work of Maron (1961). The probabilistic foundations of this technique, described further in Exercise 13.18, were elucidated by Robertson and Sparck Jones (1976). Domingos and Pazzani (1997) provide an explanation for the surprising success of naive Bayesian reasoning even in domains where the independence assumptions are clearly violated.

There are many good introductory textbooks on probability theory, including those by Chung (1979) and Ross (1988). Morris DeGroot (1989) offers a combined introduction to probability and statistics from a Bayesian standpoint, as well as a more advanced text (1970). Richard Hamming's (1991) textbook gives a mathematically sophisticated introduction to probability theory from the standpoint of a propensity interpretation based on physical symmetry. Hacking (1975) and Hald (1990) cover the early history of the concept of probability. Bernstein (1996) gives an entertaining popular account of the story of risk.

EXERCISES

13.1 Show from first principles that $P(a|b \wedge a) = 1$.

13.2 Using the axioms of probability, prove that any probability distribution on a discrete random variable must sum to 1.

13.3 Would it be rational for an agent to hold the three beliefs $P(A) = 0.4$, $P(B) = 0.3$, and $P(A \vee B) = 0.5$? If so, what range of probabilities would be rational for the agent to hold for $A \wedge B$? Make up a table like the one in Figure 13.2, and show how it supports your argument about rationality. Then draw another version of the table where $P(A \vee B) = 0.7$. Explain why it is rational to have this probability, even though the table shows one case that is a loss and three that just break even. (*Hint:* what is Agent 1 committed to about the probability of each of the four cases, especially the case that is a loss?)

13.4 This question deals with the properties of atomic events, as discussed on page 468.

a. Prove that the disjunction of all possible atomic events is logically equivalent to $true$. [*Hint:* Use a proof by induction on the number of random variables.]

b. Prove that any proposition is logically equivalent to the disjunction of the atomic events that entail its truth.

13.5 Consider the domain of dealing 5-card poker hands from a standard deck of 52 cards, under the assumption that the dealer is fair.

a. How many atomic events are there in the joint probability distribution (i.e., how many 5-card hands are there)?

b. What is the probability of each atomic event?

c. What is the probability of being dealt a royal straight flush? Four of a kind?

13.6 Given the full joint distribution shown in Figure 13.3, calculate the following:

a. $P(toothache)$

b. $\mathbf{P}(Cavity)$

c. $\mathbf{P}(Toothache|cavity)$

d. $\mathbf{P}(Cavity|toothache \vee catch)$.

13.7 Show that the three forms of independence in Equation (13.8) are equivalent.

13.8 After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease and that the test is 99% accurate (i.e., the probability of testing positive when you do have the disease is 0.99, as is the probability of testing negative when you don't have the disease). The good news is that this is a rare disease, striking only 1 in 10,000 people of your age. Why is it good news that the disease is rare? What are the chances that you actually have the disease?

13.9 It is quite often useful to consider the effect of some specific propositions in the context of some general background evidence that remains fixed, rather than in the complete absence of information. The following questions ask you to prove more general versions of the product rule and Bayes' rule, with respect to some background evidence **e**:

a. Prove the conditionalized version of the general product rule:

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

b. Prove the conditionalized version of Bayes' rule in Equation (13.10).

13.10 Show that the statement

$$\mathbf{P}(A, B|C) = \mathbf{P}(A|C)\mathbf{P}(B|C)$$

is equivalent to either of the statements

$$\mathbf{P}(A|B, C) = \mathbf{P}(A|C) \quad \text{and} \quad \mathbf{P}(B|A, C) = \mathbf{P}(B|C) \ .$$

13.11 Suppose you are given a bag containing n unbiased coins. You are told that $n - 1$ of these coins are normal, with heads on one side and tails on the other, whereas one coin is a fake, with heads on both sides.

a. Suppose you reach into the bag, pick out a coin uniformly at random, flip it, and get a head. What is the (conditional) probability that the coin you chose is the fake coin?

b. Suppose you continue flipping the coin for a total of k times after picking it and see k heads. Now what is the conditional probability that you picked the fake coin?

c. Suppose you wanted to decide whether the chosen coin was fake by flipping it k times. The decision procedure returns FAKE if all k flips come up heads, otherwise it returns NORMAL. What is the (unconditional) probability that this procedure makes an error?

13.12 In this exercise, you will complete the normalization calculation for the meningitis example. First, make up a suitable value for $P(S|\neg M)$, and use it to calculate unnormalized values for $P(M|S)$ and $P(\neg M|S)$ (i.e., ignoring the $P(S)$ term in the Bayes' rule expression). Now normalize these values so that they add to 1.

13.13 This exercise investigates the way in which conditional independence relationships affect the amount of information needed for probabilistic calculations.

a. Suppose we wish to calculate $P(h|e_1, e_2)$ and we have no conditional independence information. Which of the following sets of numbers are sufficient for the calculation?

(i) $\mathbf{P}(E_1, E_2), \mathbf{P}(H), \mathbf{P}(E_1|H), \mathbf{P}(E_2|H)$
(ii) $\mathbf{P}(E_1, E_2), \mathbf{P}(H), \mathbf{P}(E_1, E_2|H)$
(iii) $\mathbf{P}(H), \mathbf{P}(E_1|H), \mathbf{P}(E_2|H)$

b. Suppose we know that $\mathbf{P}(E_1|H, E_2) = \mathbf{P}(E_1|H)$ for all values of H, E_1, E_2. Now which of the three sets are sufficient?

13.14 Let X, Y, Z be Boolean random variables. Label the eight entries in the joint distribution $\mathbf{P}(X, Y, Z)$ as a through h. Express the statement that X and Y are conditionally

independent given Z as a set of equations relating a through h. How many *nonredundant* equations are there?

13.15 (Adapted from Pearl (1988).) Suppose you are a witness to a nighttime hit-and-run accident involving a taxi in Athens. All taxis in Athens are blue or green. You swear, under oath, that the taxi was blue. Extensive testing shows that, under the dim lighting conditions, discrimination between blue and green is 75% reliable. Is it possible to calculate the most likely color for the taxi? (*Hint:* distinguish carefully between the proposition that the taxi *is* blue and the proposition that it *appears* blue.)

What about now, given that 9 out of 10 Athenian taxis are green?

13.16 (Adapted from Pearl (1988).) Three prisoners, *A*, *B*, and *C*, are locked in their cells. It is common knowledge that one of them will be executed the next day and the others pardoned. Only the governor knows which one will be executed. Prisoner *A* asks the guard a favor: "Please ask the governor who will be executed, and then take a message to one of my friends *B* or *C* to let him know that he will be pardoned in the morning." The guard agrees, and comes back later and tells *A* that he gave the pardon message to *B*.

What are *A*'s chances of being executed, given this information? (Answer this *mathematically*, not by energetic waving of hands.)

13.17 Write out a general algorithm for answering queries of the form $\mathbf{P}(Cause|\mathbf{e})$, using a naive Bayes distribution. You should assume that the evidence $\mathbf{e}$ may assign values to *any subset* of the effect variables.

13.18 Text categorization is the task of assigning a given document to one of a fixed set of categories, on the basis of the text it contains. Naive Bayes models are often used for this task. In these models, the query variable is the document category, and the "effect" variables are the presence or absence of each word in the language; the assumption is that words occur independently in documents, with frequencies determined by the document category.

a. Explain precisely how such a model can be constructed, given as "training data" a set of documents that have been assigned to categories.

b. Explain precisely how to categorize a new document.

c. Is the independence assumption reasonable? Discuss.

13.19 In our analysis of the wumpus world, we used the fact that each square contains a pit with probability 0.2, independently of the contents of the other squares. Suppose instead that exactly $N/5$ pits are scattered uniformly at random among the N squares other than [1,1]. Are the variables $P_{i,j}$ and $P_{k,l}$ still independent? What is the joint distribution $\mathbf{P}(P_{1,1}, \ldots, P_{4,4})$ now? Redo the calculation for the probabilities of pits in [1,3] and [2,2].