Introduction

This chapter introduces the ideas behind probabilistic reasoning, and in particular Bayes' Theorem. Thomas Bayes was an English mathematician and theologian who lived from 1702 to 1761. His theorem is used extensively today in dealing with situations that lack certainty. It explains joint probability distributions and goes on to explain Bayes' theorem, using two examples.

Probabilistic Reasoning

In this section, we will present a brief introduction to probability theory and the notation that is used to express it. Probability theory is used to discuss events, categories, and hypotheses about which there is not 100% certainty.

The notation that we saw in Chapter 7 for making and analyzing logical statements does not function in situations that are lacking **certainty**.

For example, we might write

$A \rightarrow B$

which means that if *A* is true, then *B* is true. If we are unsure whether A is true, then we cannot make use of this expression. In many real-world situations, it is very useful to be able to talk about things that lack certainty. For example, what will the weather be like tomorrow? We might formulate a very simple hypothesis based on general observation, such as "it is sunny only 10% of the time, and rainy 70% of the time." We can use a notation similar to that used for predicate calculus to express such statements:

P(S) = 0.1
P(R) = 0.7

The first of these statements says that the probability of *S* ("it is sunny") is 0.1. The second says that the probability of *R* is 0.7. Probabilities are always expressed as real numbers between 0 and 1. A probability of 0 means "definitely not" and a probability of 1 means "definitely so." Hence, *P(S)* = 1 means that it is always sunny.

Many of the operators and notations that are used in prepositional logic can also be used in probabilistic notation. For example, *P(¬S)* means "the probability that it is not sunny"; *P(S ∧ R)* means "the probability that it is both sunny and rainy."

P(A ∨ B), which means "the probability that either A is true or B is true," is defined by the following rule:

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

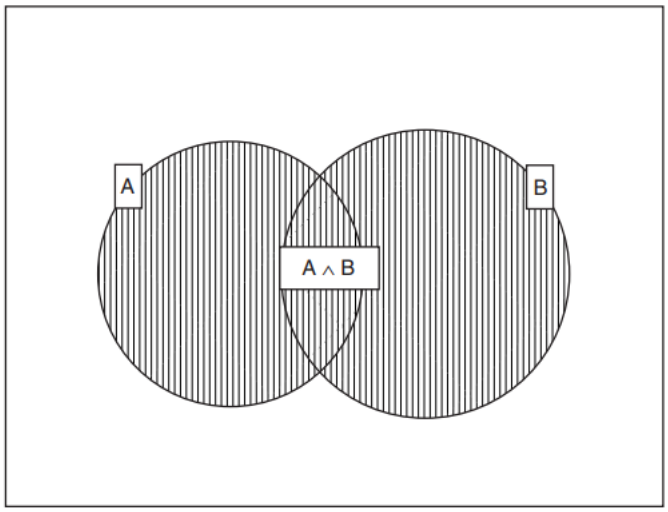

Figure 12.1
Illustrating the relationship between $A \wedge B$ and $A \vee B$

This rule can be seen to be true by examining the Venn diagram shown in Figure 12.1.

The notation P(B|A) can be read as "the probability of B, given A." This is known as conditional probability—it is conditional on A. In other words, it states the probability that B is true, given that we already know that A is true.

$P(B|A)$ is defined by the following rule:

$$P(B|A) = \frac{P(B \wedge A)}{P(A)}$$

Of course, this rule cannot be used in cases where $P(A) = 0$.

For example, let us suppose that the likelihood that it is both sunny and rainy at the same time is 0.01. Then we can calculate the probability that it is rainy, given that it is sunny as follows:

$$P(R|S) = \frac{P(R \wedge S)}{P(S)}$$

$$= \frac{0.01}{0.1}$$

$$= 0.1$$

Note that the probability that it is sunny given that it is overcast—$P(S|R)$— is different from this: 0.01/0.7 = 0.14; hence, $P(A|B) \neq P(B|A)$.

Joint Probability Distributions

A **joint probability distribution** (also known as a **joint**) can be used to represent the probabilities of combined statements, such as $A \wedge B$. For example, the following table shows a joint probability distribution of two variables, A and B:

	A	¬A
B	0.11	0.09
¬B	0.63	0.17

This shows, for example, that $P(A \wedge B) = 0.11$, and that $P(A \nearrow \neg B) = 0.63$. By summing these two values, we can find $P(A) = 0.11 + 0.63 = 0.74$. Similarly, $P(B) = 0.11 + 0.09 = 0.2$.

We can use this table to determine the probability of any logical combination of A and B.
For example, $P(A \vee B) = 0.11 + 0.09 + 0.63 = 0.83$. We could have obtained this result by noting that $P(\neg A \wedge \neg B) = 0.17$ and that $P(\neg A \wedge \neg B) = 1 — P(A \vee B) = 1 — 0.17 = 0.83$.

Similarly, we can determine conditional probabilities, such as P(B|A) using the following rule:

$$P(B|A) = \frac{P(B \wedge A)}{P(A)}$$

In this case, $P(B \wedge A) = 0.11$ and $P(A) = 0.11 + 0.63 = 0.74$, so $P(B|A) = 0.11 / 0.74 = 0.15$.

Calculations like this are easy when we use a joint probability of just two variables. Real-world problems will often involve much greater numbers of variables, and in these cases, drawing up probability distribution tables is clearly much less straightforward.

Bayes' Theorem

Bayes' theorem can be used to calculate the probability that a certain event will occur or that a certain proposition is true, given that we already know a related piece of information.

The theorem is stated as follows:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

P(B) is called the **prior probability** of *B*. *P(B|A)*, as well as being called the conditional probability, is also known as the **posterior probability** of *B*.

Let us briefly examine how Bayes' theorem is derived:

We can deduce a further equation from the rule given in Section 12.2 above. This rule is known as the **product rule**:

$P(A \wedge B) = P(A|B)P(B)$

Note that due to the commutativity of ∧, we can also write

$P(A \wedge B) = P(B|A)P(A)$

Hence, we can deduce:

$P(B|A)P(A) = P(A|B)P(B)$

This can then be rearranged to give Bayes' theorem:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

Example: Medical Diagnosis

Let us examine a simple example to illustrate the use of Bayes' theorem for the purposes of medical diagnosis.

When one has a cold, one usually has a high temperature (let us say, 80% of the time). We can use *A* to denote "I have a high temperature" and *B* to denote "I have a cold." Therefore, we can write this statement of posterior probability as

$P(A|B) = 0.8$

Note that in this case, we are using *A* and *B* to represent pieces of data that could each either be a hypothesis or a piece of evidence. It is more likely that we would use *A* as a piece of evidence to help us prove or disprove the hypothesis, *B*, but it could work equally well the other way around (at least, mathematically speaking).

Now, let us suppose that we also know that at any one time around 1 in every 10,000 people has a cold, and that 1 in every 1000 people has a high temperature. We can write these prior probabilities as

$P(A) = 0.001$
$P(B) = 0.0001$

Now suppose that you have a high temperature. What is the likelihood that you have a cold? This can be calculated very simply by using Bayes' theorem:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

$$= \frac{0.8 \cdot 0.0001}{0.001}$$

$$= 0.008$$

Hence, we have shown that just because you have a high temperature does not necessarily make it very likely that you have a cold—in fact, the chances that you have a cold are just 8 in 1000.

Bayes' theorem can be extended to express a conditional probability involving more than two variables as follows:

$$P(H|E_1 \wedge \ldots \wedge E_n) = \frac{P(E_1 \wedge \ldots \wedge E_n|H) \cdot P(H)}{P(E_1 \wedge \ldots \wedge E_n)}$$

Provided the *n* pieces of evidence $E_1 \ldots E_n$ are independent of each other, given the hypothesis H,[1] then this can be rewritten as follows:

$$P(H|E_1 \wedge \ldots \wedge E_n) = \frac{P(E_1|H) \cdot \ldots \cdot P(E_n|H) \cdot P(H)}{P(E_1 \wedge \ldots \wedge E_n)}$$

Example: Witness Reliability

Let us examine a further example. In the city of Cambridge, there are two taxi companies. One taxi company uses yellow taxis, and the other uses white taxis. The yellow taxi company has 90 cars, and the white taxi company has just 10 cars.

A hit-and-run incident has been reported, and an eye witness has stated that she is certain that the car was a white taxi.

So far, we have the following information:

P(Y) = 0.9 (the probability of any particular taxi being yellow)

P(W) = 0.1 (the probability of any particular taxi being white)

Let us further suppose that experts have asserted that given the foggy weather at the time of the incident, the witness had a 75% chance of correctly identifying the taxi.

Given that the lady has said that the taxi was white, what is the likelihood that she is right?

Let us denote by $P(C_W)$ the probability that the culprit was driving a white taxi and by $P(C_Y)$ the probability that it was a yellow car.

We will use $P(W_W)$ to denote the probability that the witness says she saw a white car and $P(W_Y)$ to denote that she says she saw a yellow car. (We assume the witness tells the truth!)

Now, if the witness really saw a yellow car, she would say that it was yellow 75% of the time, and if she says she saw a white car, she would say it was white 75% of the time. Hence, we now know the following:

$P(C_Y) = 0.9$

$P(C_W) = 0.1$

$P(W_W \mid C_W) = 0.75$

$P(W_Y \mid C_Y) = 0.75$

Hence, we can apply Bayes' theorem to find the probability, given that she is saying that the car was white, that she is correct:

$$P(C_W|W_W) = \frac{0.75 \cdot 0.1}{P(W_W)}$$

We now need to calculate $P(W_W)$—the prior probability that the lady would say she saw a white car.

Let us imagine that the lady is later shown a random sequence of 1000 cars. We expect 900 of these cars to be yellow and 100 of them to be white. The witness will misidentify 250 of the cars: Of the 900 yellow cars, she will incorrectly say that 225 are white. Of the 100 white cars, she will incorrectly say that 25 are yellow. Hence, in total, she will believe she sees 300 white cars—even though only 100 of them are really white. So, $P(W_W)$ is 300/1000 = 0.3.

We can now complete our equation to find $P(C_W|W_W)$:

$$P(C_W|W_W) = \frac{0.75 \cdot 0.1}{0.3}$$
$$= 0.25$$

In other words, if the lady says that the car was white, the probability that it was in fact white is only 0.25—it is three times more likely that it was actually yellow!

In this example, Bayes' theorem takes into account the actual number of each color of taxi in the city. If the witness had said she saw a yellow taxi, it would be very likely that she was right—but this is likely anyway because there are so many more yellow taxis than white taxis. If the witness were a

perfect observer who made no errors, then the probability $P(C_W|W_W)$ would, of course, be 1.

This example also helps to illustrate the fact that in many real-world situations we do have enough information to be able to use Bayes' theorem. It can look as though Bayes' theorem will apply only in contrived situations, but in fact it is usually the case that obtaining the data needed to use Bayes' theorem is easier than obtaining the posterior probability by other means. This is particularly true in cases where there are a large number of individuals being discussed.

Comparing Conditional Probabilities

In many situations, it can be useful to compare two probabilities. In particular, in making a diagnosis from a set of evidence, one will often have to choose from a number of possible hypotheses.

For example, let us extend the medical example given in Section 12.4.1. There we used A to represent the piece of evidence "I have a high temperature" and B to represent the hypothesis "I have a cold," where

P(A) = 0.001
P(B) = 0.0001
P(A|B) = 0.8

Let us further use C to represent the hypothesis "I have plague," where

P(C) = 0.000000001
P(A|C) = 0.99

In other words, it is highly unlikely for anyone to have plague, but if they do, they will almost certainly have a high temperature.

In this case, when carrying out a diagnosis of a patient that has a high temperature, it will be useful to determine which is the more likely hypothesis—*B* or *C*.

Bayes' theorem gives us the following:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

$$P(C|A) = \frac{P(A|C) \cdot P(C)}{P(A)}$$

Clearly, to find the more likely of *B* and *C*, given *A*, we can eliminate *P(A)* from these equations and can determine the **relative likelihood** of *B* and *C* as follows:

$$\frac{P(B|A)}{P(C|A)} = \frac{P(A|B) \cdot P(B)}{P(A|C) \cdot P(C)}$$

$$= \frac{0.8 \cdot 0.001}{0.95 \cdot 0.000000001}$$

$$= 842,105$$

Hence, it is hundreds of thousands of times more likely given that a patient has a high temperature that he has a cold than that he has plague.

Normalization

Normalization is the process whereby the posterior probabilities of a pair of variables are divided by a fixed value to ensure that they sum to 1.

This can be done by considering the following two equations:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

$$P(\neg B|A) = \frac{P(A|\neg B) \cdot P(\neg B)}{P(A)}$$

Given that *A* is true, *B* must either be true or false, which means that $P(B|A) + P(\neg B|A) = 1$.

Hence, we can add the two equations above to give

$$1 = \frac{P(A|B) \cdot P(B)}{P(A)} + \frac{P(A|\neg B) \cdot P(B)}{P(A)}$$

$$\therefore P(A) = P(A|B) \cdot P(B) + P(A|\neg B) \cdot P(\neg B)$$

Now we can replace *P(A)* in the equation for Bayes' theorem, to give

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|\neg B) \cdot P(\neg B)}$$

Hence, it is possible to use Bayes' theorem to obtain the conditional probability *P(B|A)* without needing to know or calculate *P(A)*, providing we can obtain *P(A|¬B)*. [*P(¬B)* is simply 1–*P(B)*].

This equation is often written as follows:

$$P(B|A) = \alpha \cdot P(A|B) \cdot P(B)$$

where α represents the normalizing constant:

$$\alpha = \frac{1}{P(A|B) \cdot P(B) + P(A|\neg B) \cdot P(\neg B)}$$

Let us examine our diagnosis example again. The facts we have are as follows:

P(A) = 0.001

P(B) = 0.0001

P(A|B) = 0.8

Let us now suppose that $P(A|\neg B) = 0.00099$. This conditional probability states the likelihood that a person will have a high temperature if she does not have a cold ($\neg B$). We can now thus use the following equation to calculate $P(B|A)$:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|\neg B) \cdot P(\neg B)}$$

$$= \frac{0.8 \cdot 0.0001}{0.8 \cdot 0.001 + 0.00099 \cdot 0.9999}$$

$$= \frac{0.00008}{0.001069901}$$

$$= 0.075$$

Similarly, we can calculate $P(\neg B|A)$:

$$P(\neg B|A) = \frac{P(A|\neg B) \cdot P(\neg B)}{P(A|\neg B) \cdot P(\neg B) + P(A|B) \cdot P(B)}$$

$$= \frac{0.00099 \cdot 0.9999}{0.00099 \cdot 0.9999 + 0.8 \cdot 0.0001}$$

$$= \frac{0.000989901}{0.001069901}$$

$$= 0.925$$

The net result of this normalization process has been to ensure that $P(B|A) + P(\neg B|A) = 1$. We could now carry out a similar process to calculate $P(C|A)$ and $P(\neg C|A)$, which would enable us to ensure that they also sum to 1.

Simple Bayesian Concept Learning

A very simple model for learning can be developed using Bayes' rule.

Throughout the above discussion we have been talking about probabilities of hypotheses or of specific pieces of evidence. To use probability theory in learning, it is useful to talk about the probability that some hypothesis is true, given a particular set of evidence. We can use the same notation for this, and write

P(H|E)

Hence, given a set of evidence, the learner can determine which hypothesis to believe in by identifying the posterior probability of each. Let us suppose that there are n possible hypotheses, H1 ... Hn. Hence, for each Hi

$$P(H_i|E) = \frac{P(E|H_i) \cdot P(H_i)}{P(E)}$$

So, the algorithm could calculate $P(H_i|E)$ for each possible hypothesis and select the one that has the highest probability. Similarly, the system could use this method to determine an action to take, where H_i is the hypothesis that the best action to take in the current situation is action A_i.

In fact, the formula above can be simplified in this situation: because $P(E)$ is **independent** of H_i, it will have the same value for each hypothesis. So, because we are simply looking for the hypothesis with the maximum posterior probability, we can eliminate $P(E)$ from the calculation and simply aim to maximize the following value:

$$P(E|H_i) \cdot P(H_i)$$

In fact, if we assume that all hypotheses are equally likely, given no additional information (i.e., $P(H_i) = P(H_j)$ for any i and j), we can in fact reduce this further and simply choose the hypothesis for whom the value $P(E|H_i)$ is the highest. This value is known as the **likelihood** of the evidence E, given hypothesis H_i. Of course, by learning from observations what the prior probabilities are of each of the hypotheses, more accurate results can be obtained, but the simpler formula is more efficient in calculation time.

Recall the discussion from Chapter 7 of abduction and inductive reasoning. These are really a form of learning: by observing the events that occur, we are able to make reasonable guesses about future events, and these guesses can often guide our actions. For example, if a robot observed that every time it heard a particular noise, an enemy robot appeared, it might learn to hide when it heard that noise. In doing so, it is learning from experience and using Bayesian reasoning to decide upon the correct course of action. The robot is not using rules of logical deduction, such as modus ponens, which was explained in Chapter 7, but a rather more probabilistic form of reasoning, along the lines of "I have noticed in the past that when this noise occurs, an enemy appears. I have also noticed in the past that if I do not hide when an enemy appears, I get hurt by the enemy. Hence, I should probably hide when I hear the noise."

Humans use learning of this kind all the time, and it is essential for learning in situations in which there is very little certainty, such as the real world.

Bayesian Belief Networks

The concept of **dependence** is very important in probability theory. Two events, A and B, are **independent** if the likelihood of occurrence of A is entirely unrelated to whether or not B occurs.

For example, in tossing two coins, the likelihood that the first coin will come up heads and the likelihood that the second coin will come up heads are two independent probabilities because neither one depends on the other.

If A and B are independent, then the probability that A and B will both occur can be calculated very simply:

$$P(A \wedge B) = P(A).P(B)$$

We know that this equation does not hold if A depends on B because we have already seen the following equation:

$$P(B|A) = \frac{P(B \wedge A)}{P(A)}$$

By comparing these two equations, we can see that A and B are independent if $P(B|A) = P(B)$. In other words, the likelihood of B is unaffected by whether or not A occurs. B is independent of A. If B is dependent on A, then $P(B|A)$ will be different from $P(B)$.

These relationships can be expressed extremely succinctly in a belief net- work, such as the one shown in Figure.

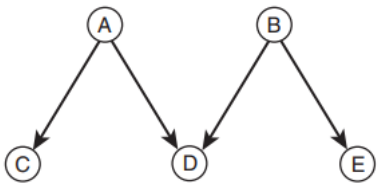

Figure 12.2
A simple belief network

A **Bayesian belief network** is an acyclic directed graph, where the nodes in the graph represent evidence or hypotheses, and where an arc that connects two nodes represents a dependence between those two nodes. The belief network in Figure 12.2 contains five nodes that represent pieces of evidence (A and B) and three hypotheses (C, D, and E). The arcs between these nodes represent the interdependence of the hypotheses. According to this diagram, C and D are both dependent on A, and D and E are both dependent on B. Two nodes that do not have an arc between them are independent of each other. For example, B is independent of A.

Each node in the network has a set of probabilities associated with it, based on the values of the nodes on which it is dependent. Hence, *A* and *B* both have just prior probabilities, *P(A)* and *P(B)*, because they are not dependent on any other nodes. *C* and *E* are each dependent on just one other node. Hence, for example, *P(C)* must be represented in the two cases—*A* is true and *A* is false. *P(D)* must be represented in four cases, depending on the values of *A* and *B*.

For example, the following conditional probabilities might be used in the network shown in Figure 12.2:

P(A) = 0.1
P(B) = 0.7
P(C|A) = 0.2
P(C|¬A) = 0.4
P(D|A ∧ B) = 0.5
P(D|A ∧ ¬B) = 0.4
P(D|¬A ∧ B) = 0.2
P(D|¬A ∧ ¬B) = 0.0001
P(E|B) = 0.2
P(E|¬B) = 0.1

The above list of probabilities, combined with the diagram shown in Figure 12.2, represent a complete (rather simple) Bayesian belief network. The network states beliefs about a set of hypotheses or pieces of evidence and the ways that they interact.

These probabilities can also be expressed in the form of **conditional probability tables**, as follows:

P(A)
0.1

P(B)
0.7

A	P(C)
true	0.2
false	0.4

B	P(E)
true	0.2
false	0.1

A	B	P(D)
true	true	0.5
true	false	0.4
false	true	0.2
false	false	0.0001

Compare these tables with the logical truth tables described in Chapter 7. In those tables, a logical value (*true* or *false*) was given for a variable that depended on the values of one or more other variables. Hence, a conditional probability table is very similar to a truth table, except that it expresses the probability of one variable, given the truth values of one or more other variables.

A joint probability can be calculated from the Bayesian belief network using the definition of conditional probability:

$$P(B|A) = \frac{P(B \wedge A)}{P(A)}$$

Hence,

$$P(A,B,C,D,E) = P(E|A,B,C,D) \cdot P(A,B,C,D)$$

We can apply this rule recursively to obtain

$$P(A,B,C,D,E) = P(E|A,B,C,D,) \cdot P(D|A,B,C,) \cdot P(C|A,B) \cdot P(B|A) \cdot P(A)$$

In fact, the nature of our belief network allows us to simplify this expression, and because we know that, for example, E is not dependent on *A, C,* or *D*, we can reduce *P(E|A,B,C,D)* to *P(E|B).*

$$P(A,B,C,D,E) = P(E|B) \cdot P(D|A,B) \cdot P(C|A) \cdot P(B) \cdot P(A)$$

We have now greatly reduced the complexity of the calculation needed to compute the joint probability. This has only been possible due to the way in which the nodes were ordered in the original expression. For example, if we used the same method blindly on the expression

P(E,D,C,B,A)

we would be left with the following expression:

$$P(E,D,C,B,A) = P(A|E,D,C,B) \cdot P(B|E,D,C) \cdot P(C|E,D) \cdot P(D|E) \cdot P(E)$$

This is not correct because E is dependent on B, and so we need to include P(E|B). Similarly, D is dependent on A and B, which is not reflected in this expression.

In other words, to calculate the joint probability, the nodes must be ordered in the expression in such a way that if a node *X* is dependent on another node *Y*, then *X* appears before *Y* in the joint. Hence, we could have used any ordering in which *A* and *B* appear before *C*, *D*, and *E*; *B,A,E,D,C* would have worked equally well, for example.

As a result of this, when constructing a Bayesian belief network, it is essential that the graph be constructed in the correct order—in other words, in an order such that the connections between nodes makes logical sense. This usually means starting with causes and then adding the events they cause, and then treating those events as causes, and adding any further events they cause.

The nature of Bayesian belief networks means that in general they are an efficient way of storing a joint probability distribution. The network does not store the conditional probability *P(X|Y)* if *X* and *Y* are independent of each other, given the parents of *X*. In the network shown in Figure 12.2, for example, this means that *P(E|A)* does not need to be stored.

Example: Life at College

Let us examine the simple Bayesian belief network shown in Figure 12.3. In Figure 12.3, the five nodes represent the following statements:

C = that you will go to college
S = that you will study
P = that you will party
E = that you will be successful in your exams
F = that you will have fun

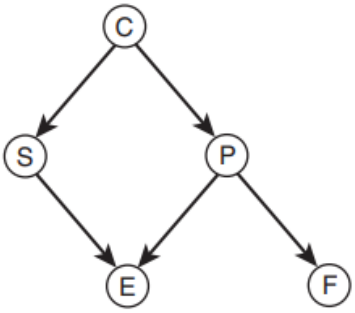

Figure 12.3
A Bayesian network to represent activities at college

This network shows us at a glance that if you go to college, this will affect the likelihood that you will study and the likelihood that you will party. Studying and partying affect your chances of exam success, and partying affects your chances of having fun.

To complete the Bayesian belief network, we need to include the conditional probability tables. Let us define these as follows:

P(C)
0.2

C	P(S)
true	0.8
false	0.2

C	P(P)
true	0.6
false	0.5

S	P	P(E)
true	true	0.6
true	false	0.9
false	true	0.1
false	false	0.2

P	P(F)
true	0.9
false	0.7

Note that according to this belief network there is a dependence between F and C, but because it is not a direct dependence, no information needs to be stored about it.

These conditional probability tables give us all the information we need to carry out any reasoning about this particular domain. For example, we can clearly obtain values such as P(¬C) by using the fact that

$$P(\neg C) = 1 - P(C) = 1 - 0.2 = 0.8.$$

We can use the network to determine conditional probabilities, such as $P(F|P)$ by observing that in the final table, if P is true, then $P(F) = 0.9$. Hence, $P(F|P) = 0.9$.

The joint probability distribution for this domain represents the entire state of the domain. We can represent such a state using the notation as used in the following example:

P(C = true, S = true, P = false, E = true, F = false)

We can simplify this notation as follows:

P(C, S, ¬P, E, ¬F)

This represents the probability that you will go to college and that you will study and be successful in your exams, but will not party or have fun. This probability can be calculated using the following rule:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|E)$$

where E is the evidence on which each x_i is dependent—in other words, in the Bayesian belief network, E consists of the nodes that are parents of x_i. For example, using the network shown in Figure 12.3, we can calculate the following probability:

$$P(C,S,\neg P,E,\neg F) = P(C) \cdot P(S|C) \cdot P(\neg P|C) \cdot P(E|S \wedge \neg P) \cdot P(\neg F|\neg P)$$
$$= 0.2 \cdot 0.8 \cdot 0.4 \cdot 0.9 \cdot 0.3$$
$$= 0.01728$$

Hence, for S we need to include in the product $P(S|C)$ because S is only dependent on C, and C is true in the situation we are examining. Similarly, for E we need to include $P(E|S \wedge \neg P)$ because E is dependent on S and on P, and S is true and P is not true in the scenario.

We can also calculate more complex conditional probabilities. In fact, this is an extremely simple process, due to the way in which the belief network has been created. For example, let us look at the following conditional probability:

P(E|F ∧ ¬P ∧ S ∧ C)

This is the probability that you will have success in your exams if you have fun and study at college, but don't party.

The assumption behind the Bayesian belief network is that because there is no direct connection between *E* and *C*, *E* is independent of *C*, given *S* and *P*. In other words, if we wish to calculate the following:

$$P(E|C \wedge S \wedge P)$$

we can in fact drop C from this altogether, and simply obtain

$$P(E|S \wedge P) = 0.6$$

Similarly, the more complex conditional probability above can be simplified by dropping *F* and *C* to give

$$P(E|S \wedge \neg P) = 0.9$$

Hence, any calculation that we might need to make about this domain can be made simply using the conditional probability tables of the belief network.

Similarly, we can make diagnoses about your college life by determining posterior probabilities. For example, let us say that we know that you had fun and studied hard while at college and we know that you succeeded in your exams, but we want to know whether you partied or not.

Clearly, we know *C*, *S*, *E*, and *F*, but we do not know *P*. We need to deter- mine the most likely value for *P*. Hence, we can compare the values of the following two expressions:

$$\begin{aligned} P(C \wedge S \wedge P \wedge E \wedge F) &= P(C) \cdot P(S|C) \cdot P(P|C) \cdot P(E|S \wedge P) \cdot P(F|P) \\ &= 0.2 \cdot 0.8 \cdot 0.6 \cdot 0.6 \cdot 0.9 \\ &= 0.05184 \end{aligned}$$

$$\begin{aligned} P(C \wedge S \wedge \neg P \wedge E \wedge F) &= P(C) \cdot P(S|C) \cdot P(\neg P|C) \cdot P(E|S \wedge \neg P) \cdot P(F|\neg P) \\ &= 0.2 \cdot 0.8 \cdot 0.4 \cdot 0.9 \cdot 0.7 \\ &= 0.04032 \end{aligned}$$

Hence, it is slightly more likely that you *did* party while at college than that you did not.

Example: Chapter Dependencies

We will now examine a simple example of a slightly unusual Bayesian net- work. Rather than each node representing a hypothesis or a piece of diagnostic information, each node in the Bayesian network shown in Figure 12.4 represents a chapter of this book. The arcs between nodes represent the dependencies between chapters. For example, the network shows that if you plan to read Chapter 8, which covers logical proof by resolution, it is a good idea to have read Chapter 7 on propositional and predicate logic first.

To see this as a more standard belief network, we can consider each node to represent the likelihood that you have read a given chapter and that a dependency from Chapter 8 to Chapter 7, for example, represents the fact that, if you have read Chapter 8, it is likely that you have also read Chapter 7. For this network to be useful to you in deciding which order to read the chapters, you can think of the dependencies as being advice about whether you should read a particular chapter before reading another.

The Noisy-V Function

Thus far, we have assumed that the probabilities contained with a joint probability distribution are unrelated to each other, in the sense that they have been determined by observing the way in which events occur. In some situations, it can be possible to use the fact that events in a Bayesian belief network are related to each other by some kind of mathematical or logical relation.

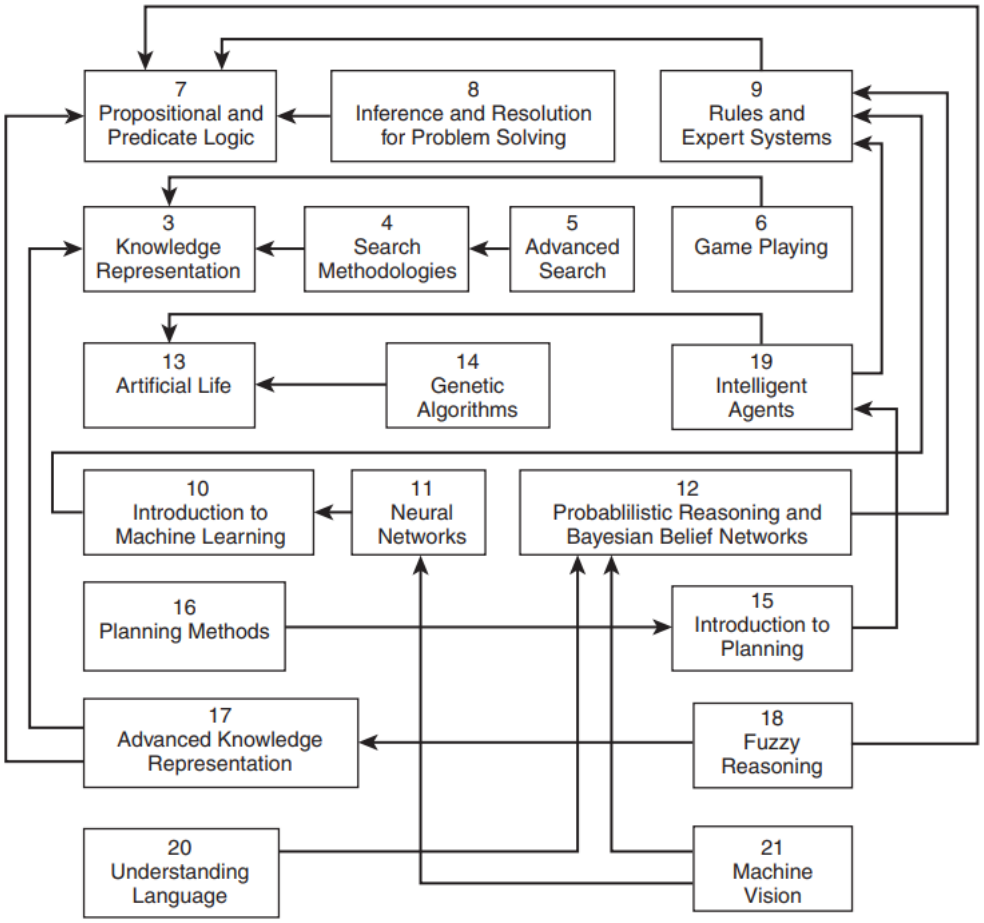

Figure 12.4

A Bayesian belief network that shows dependencies between chapters in this book

Clearly, logical relations such as ∧ and ∨, as defined in propositional logic, will not do because they do not provide a way to handle probabilities. **Fuzzy logic** (which is described in Chapter 18) could provide suitable relations. Another useful class of relations is **noisy logical relationships**.

Let us return to our diagnosis example. We use $P(A|B)$ to represent the probability that if one has a cold, then one will also have a high temperature. Similarly, we used $P(A|C)$ to represent the probability that if one has the plague, then one will also have a high temperature.

We have the following:

$$P(A|B) = 0.8$$

$$P(A|C) = 0.99$$

The noisy-∨ function is based on the assumption that the only possible causes of a high temperature are a cold and the plague (i.e., that $P(A|B \vee C) = 1$. Clearly this is not true for our example, but we can fix this by including a **leak node** in the network, which represents all other possible causes. Hence, we will further include $P(A|D) = 0.9$, where D is the leak node, which represents other causes of a high temperature.

Let us now define the **noise parameters** for these relationships. The noise parameters are simply defined as the conditional probabilities for ¬*A*, rather than for *A*, and can be obtained as follows:

$$P(\neg A|B) = 1 - P(A|B) = 0.2$$

$$P(\neg A|C) = 1 - P(A|C) = 0.01$$

$$P(\neg A|D) = 1 - P(A|D) = 0.1$$

A further assumption in using the noisy-∨ function is that the causes of a high temperature are independent of each other and, similarly, that the noise parameters (whatever it is that stops each illness from causing a high temperature) are independent of each other.

The noisy-∨ function for *B*, *C*, and *D* is defined as follows:

If *B*, *C*, and *D* are all false, then $P(A) = 0$. Otherwise, $P(\neg A)$ is equal to the product of the noise parameters for all the variables that are true. For example, if *B* is true and *C* and *D* are false, then $P(\neg A)$ is equal to the noise parameter for *B*, and so

$$P(A) = 1 - 0.2$$

$$= 0.8$$

If *C* and *D* are both true, and *B* is false, then $P(\neg A)$ is equal to the product of the noise parameters for *C* and *D*, and so

$$P(A) = 1 - (0.01 \times 0.1)$$

$$= 0.999$$

Now we can define the noisy-∨ function for our diagnosis example:

B	C	D	P(A)	P(¬A)
false	false	false	0	1
false	false	true	0.9	0.1
false	true	false	0.99	0.01
false	true	true	0.999	$0.01 \times 0.1 = 0.001$
true	false	false	0.8	0.2
true	false	true	0.98	$0.2 \times 0.1 = 0.02$
true	true	false	0.998	$0.2 \times 0.01 = 0.002$
true	true	true	0.9998	$0.2 \times 0.01 \times 0.1 = 0.0002$

Note that this noisy logical function is defined by just three conditional probabilities, as opposed to needing to store eight values. For Bayesian belief networks used in the real world with hundreds or even thousands of nodes, this can make a significant difference.

Bayes' Optimal Classifier

It is possible to use Bayesian reasoning to build a system that learns to classify data.

For example, let us suppose that for a given piece of data, *y*, there are five possible hypotheses, $H_1 \ldots H_5$, each of which assigns a classification to *y*. The classification, *c*, can be any value from a set *C*. For this example, let us assume that *C* consists of the values true and false.

Our classifier knows the posterior probabilities of each of the five hypotheses to be the following:

$P(H_1|x_1, \ldots, x_n) = 0.2$
$P(H_2|x_1, \ldots, x_n) = 0.3$
$P(H_3|x_1, \ldots, x_n) = 0.1$
$P(H_4|x_1, \ldots, x_n) = 0.25$
$P(H_5|x_1, \ldots, x_n) = 0.15$

where x_1 to x_n are the training data.

The probability that the new item of data, *y*, should be classified with classification c_j is defined by the following:

$$P(c_j|x_1 \ldots x_n) = \sum_{i=1}^{m} P(c_j|h_i) \cdot P(h_i|x_1 \ldots x_n)$$

where m is the number of available hypotheses, which in this case is 5. The optimal classification for *y* is the classification c_j for which $P(c_j|x_1 \ldots x_n)$ is the highest.

In our case, there are two classifications:

c_1 = true
c_2 = false

Let us suppose that hypotheses H_3 and H_5 each define *y* as true, while H_1, H_2, and H_4 define *y* as false.

Hence, we have the following posterior probabilities:

$P(false|H_1) = 0$	$P(true|H_1) = 1$
$P(false|H_2) = 0$	$P(true|H_2) = 1$
$P(false|H_3) = 1$	$P(true|H_3) = 0$
$P(false|H_4) = 0$	$P(true|H_4) = 1$
$P(false|H_5) = 1$	$P(true|H_5) = 0$

Thus, we can calculate the posterior probabilities for each of the two possible classifications for *y* as follows:

$$P(true|x_1 \ldots x_n) = \sum_{i=1}^{5} P(true|H_i) \cdot P(H_i|x_1 \ldots x_n)$$
$$= 0.2 + 0.3 + 0.25$$
$$= 0.75$$

$$P(false|x_1 \ldots x_n) = \sum_{i=1}^{5} P(false|H_i) \cdot P(H_i|x_1 \ldots x_n)$$
$$= 0.1 + 0.15$$
$$= 0.25$$

Hence, the optimal classification for *y* is true.

This method is known as an optimal classifier because it provides the best possible classification system. Another classification system, given the same data, can only hope to classify unseen data as well as this method—it can- not do better than the optimal classifier, on average.

The Naive Bayes Classifier

The naïve Bayes classifier is a simple but effective learning system. Each piece of data that is to be classified consists of a set of attributes, each of which can take on a number of possible values. The data are then classified into a single classification.

To identify the best classification for a particular instance of data ($d_1, \ldots, d_n$), the posterior probability of each possible classification is calculated:

$$P(c_i | d_1, \ldots, d_n)$$

where c_i is the ith classification, from a set of $|c|$ classifications.

The classification whose posterior probability is highest is chosen as the correct classification for this set of data. The hypothesis that has the highest posterior probability is often known as the maximum a posteriori, or MAP hypothesis. In this case, we are looking for the MAP classification.

To calculate the posterior probability, we can use Bayes' theorem and rewrite it as

$$\frac{P(d_1, \ldots, d_n | c_i) \cdot P(c_i)}{P(d_1, \ldots, d_n)}$$

Because we are simply trying to find the highest probability, and because $P(d_1, \ldots, d_n)$ is a constant independent of c_i, we can eliminate it and simply aim to find the classification c_i, for which the following is maximized:

$$P(d_1, \ldots, d_n | c_i) \cdot P(c_i)$$

The naïve Bayes classifier now assumes that each of the attributes in the data item is independent of the others, in which case $P(d_1, \ldots, d_n | c_i)$ can be rewritten and the following value obtained:

$$P(c_i) \cdot \prod_{j=1}^{n} P(d_j | c_i)$$

The naïve Bayes classifier selects a classification for a data set by finding the classification c_i for which the above calculation is a maximum.

For example, let us suppose that each data item consists of the attributes x, y, and z, where x, y, and z are each integers in the range 1 to 4.

The available classifications are A, B, and C. The example training data are as follows:

x	y	z	Classification
2	3	2	A
4	1	4	B
1	3	2	A
2	4	3	A
4	2	4	B
2	1	3	C
1	2	4	A
2	3	3	B
2	2	4	A
3	3	3	C
3	2	1	A
1	2	1	B
2	1	4	A
4	3	4	C
2	2	4	A

Hence, we have 15 pieces of training data, each of which has been classified. Eight of the training data are classified as *A*, four as *B*, and three as *C*.

Now let us suppose that we are presented with a new piece of data, which is (x = 2, y = 3, z = 4)

We need to obtain the posterior probability of each of the three classifications, given this piece of training data. Note that if we were to attempt to calculate $P(c_i|x = 2, y = 3, z = 4)$ without having made the simplifying step that we took above, in assuming that the attribute values are independent of each other, then we would need to have had many more items of training data to proceed. The naïve Bayes classifier requires far fewer items of training data.

We must now calculate each of the following:

$$P(A) \cdot P(x = 2|A) \cdot P(y = 3|A) \cdot P(z = 4|A)$$
$$P(B) \cdot P(x = 2|B) \cdot P(y = 3|B) \cdot P(z = 4|B)$$
$$P(C) \cdot P(x = 2|C) \cdot P(y = 3|C) \cdot P(z = 4|C)$$

Hence, for classification A, we obtain the following:

$$\frac{8}{15} \cdot \frac{5}{8} \cdot \frac{2}{8} \cdot \frac{4}{8} = 0.0417$$

This was calculated by observing that of the 15 items of training data, 8 were classified as *A*, and so $P(A) = 8/15$. Similarly, of the eight items of training data that were classified as *A*, five had $x = 2$, two had $y = 3$, and four had $z = 4$, and so $P(x = 2|A) = 5/8$, $P(y = 3|A) = 2/8$, and $P(z = 4|A) = 4/8$.

Similarly, we obtain the posterior probability for category *B*:

$$\frac{4}{15} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{2}{4} = 0.0083$$

and for category *C*:

$$\frac{3}{15} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} = 0.015$$

Hence, category A is chosen as the best category for this new piece of data, with category *C* as the second-best choice.

Let us now suppose that we are to classify the following piece of unseen data: (x = 1, y = 2, z = 2)

As before, we would calculate the posterior probability for *A*. However, in calculating the probabilities for *B* and *C*, we would have problems. In the case of category *B*, we would have

$$P(x = 1|B) = 1/5$$

$$P(y = 2|B) = 1/5$$

$$P(z = 2|B) = 0$$

Because there are no training examples with $z = 2$ that were classified as *B*, we have a posterior probability of 0. Similarly, for category *C*, we end up with

$$P(x = 1|C) = 0$$

$$P(y = 2|C) = 0$$

$$P(z = 2|C) = 0$$

In this case, we clearly must select category *A* as the best choice for the data, but it appears to be based on a fairly inadequate comparison because insufficient training data were available to properly compute posterior probabilities for the other categories.

This problem can be avoided by using the **m-estimate**, as follows:

We wish to determine the probability of a particular attribute value, given a particular classification, such as $P(x = 1|C)$. We will estimate this probability according to the following formula:

$$\frac{a + mp}{b + m}$$

where a = the number of training examples that exactly match our requirements (e.g., for P(x = 1|C), a is the number of training examples where x = 1 and that have been categorized as C. In this example, a is 0); b = the number of training examples that were classified in the current classification (i.e., for P(x = 1|C), b is the number of items of training data that were given classification C); p = an estimate of the probability that we are trying to obtain (usually this is obtained by simply assuming that each possible value is equally likely—hence, in our example, for P(x = 1|C), p = 1/4 = 0.25, as it would be for each of the other three possible values for x); m is a constant value, known as the **equivalent sample size**.

For example, let us use an equivalent sample size of 5 and determine the best classification for (x = 1, y = 2, z = 2):

For category A, we first need to calculate the probability for each of the three attributes.

Hence, for $x = 1$:

$$\frac{2 + 5/4}{8 + 5} = 0.25$$

For $y = 2$:

$$\frac{3 + 5/4}{8 + 5} = 0.33$$

For $z = 2$:

$$\frac{1 + 5/4}{8 + 5} = 0.17$$

Hence, the posterior probability estimate for A is

$$\frac{8}{15} \cdot 0.25 \cdot 0.33 \cdot 0.17 = 0.0076$$

Similarly, we can now obtain posterior probability estimates for categories *B* and *C*:

For category *B*, we obtain the following three probabilities:

$$\frac{1 + 5/4}{5 + 5} = 0.225, \quad \frac{2 + 5/4}{5 + 5} = 0.325, \quad \frac{0 + 5/4}{5 + 5} = 0.125$$

This gives us a posterior probability for category *B* as follows:

$$\frac{5}{15} \cdot 0.225 \cdot 0.325 \cdot 0.125 = 0.0091$$

Finally, the posterior probability for category *C* can be obtained. We note first that each of the three probabilities is the same because none of the attribute values occur in the training data with category *C*. Hence, the probability we use will be

$$\frac{0 + 5/4}{3 + 5} = 0.156$$

Hence, the posterior probability for category *C* is as follows:

$$\frac{3}{15} \cdot 0.156 \cdot 0.156 \cdot 0.156 = 0.0008$$

Hence, using this estimate for probability, we find that category *B* is the best match for the new data, and not category *A* as would have been obtained using the simpler probability estimates.

It is possible to further simplify the naïve Bayes classifier by considering the values to be position less within each item of data. In other words, when considering a new item of data, rather than assigning values to three attributes, we can simply think of the data as consisting of three values, whose order is arbitrary.

For example, consider the piece of new data (2, 3, 4).

In this case, we use the same method as before, but rather than considering the probability that, for example, $x = 2$ when an item is classified as A, we simply consider the probability that any attribute has value 2.

This simplified version of the naïve Bayes classifier is often used in text classification applications. Here, the categories are often simply "relevant" and "irrelevant," and the data to be classified consist of the words contained within textual documents. For example, an item of data might be ("the," "cat," "sat," "on," "the," "mat"). Training data would be presented in the form of a set of documents that has been preclassified as relevant and a set that has been preclassified as irrelevant. This form of textual analysis is dis- cussed in more detail in Chapter 20, which is concerned with information retrieval and natural language processing.

Collaborative Filtering

A further practical use for Bayesian reasoning is in **collaborative filtering**. Collaborative filtering is a technique that is increasingly used by online stores (such as Amazon.com) to provide plausible suggestions to customers based on their previous purchases. The idea behind collaborative filtering can be stated very simply: if we know that Anne and Bob both like items *A*, *B*, and *C*, and that Anne likes *D*, then it is reasonable to suppose that Bob would also like *D*.

Collaborative filtering can be implemented in a number of ways, and the Bayesian inference has proved to be a successful method. This involves working with posterior probabilities such as the following:

P(Bob Likes Z | Bob likes A, Bob likes B, ... , Bob Likes Y)

Clearly, for this mechanism to work accurately, large amounts of data must be collected. Information about thousands of individuals is needed, and information is required about dozens or hundreds of items for each individual. In the case of commerce sites, this information can be collected on

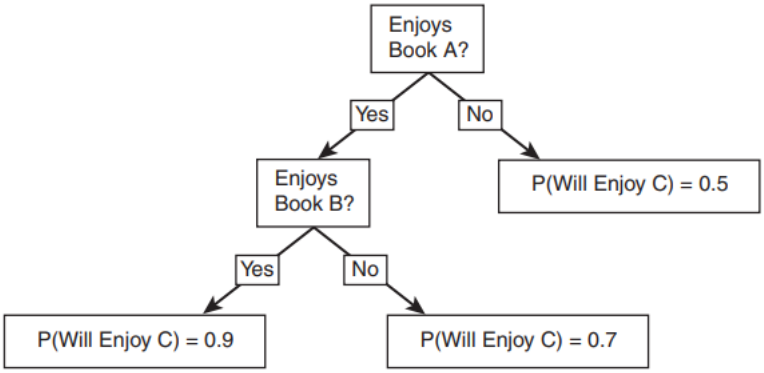

Figure 12.5
A decision tree for collaborative filtering

the basis of assuming that if a user buys a book or a CD, then he probably likes it. More accurate data can be collected by asking users to rate products.

The decision tree in Figure 12.5 relates enjoyment of book *C* to information about enjoyment of books *A* and *B*. It states that if you did not enjoy book A, then you will only have a 0.5 probability of enjoying book *C*. On the other hand, if you did enjoy book *A* and also enjoyed book *B*, then you will have a 0.9 chance of enjoying book *C*.

A full collaborative filtering system would have one decision tree for each item. A full Bayesian belief network would then be built from these decision trees, which can be used to make inferences about a new person on the basis of their likes or dislikes.