1

Prerequisites on Probability Theory

In this chapter we review some standard results and definitions from probability theory. The reader is assumed to have had some contact with probability theory before, and the purpose of this section is simply to brush up on some of the basic concepts and to introduce some of the notation used in the later chapters. Sections 1.1–1.3 are prerequisites for Section 2.3 and thereafter, Section 1.4 is a prerequisite for Chapter 4, and Section 1.5 is a prerequisite for Chapter 6 and Chapter 7.

1.1 Two Perspectives on Probability Theory

In many domains, the probability of seeing a certain outcome of an experiment can be interpreted as the *relative frequency* of seeing this particular outcome in all of the experiments performed. For instance, if you throw a six-sided die, then you would say that the probability of obtaining a three is 1/6, because if we throw this die a large number of times we would expect to see a three in approximately 1/6 of the throws. Along the same line of reasoning, we would also say that if we randomly draw a card from a deck consisting of 52 cards, then the probability that it will be a spade is 13/52. This interpretation of probability rests on the assumption that there is some stochastic process that can be repeated several times and from which the relative frequencies can be counted. On the other hand, we often talk about the probability of seeing a certain event although we cannot specify a frequency for it. For example, I may estimate that the probability that the Danish soccer team will win the World Cup in 2010 is p. This probability is my own personal judgment of how likely it is that the Danish team will actually win, and it is based on my belief, experience, and current state of information. However, another person may specify another probability for the same event, and it has no meaning to look for ways of determining which of us is right, if either. These probabilities are referred to as *subjective probabilities*. One way to interpret my subjective probability of Denmark winning the world cup in 2010 is to imagine the following two wagers:

1. If the Danish soccer team wins the world cup in 2010, I will receive $100.
2. I will draw a ball from an urn containing 100 balls out of which n are white and $100 - n$ are black. If the ball drawn is white then I will receive $100 in 2010.

If all the balls are white then I will prefer the second wager, and if all the balls are black then I will prefer the first. However, for a certain n between 0 and 100 I will be indifferent about the two wagers, and for this n, $n/100$ will be my subjective probability that the Danish soccer team will win the World Cup.

1.2 Fundamentals of Probability Theory

For both views on probability described above, we will refer to the set of possible outcomes of an experiment as the *sample space* of the experiment. Here we use the somewhat abstract term "experiment" to refer to any type of process for which the outcome is uncertain, e.g., the throw of a die and the winner of the World Cup. We shall also assume that the sample space of an experiment contains all possible outcomes of the experiment, and that each pair of outcomes are mutually exclusive. These assumptions ensure that the experiment is guaranteed to end up in exactly one of the specified outcomes in the sample space. For instance, for the die example above, the sample space would be $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, and for the soccer example the sample space would be $\mathcal{S} = \{\text{yes}, \text{no}\}$, assuming that I am interested only in whether the Danish team will win; both of the sample spaces satisfy the assumptions above. A subset of a sample space is called an *event*. For example, the event that we will get a value of three or higher with a six-sided die corresponds to the subset $\{3, 4, 5, 6\} \subseteq \{1, 2, 3, 4, 5, 6\}$, and the event will occur if the outcome of the throw is an element in the set. In general, we say that an event $\mathcal{A}$ is *true* for an experiment if the outcome of the experiment is an element of $\mathcal{A}$. When an event contains only one element, we will also refer to the event as an outcome.

To measure our degree of uncertainty about an experiment we assign a probability $P(\mathcal{A})$ to each event $\mathcal{A} \subseteq \mathcal{S}$. These probabilities must obey the following three axioms:

The event $\mathcal{S}$ that we will get an outcome in the sample space is certain to occur and is therefore assigned the probability 1.

Axiom 1 $P(\mathcal{S}) = 1$.

Any event $\mathcal{A}$ must have a nonnegative probability.

Axiom 2 *For all $\mathcal{A} \subseteq \mathcal{S}$ it holds that $P(\mathcal{A}) \geq 0$.*

If two events $\mathcal{A}$ and $\mathcal{B}$ are disjoint (see Figure 1.1(a)), then the probability of the combined event is the sum of the probabilities for the two individual events:

Axiom 3 *If $\mathcal{A} \subseteq \mathcal{S}$, $\mathcal{B} \subseteq \mathcal{S}$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$, then $P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B})$.*

For example, the event that a die will turn up 3, $\mathcal{B} = \{3\}$, and the event that the die will have an even number, $\mathcal{A} = \{2, 4, 6\}$, are two disjoint events, and the probability that one of these two events will occur is therefore

$$P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) = \frac{1}{6} + \frac{3}{6} = \frac{4}{6}.$$

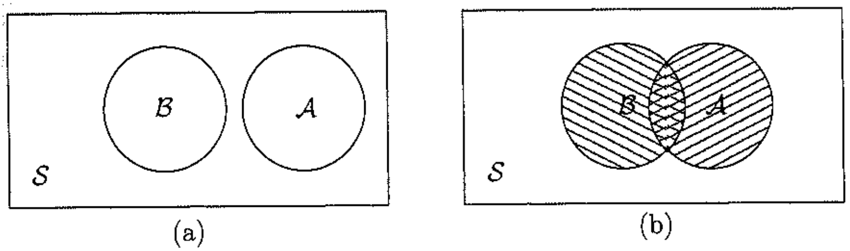


Fig. 1.1. In figure (a) the two events $\mathcal{A}$ and $\mathcal{B}$ are disjoint, whereas in figure (b), $\mathcal{A} \cap \mathcal{B} \neq \emptyset$.

On the other hand, if $\mathcal{A}$ and $\mathcal{B}$ are not disjoint (see Figure 1.1(b)), then it can easily be shown that

$$P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \cap \mathcal{B}),$$

where $\mathcal{A} \cap \mathcal{B}$ is the intersection between $\mathcal{A}$ and $\mathcal{B}$ and it represents the event that *both* $\mathcal{A}$ and $\mathcal{B}$ will occur. Consider again a deck with 52 cards. The event $\mathcal{A}$ that I will draw a spade and the event $\mathcal{B}$ that I will draw a king are clearly not disjoint events; their intersection specifies the event that I will draw the king of spades, $\mathcal{A} \cap \mathcal{B} = \{\text{king of spades}\}$. Thus, the probability that I will draw either a king or a spade is

$$P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \cap \mathcal{B}) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52}.$$

Notation: Sometimes we will emphasize that a probability is based on a frequency (rather than being a subjective probability), in which case we will use the notation $P^{\#}$. If the event $\mathcal{A}$ contains only one outcome a, we write $P(a)$ rather than $P(\{a\})$.

1.2.1 Conditional Probabilities

Whenever a statement about the probability $P(\mathcal{A})$ of an event $\mathcal{A}$ is given, then it is implicitly given conditioned on other known factors. For example, a statement such as "the probability of the die turning up 6 is $\frac{1}{6}$" usually has the unsaid prerequisite that it is a fair die, or rather, as long as I know nothing further, I assume it to be a fair die. This means that the statement should be "given that it is a fair die, the probability ...." In this way, any statement on probabilities is a statement conditioned on what else is known. These types of probabilities are called *conditional probabilities* and are generally statements of the following kind:

"Given the event $\mathcal{B}$, the probability of the event $\mathcal{A}$ is p."

The notation for the preceding statement is $P(\mathcal{A}|\mathcal{B}) = p$. It should be stressed that $P(\mathcal{A}|\mathcal{B}) = p$ does not mean that whenever $\mathcal{B}$ is true, then the probability of $\mathcal{A}$ is p. It means that if $\mathcal{B}$ is true, and *everything else is irrelevant for $\mathcal{A}$*, then the probability of $\mathcal{A}$ is p.

Assume that we have assigned probabilities to all subsets of the sample space $\mathcal{S}$, and let $\mathcal{A}$ and $\mathcal{B}$ be subsets of $\mathcal{S}$ (Figure 1.1(b)). The question is whether the probability assignment for $\mathcal{S}$ can be used to calculate $P(\mathcal{A}|\mathcal{B})$. If we know the event $\mathcal{B}$, then all possible outcomes are elements of $\mathcal{B}$, and the outcomes for which $\mathcal{A}$ can be true are $\mathcal{A} \cap \mathcal{B}$. So, we look for the probability assignment for $\mathcal{A} \cap \mathcal{B}$ given that we know $\mathcal{B}$. Knowing $\mathcal{B}$ does not change the proportion between the probabilities of $\mathcal{A} \cap \mathcal{B}$ and another set $\mathcal{C} \cap \mathcal{B}$ (if, for example, I will bet twice as much on $\mathcal{A} \cap \mathcal{B}$ as on $\mathcal{C} \cap \mathcal{B}$, then after knowing $\mathcal{B}$, I will still bet twice as much on $\mathcal{A} \cap \mathcal{B}$ as on $\mathcal{C} \cap \mathcal{B}$). We can conclude that the proportions $P(\mathcal{A} \cap \mathcal{B})/P(\mathcal{C} \cap \mathcal{B})$ and $P(\mathcal{A}|\mathcal{B})/P(\mathcal{C}|\mathcal{B})$ must be the same. Setting $\mathcal{C} = \mathcal{B}$, and since we know from Axiom 1 that $P(\mathcal{B}|\mathcal{B}) = 1$, we have justified the following property, which should be considered an axiom.

Property 1.1 (Conditional probability). For two events $\mathcal{A}$ and $\mathcal{B}$, with $P(\mathcal{B}) > 0$, the conditional probability for $\mathcal{A}$ given $\mathcal{B}$ is

$$P(\mathcal{A} \,|\, \mathcal{B}) = \frac{P(\mathcal{A} \cap \mathcal{B})}{P(\mathcal{B})}.$$

For example, the conditional probability that a die will come up 4 given that we get an even number is $P(\mathcal{A} = \{4\} \,|\, \mathcal{B} = \{2,4,6\}) = P(\{4\})/P(\{2,4,6\})$, and by assuming that the die is fair we get $\frac{1/6}{3/6} = \frac{1}{3}$.

Obviously, when working with conditional probabilities we can also condition on more than one event, in which case the definition of a conditional probability generalizes as

$$P(\mathcal{A} \,|\, \mathcal{B} \cap \mathcal{C}) = \frac{P(\mathcal{A} \cap \mathcal{B} \cap \mathcal{C})}{P(\mathcal{B} \cap \mathcal{C})}.$$

1.2.2 Probability Calculus

The expression in Property 1.1 can be rewritten so that we obtain the so-called *fundamental rule* for probability calculus:

Theorem 1.1 (The fundamental rule).

$$P(\mathcal{A} \,|\, \mathcal{B})P(\mathcal{B}) = P(\mathcal{A} \cap \mathcal{B}). \tag{1.1}$$

That is, the fundamental rule tells us how to calculate the probability of seeing both $\mathcal{A}$ and $\mathcal{B}$ when we know the probability of $\mathcal{A}$ given $\mathcal{B}$ and the probability of $\mathcal{B}$.

By conditioning on another event $\mathcal{C}$, the fundamental rule can also be written as

$$P(\mathcal{A} \,|\, \mathcal{B} \cap \mathcal{C})P(\mathcal{B} \,|\, \mathcal{C}) = P(\mathcal{A} \cap \mathcal{B} \,|\, \mathcal{C}).$$

Since $P(\mathcal{A} \cap \mathcal{B}) = P(\mathcal{B} \cap \mathcal{A})$ (and also $P(\mathcal{A} \cap \mathcal{B} \,|\, \mathcal{C}) = P(\mathcal{B} \cap \mathcal{A} \,|\, \mathcal{C})$), we get that $P(\mathcal{A} \,|\, \mathcal{B})P(\mathcal{B}) = P(\mathcal{A} \cap \mathcal{B}) = P(\mathcal{B} \,|\, \mathcal{A})P(\mathcal{A})$ from the fundamental rule. This yields the well-known *Bayes' rule*:

Theorem 1.2 (Bayes' rule).

$$P(\mathcal{A} \,|\, \mathcal{B}) = \frac{P(\mathcal{B} \,|\, \mathcal{A})P(\mathcal{A})}{P(\mathcal{B})}.$$

Bayes' rule provides us with a method for updating our beliefs about an event $\mathcal{A}$ given that we get information about another event $\mathcal{B}$. For this reason $P(\mathcal{A})$ is usually called the *prior* probability of $\mathcal{A}$, whereas $P(\mathcal{A} \,|\, \mathcal{B})$ is called the *posterior* probability of $\mathcal{A}$ given $\mathcal{B}$; the probability $P(\mathcal{B} \,|\, \mathcal{A})$ is called the *likelihood* of $\mathcal{A}$ given $\mathcal{B}$. For an explanation of this strange use of the term, see Example 1.1.

Finally, as for the fundamental rule, we can also state Bayes' rule in a context $\mathcal{C}$:

$$P(\mathcal{A} \,|\, \mathcal{B}, \mathcal{C}) = \frac{P(\mathcal{B} \,|\, \mathcal{A}, \mathcal{C})P(\mathcal{A} \,|\, \mathcal{C})}{P(\mathcal{B} \,|\, \mathcal{C})}.$$

Example 1.1. We have two diseases a_1 and a_2, both of which can cause the symptom b. Let $P(b \,|\, a_1) = 0.9$ and $P(b \,|\, a_2) = 0.3$. Assume that the prior probabilities for a_1 and a_2 are the same ($P(a_1) = P(a_2)$). Now, if b occurs, Bayes' rule gives

$$P(a_1 \,|\, b) = \frac{P(b \,|\, a_1)P(a_1)}{P(b)} = 0.9 \cdot \frac{P(a_1)}{P(b)};$$

$$P(a_2 \,|\, b) = \frac{P(b \,|\, a_2)P(a_2)}{P(b)} = 0.3 \cdot \frac{P(a_2)}{P(b)}.$$

Even though we cannot calculate the posterior probabilities, we can conclude that a_1 is three times as likely as a_2 given the symptom b.

If we furthermore know that a_1 and a_2 are the only possible causes of b, we can go even further (assuming that the probability of having both diseases is 0). Then $P(a_1 \mid b) + P(a_2 \mid b) = 1$, and we get

$$\frac{P(a_1)}{P(b)} = \frac{P(a_2)}{P(b)} = \frac{1}{0.9+0.3} = \frac{1}{1.2},$$

$P(a_1 \mid b) = 0.9/1.2 = 0.75$, and $P(a_2 \mid b) = 0.3/1.2 = 0.25$.

1.2.3 Conditional Independence

Sometimes information on one event $\mathcal{B}$ does not change our belief about the occurrence of another event $\mathcal{A}$, and in this case we say that $\mathcal{A}$ and $\mathcal{B}$ are *independent*.

Definition 1.1 (Independence). *The events $\mathcal{A}$ and $\mathcal{B}$ are* independent *if*

$$P(\mathcal{A} \mid \mathcal{B}) = P(\mathcal{A}).$$

For example, if we throw two fair dice, then seeing that the first die turns up 2 will not change our beliefs about the outcome of the second die.

This notion of independence is symmetric, so that if $\mathcal{A}$ is independent of $\mathcal{B}$, then $\mathcal{B}$ is independent of $\mathcal{A}$:

$$P(\mathcal{B} \mid \mathcal{A}) = \frac{P(\mathcal{A} \cap \mathcal{B})}{P(\mathcal{A})} = \frac{P(\mathcal{A} \mid \mathcal{B})P(\mathcal{B})}{P(\mathcal{A})} = \frac{P(\mathcal{A})P(\mathcal{B})}{P(\mathcal{A})} = P(\mathcal{B}).$$

The proof requires that $P(\mathcal{A}) > 0$, so if $P(\mathcal{A}) = 0$, the calculations are not valid. However, for our considerations it does not matter; if $\mathcal{A}$ is impossible why bother considering it?

When two events are independent, then the fundamental rule can be rewritten as

$$P(\mathcal{A} \cap \mathcal{B}) = P(\mathcal{A} \mid \mathcal{B})P(\mathcal{B}) = P(\mathcal{A}) \cdot P(\mathcal{B}).$$

That is, we can calculate the probability that both events will occur by multiplying the probabilities for the individual events.

The concept of independence also appears when we are conditioning on several events. Specifically, if information about the event $\mathcal{B}$ does not change our belief about the event $\mathcal{A}$ when we already know the event $\mathcal{C}$, then we say that $\mathcal{A}$ and $\mathcal{B}$ are *conditionally independent* given $\mathcal{C}$.

Definition 1.2 (Conditional independence). *The events $\mathcal{A}$ and $\mathcal{B}$ are* conditionally independent *given the event $\mathcal{C}$ if*

$$P(\mathcal{A} \mid \mathcal{B} \cap \mathcal{C}) = P(\mathcal{A} \mid \mathcal{C}).$$

Similar to the situation above, the conditional independence statement is symmetric. If $\mathcal{A}$ is conditionally independent of $\mathcal{B}$ given $\mathcal{C}$, then $\mathcal{B}$ is conditionally independent of $\mathcal{A}$ given $\mathcal{C}$:

$$\begin{aligned} P(\mathcal{B} \mid \mathcal{A} \cap \mathcal{C}) &= \frac{P(\mathcal{A} \cap \mathcal{B} \mid \mathcal{C})P(\mathcal{C})}{P(\mathcal{A} \mid \mathcal{C})P(\mathcal{C})} = \frac{P(\mathcal{A} \mid \mathcal{B} \cap \mathcal{C})P(\mathcal{B} \mid \mathcal{C})}{P(\mathcal{A} \mid \mathcal{C})} = \frac{P(\mathcal{A} \mid \mathcal{C})P(\mathcal{B} \mid \mathcal{C})}{P(\mathcal{A} \mid \mathcal{C})} \\ &= P(\mathcal{B} \mid \mathcal{C}). \end{aligned}$$

Furthermore, when two events are conditionally independent, then we can use a multiplication rule similar to the one above when calculating the probability that both of the events will occur:

$$P(\mathcal{A} \cap \mathcal{B} \mid \mathcal{C}) = P(\mathcal{A} \mid \mathcal{C}) \cdot P(\mathcal{B} \mid \mathcal{C}).$$

Note that when two events are independent it is actually a special case of conditional independence but with $\mathcal{C} = \emptyset$.

1.3 Probability Calculus for Variables

So far we have talked about probabilities of simple events and outcomes with respect to a certain sample space. In this book, however, we will be working with a collection of sample spaces, also called *variables*, and we will now extend the concepts above to probabilities over variables. A variable can be considered an experiment, and for each outcome of the experiment the variable has a corresponding *state*. The set of states associated with a variable A is denoted by $\mathrm{sp}(A) = (a_1, a_2, \ldots, a_n)$, and similar to the sample space these states should be *mutually exclusive* and *exhaustive*. The last assumption ensures that the variable is in one of its states (although we may not know which one), and the first assumption ensures that the variable is in only one state. For example, if we let D be a variable representing the outcome of rolling a die, then its state space would be $\mathrm{sp}(D) = (1, 2, 3, 4, 5, 6)$. We will use uppercase letters for variables and lowercase letters for states, and unless otherwise stated, a variable has a finite number of states.

For a variable A with states $a_1, \ldots, a_n$, we express our uncertainty about its state through a probability distribution $P(A)$ over these states:

$$P(A) = (x_1, \ldots, x_n); \qquad x_i \geq 0; \qquad \sum_{i=1}^{n} x_i = x_1 + \cdots + x_n = 1,$$

where x_i is the probability of A being in state a_i. A distribution is called *uniform* (or *even*) if all probabilities are equal.

Notation: In general, the probability of A being in state a_i is denoted by $P(A = a_i)$, and denoted by $P(a_i)$ if the variable is obvious from the context.

As we talked about conditional probabilities for events, we can also talk about *conditional probabilities* for variables: If the variable B has states $b_1, \ldots, b_m$, then $P(A \mid B)$ contains $n \cdot m$ conditional probabilities $P(a_i \mid b_j)$ that specify the probability of seeing a_i given b_j. That is, the conditional probability for a variable given another variable is a set of probabilities (usually organized in an $n \times m$ table) with one probability for each configuration of the states of the variables involved (see Table 1.1 for an example). Moreover, since $P(A \mid B)$ specifies a probability distribution for each event $B = b_j$, we know from Axiom 1 that the probabilities over A should sum to 1 for each state of B:

$$\sum_{i=1}^{n} P(A = a_i \mid B = b_j) = 1 \text{ for each } b_j.$$

	b_1	b_2	b_3
a_1	0.4	0.3	0.6
a_2	0.6	0.7	0.4

Table 1.1. An example of a conditional probability table $P(A \mid B)$ for the binary variable A given the ternary variable B. Note that for each state of B the probabilities of A sum up to 1.

The probability of seeing joint outcomes for different experiments can be expressed by the *joint probability* for two or more variables: For each configuration (a_i, b_j) of the variables A and B, $P(A, B)$ specifies the probability of seeing both $A = a_i$ *and* $B = b_j$. Hence, $P(A, B)$ consists of $n \cdot m$ numbers, and, similar to $P(A \mid B)$, $P(A, B)$ is usually represented in an $n \times m$ table (see Table 1.2 for an example). Note that since the state spaces of both A and B are mutually exclusive and exhaustive, it follows that all combinations of their states (the Cartesian product) are also mutually exclusive and exhaustive, and they can therefore be considered a sample space. Hence, by Axiom 1,

$$P(A, B) = \sum_{i=1}^{n} \sum_{j=1}^{m} P(A = a_i, B = b_j) = 1.$$

	b_1	b_2	b_3
a_1	0.16	0.12	0.12
a_2	0.24	0.28	0.08

Table 1.2. An example of a joint probability table $P(A, B)$ for the binary variable A and the ternary variable B. Note that the sum of all entries is 1.

When the fundamental rule (equation (1.1)) is used on variables A and B, the procedure is to apply the rule to each of the $n \cdot m$ configurations (a_i, b_j) of the two variables:

$$P(a_i \mid b_j) P(b_j) = P(a_i, b_j).$$

This means that in the table $P(A \mid B)$, each probability in $P(A \mid b_j)$ is multiplied by $P(b_j)$ to obtain the table $P(A, b_j)$, and by doing this for each b_j we get $P(A, B)$. If $P(B) = (0.4, 0.4, 0.2)$, then Table 1.2 is the result of using the fundamental rule on Table 1.1 (see also Table 1.3).

$P(A, B) =$

	b_1	b_2	b_3
a_1	$0.4 \cdot 0.4$	$0.3 \cdot 0.4$	$0.6 \cdot 0.2$
a_2	$0.6 \cdot 0.4$	$0.7 \cdot 0.4$	$0.4 \cdot 0.2$

$=$

	b_1	b_2	b_3
a_1	0.16	0.12	0.12
a_2	0.24	0.28	0.08

Table 1.3. The joint probability table $P(A, B)$ in Table 1.2 can be found by multiplying $P(B) = (0.4, 0.4, 0.2)$ by $P(A \mid B)$ in Table 1.1.

When applied to variables, the fundamental rule is expressed as follows:

Theorem 1.3 (The fundamental rule for variables).

$$P(A, B) = P(A \mid B) P(B),$$

and conditioned on another variable C we have

$$P(A, B \mid C) = P(A \mid B, C) P(B \mid C).$$

From a joint probability table $P(A, B)$, the probability distribution $P(A)$ can be calculated by considering the outcomes of B that can occur together with each state a_i of A. There are exactly m different outcomes for which A is in state a_i, namely the mutually exclusive outcomes $(a_i, b_1), \ldots, (a_i, b_m)$. Therefore, by Axiom 3,

$$P(a_i) = \sum_{j=1}^{m} P(a_i, b_j).$$

This calculation is called *marginalization*, and we say that the variable B is marginalized out of $P(A, B)$ (resulting in $P(A)$). The notation is

$$P(A) = \sum_{B} P(A, B).$$

By marginalizing B out of Table 1.2, we get

$$P(A) = (0.16 + 0.12 + 0.12, 0.24 + 0.28 + 0.08) = (0.4, 0.6),$$

and by marginalizing out A we get

$$P(B) = (0.16+0.24, 0.12+0.28, 0.12+0.08) = (0.4, 0.4, 0.2).$$

That is, the marginalization operation allows us to remove variables from a joint probability distribution.

Bayes' rule for events (Theorem 1.2) can also be extended to variables, by treating the division in the same way as we treated multiplication above.

Theorem 1.4 (Bayes' rule for variables).

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)} = \frac{P(A,B)}{\sum_B P(A,B)},$$

and conditioned on another variable C we have

$$P(B \mid A,C) = \frac{P(A \mid B,C)P(B \mid C)}{P(A \mid C)} = \frac{P(A,B \mid C)}{\sum_B P(A,B \mid C)}.$$

Note that the two equalities in the equations follow from (1) the fundamental rule and (2) the marginalization operator described above.

By applying Bayes' rule using $P(A)$, $P(B)$, and $P(A \mid B)$ as specified above, we get $P(B \mid A)$ shown in Table 1.4.

$$P(B \mid A) = \tfrac{P(A \mid B)P(B)}{P(A)} = \begin{array}{|c|cc|}\hline & a_1 & a_2 \\ \hline b_1 & \frac{0.4 \cdot 0.4}{0.4} & \frac{0.6 \cdot 0.4}{0.6} \\ b_2 & \frac{0.3 \cdot 0.4}{0.4} & \frac{0.7 \cdot 0.4}{0.6} \\ b_3 & \frac{0.6 \cdot 0.2}{0.4} & \frac{0.4 \cdot 0.2}{0.6} \\ \hline \end{array} = \begin{array}{|c|cc|}\hline & a_1 & a_2 \\ \hline b_1 & 0.4 & 0.4 \\ b_2 & 0.3 & 0.47 \\ b_3 & 0.3 & 0.13 \\ \hline \end{array}$$

Table 1.4. The conditional probability $P(B \mid A)$ obtained by applying Bayes' rule to $P(A \mid B)$ in Table 1.1, $P(A) = (0.4, 0.6)$, and $P(B) = (0.4, 0.4, 0.2)$. Note that the probabilities over B sum to 1 for each state of A.

The concept of (conditional) independence is also defined for variables.

Definition 1.3 (Conditional independence for variables). *Two variables A and C are said to be* conditionally independent *given the variable B if*

$$P(a_i \mid c_k, b_j) = P(a_i \mid b_j)$$

for each $a_i \in \mathrm{sp}(A)$, $b_j \in \mathrm{sp}(B)$, and $c_k \in \mathrm{sp}(C)$.

As a shorthand notation we will sometimes write $P(A \mid C, B) = P(A \mid B)$.

This means that when the state of B is known, then no knowledge of C will alter the probability of A. Observe that we require the independence statement to hold for each state of B; if the conditioning set is empty then we say that A and C are *marginally independent* or just independent (written as $P(A \mid C) = P(A)$).

When two variables A and C are conditionally independent given B, then the fundamental rule (Theorem 1.3) can be simplified:

$$P(A, C \mid B) = P(A \mid B, C)P(C \mid B) = P(A \mid B)P(C \mid B).$$

In the expression above, we multiply two conditional probability tables over different domains. Fortunately, the method for doing this multiplication is a straightforward extension of what we have done so far:

$$P(a_i, c_k \mid b_j) = P(a_i \mid b_j)P(c_k \mid b_j).$$

For example, by multiplying $P(A \mid B)$ and $P(C \mid B)$ (specified in Table 1.1 and Table 1.5, respectively) we get the joint probability $P(A, C \mid B)$ in Table 1.6.

	b_1	b_2	b_3
c_1	0.2	0.9	0.3
c_2	0.05	0.05	0.2
c_3	0.75	0.05	0.5

Table 1.5. The conditional probability table $P(C \mid B)$ for the ternary variable C given the ternary variable B.

$P(A, C \mid B) = P(A \mid B)P(C \mid B)$

=

	b_1	b_2	b_3
c_1	$(0.2 \cdot 0.4, 0.2 \cdot 0.6)$	$(0.9 \cdot 0.3, 0.9 \cdot 0.7)$	$(0.3 \cdot 0.6, 0.3 \cdot 0.4)$
c_2	$(0.05 \cdot 0.4, 0.05 \cdot 0.6)$	$(0.05 \cdot 0.3, 0.05 \cdot 0.7)$	$(0.2 \cdot 0.6, 0.2 \cdot 0.4)$
c_3	$(0.75 \cdot 0.4, 0.75 \cdot 0.6)$	$(0.05 \cdot 0.3, 0.05 \cdot 0.7)$	$(0.5 \cdot 0.6, 0.5 \cdot 0.4)$

=

	b_1	b_2	b_3
c_1	$(0.08, 0.12)$	$(0.27, 0.63)$	$(0.18, 0.12)$
c_2	$(0.02, 0.03)$	$(0.015, 0.035)$	$(0.12, 0.08)$
c_3	$(0.3, 0.45)$	$(0.015, 0.035)$	$(0.3, 0.2)$

Table 1.6. If A and C are conditionally independent given B, then $P(A, C \mid B)$ can be found by multiplying $P(A \mid B)$ and $P(C \mid B)$ as specified in Table 1.1 and Table 1.5, respectively.

1.3.1 Calculations with Probability Tables: An Example

To illustrate the theorems above, assume that we have three variables, A, B, and C, with the probabilities as in Table 1.7. We receive evidence $A = a_2$ and

$C = c_1$ and we would now like to calculate the conditional probability table $P(B \,|\, a_2, c_1)$.

	b_1	b_2	b_3
a_1	(0, 0.05, 0.05)	(0.05, 0.05, 0)	(0.05, 0.05, 0.05)
a_2	(0.1, 0.1, 0)	(0.1, 0, 0.1)	(0.2, 0, 0.05)

Table 1.7. A joint probability table for the variables A, B, and C. The three numbers in each entry correspond to the states c_1, c_2, and c_3.

First, we focus on the part of the table corresponding to $A = a_2$ and $C = c_1$, and we get

$$P(a_2, B, c_1) = (0.1, 0.1, 0.2). \tag{1.2}$$

To calculate $P(B \,|\, a_2, c_1)$, we can use Theorem 1.4:

$$P(B \,|\, a_2, c_1) = \frac{P(a_2, B, c_1)}{P(a_2, c_1)} = \frac{P(a_2, B, c_1)}{\sum_B P(a_2, B, c_1)}. \tag{1.3}$$

By marginalizing B out of equation (1.2) we get

$$P(a_2, c_1) = 0.1 + 0.1 + 0.2 = 0.4.$$

Finally, by performing the division in equation (1.3) we get

$$P(B \,|\, a_2, c_1) = \left(\frac{0.1}{0.4}, \frac{0.1}{0.4}, \frac{0.2}{0.4}\right) = (0.25, 0.25, 0.5).$$

Another way of doing the same is to say that we wish to transform $P(a_2, B, c_1)$ into a probability distribution. Because the numbers do not add up to one, we *normalize* the distribution by dividing each number by the sum of all the numbers.

Suppose now that we were given only the evidence $A = a_2$, and we want to calculate $P(B \,|\, a_2, C)$. The calculation of this probability table follows the same steps as above, except that we now work with tables during the calculations. As before, we start by focusing on the part of $P(A, B, C)$ corresponding to $A = a_2$ and we get the result in Table 1.8.

To calculate $P(B \,|\, a_2, C)$ we use

$$P(B \,|\, a_2, C) = \frac{P(a_2, B, C)}{P(a_2, C)} = \frac{P(a_2, B, C)}{\sum_B P(a_2, B, C)}. \tag{1.4}$$

The probability $P(a_2, C)$ is found by marginalizing B out of Table 1.8:

$$P(a_2, C) = (0.1 + 0.1 + 0.2, 0.1 + 0 + 0, 0 + 0.1 + 0.05) = (0.4, 0.1, 0.15), \tag{1.5}$$

and by inserting this in equation (1.4) we get the result shown in Table 1.2.

	b_1	b_2	b_3
c_1	0.1	0.1	0.2
c_2	0.1	0	0
c_2	0	0.1	0.05

Table 1.8. The probability table $P(a_2, B, C)$ that corresponds to the part of the probability table in Table 1.8 restricted to $A = a_2$.

$$P(B \,|\, a_2, C) = \begin{array}{c|ccc} & b_1 & b_2 & b_3 \\ \hline c_1 & \frac{0.1}{0.4} & \frac{0.1}{0.4} & \frac{0.2}{0.4} \\ c_2 & \frac{0.1}{0.1} & \frac{0}{0.1} & \frac{0}{0.1} \\ c_2 & \frac{0}{0.15} & \frac{0.1}{0.15} & \frac{0.05}{0.15} \end{array} = \begin{array}{c|ccc} & b_1 & b_2 & b_3 \\ \hline c_1 & 0.25 & 0.25 & 0.5 \\ c_2 & 1 & 0 & 0 \\ c_2 & 0 & 2/3 & 1/3 \end{array}$$

Table 1.9. The calculation of $P(B \,|\, a_2, C)$ using $P(a_2, B, C)$ (Table 1.1) and $P(a_2, C)$ (equation (1.5)).

1.4 An Algebra of Potentials

Below we list some properties of the algebra of multiplication and marginalization of tables. The tables need not be (conditional) probabilities, and they are generally called *potentials*.

A potential ϕ is a real-valued function over a *domain* of finite variables $\mathcal{X}$:

$$\phi : \mathrm{sp}(\mathcal{X}) \to \mathbb{R}$$

The domain of a potential is denoted by $\mathrm{dom}\,(\phi)$. For example, the domain of the potential $P(A, B \,|\, C)$ is $\mathrm{dom}\,(P(A, B \,|\, C)) = \{A, B, C\}$.

Two potentials can be *multiplied*, denoted by an (often suppressed) dot. Multiplication has the following properties:

1. $\mathrm{dom}\,(\phi_1\phi_2) = \mathrm{dom}\,(\phi_1) \cup \mathrm{dom}\,(\phi_2)$.
2. **The commutative law**: $\phi_1\phi_2 = \phi_2\phi_1$.
3. **The associative law**: $(\phi_1\phi_2)\phi_3 = \phi_1(\phi_2\phi_3)$.
4. **Existence of unit**: The unit potential $\mathbf{1}$ is a potential that contains only 1's and is defined over any domain such that $\mathbf{1} \cdot \phi = \phi$, for all potentials ϕ.

The marginalization operator defined in Section 1.3 can be generalized to potentials so that $\sum_A \phi$ is a potential over $\mathrm{dom}(\phi) \setminus \{A\}$. Furthermore, marginalization is *commutative*:

$$\sum_A \sum_B \phi = \sum_B \sum_A \phi.$$

For potentials of the form $P(A \,|\, \mathcal{V})$, where $\mathcal{V}$ is a set of variables, we have

5. **The unit potential property**: $\sum_A P(A \,|\, \mathcal{V}) = 1$.

For marginalization of a product, the following holds

6. **The distributive law**: If $A \notin \mathrm{dom}(\phi_1)$, then $\sum_A \phi_1\phi_2 = \phi_1 \sum_A \phi_2$.

The distributive law is usually known as $ab + ac = a(b + c)$, and the preceding formula is actually the same law applied to tables. To verify it, consider the calculations in Tables 1.10–1.14. Here we see that Table 1.12 and Table 1.14 are equal and correspond to the left-hand and right-hand sides of the distributive law.

$B \setminus A$	a_1	a_2
b_1	x_1	x_2
b_2	x_3	x_4

$B \setminus C$	c_1	c_2
b_1	y_1	y_2
b_2	y_3	y_4

Table 1.10. $\phi_1(A,B)$ and $\phi_2(C,B)$.

$B \setminus A$	a_1	a_2
b_1	(x_1y_1, x_1y_2)	(x_2y_1, x_2y_2)
b_2	(x_3y_3, x_3y_4)	(x_4y_3, x_4y_4)

Table 1.11. $\phi_1(A,B) \cdot \phi_2(C,B)$. The two numbers in each entry correspond to the states c_1 and c_2.

$B \setminus A$	a_1	a_2
b_1	$x_1y_1 + x_1y_2$	$x_2y_1 + x_2y_2$
b_2	$x_3y_3 + x_3y_4$	$x_4y_3 + x_4y_4$

Table 1.12. $\sum_C \phi_1(A,B) \cdot \phi_2(C,B)$.

B	
b_1	$y_1 + y_2$
b_2	$y_3 + y_4$

Table 1.13. $\sum_C \phi_2(C,B)$.

We also use the term *projection* for marginalization. For example, if A and B are marginalized out of $\phi(A,B,C)$, we may say that ϕ is *projected* down to C, and we use the notation $\phi^{\downarrow C}$. With this notation, the properties of marginalization look as follows ($\mathcal{V}$ and $\mathcal{W}$ denote sets of variables):

$B \setminus A$	a_1	a_2
b_1	$x_1(y_1 + y_2)$	$x_2(y_1 + y_2)$
b_2	$x_3(y_3 + y_4)$	$x_4(y_3 + y_4)$

Table 1.14. $\phi_1(A,B) \sum_C \phi_2(C,B)$.

7. **The commutative law:** $(\phi^{\downarrow\mathcal{V}})^{\downarrow\mathcal{W}} = (\phi^{\downarrow\mathcal{W}})^{\downarrow\mathcal{V}}$.
8. **The distributive law:** If $\mathrm{dom}(\phi_1) \subseteq \mathcal{V}$, then $(\phi_1\phi_2)^{\downarrow\mathcal{V}} = \phi_1(\phi_2^{\downarrow\mathcal{V}})$.

1.5 Random Variables

Let $\mathcal{S}$ be a sample space. A *random variable* is a real-valued function on $\mathcal{S}$; $V : \mathcal{S} \to \mathbb{R}$. If, for example, you throw a die, and you win \$1 if you get 4 or above, and you lose \$1 if you get 3 or below, then the corresponding random variable is a function with value -1 on $\{1,2,3\}$ and 1 on $\{4,5,6\}$.

The *mean value* of a random variable V on $\mathcal{S}$ is defined as

$$\mu(V) = \sum_{s\in\mathcal{S}} V(s)P(s). \tag{1.6}$$

For the example above, the mean value is $-1\frac{1}{6}+-1\frac{1}{6}+-1\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=0$ (provided that the die is fair). The mean value is also called the *expected value*.

A measure of how much a random variable varies between its values is the *variance*, σ^2. It is defined as the mean of the square of the difference between value and mean:

$$\sigma^2(V) = \sum_{s\in\mathcal{S}} (V(s) - \mu(V))^2 P(s). \tag{1.7}$$

For the example above we have

$$\sigma^2 = 3(-1-0)^2\frac{1}{6} + 3(1-0)^2\frac{1}{6} = 1.$$

1.5.1 Continuous Distributions

Consider an experiment, where an arrow is thrown at the $[0,1] \times [0,1]$ square. The possible outcomes are the points (x,y) in the unit square. Since the probability is zero for any particular outcome, the probability distribution is assigned to subsets of the unit square. We may think of this assignment as a process of distributing a probability mass of 1 over the sample space. We may, for example, assign a probability for landing in the small square $[x, x+\epsilon] \times [y, y+\epsilon]$. To be more systematic, let n be a natural number, then the unit square can be partitioned into small squares of the type $[\frac{i}{n}, \frac{i+1}{n}] \times [\frac{j}{n}, \frac{j+1}{n}]$, and we can assign probabilities $P([\frac{i}{n}, \frac{i+1}{n}] \times [\frac{j}{n}, \frac{j+1}{n}])$ to these squares with area

$\frac{1}{n^2}$. Now, if $P([\frac{i}{n}, \frac{i+1}{n}] \times [\frac{j}{n}, \frac{j+1}{n}]) = x$, then you can say that the probability mass x is distributed over the small square with an average density of $n^2 x$, and we define the *density function* (also called the *frequency function*) $f(x,y)$ as

$$f(x,y) = \lim_{n\to\infty} n^2 P\left(\left[x, x+\frac{1}{n}\right] \times \left[y, y+\frac{1}{n}\right]\right).$$

In general, if $\mathcal{S}$ is a continuous sample space, the density function is a nonnegative real-valued function f on $\mathcal{S}$, for which it holds that for any subset $\mathcal{A}$ of $\mathcal{S}$,

$$\int_{\mathcal{A}} f(s)ds = P(\mathcal{A}).$$

In particular,

$$\int_{\mathcal{S}} f(s)ds = 1.$$

When $\mathcal{S}$ is an interval $[a,b]$ (possibly infinite), the outcomes are real numbers (such as height or weight), and you may be interested in the mean (height or weight). It is defined as

$$\mu = \int_a^b x f(x)dx,$$

and the variance is given by

$$\sigma^2 = \int_a^b (\mu - x)^2 f(x)dx.$$

Mathematically, the mean and variance are the mean and variance of the identity function $I(x) = x$, but we use the term "mean and variance of the *distribution*."

1.6 Exercises

Exercise 1.1. Given Axioms 1 to 3, prove that

$$P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \cap \mathcal{B}).$$

Exercise 1.2. Consider the experiment of rolling a red and a blue fair six-sided die. Give an example of a sample space for the experiment along with probabilities for each outcome. Suppose then that we are interested only in the sum of the dice (that is, the experiment consists in rolling the dice and adding up the numbers). Give another example of a sample space for this experiment and probabilities for the outcomes.

Exercise 1.3. Consider the experiment of flipping a fair coin, and if it lands heads, rolling a fair four-sided die, and if it lands tails, rolling a fair six-sided die. Suppose that we are interested only in the number rolled by the die, and a sample space $\mathcal{S}_A$ for the experiment could thus be the numbers from 1 to 6. Another sample space could be $\mathcal{S}_B = \{t1, \ldots, t6, h1, \ldots, h4\}$, with for example $t2$ meaning "tails and a roll of 2" and $h4$ meaning "heads and a roll of 4." Choose either $\mathcal{S}_A$ or $\mathcal{S}_B$ and associate probabilities with it. According to your sample space and probability distribution, what is the probability of rolling either 3 or 5.

Exercise 1.4. Draw a Venn diagram (like that in Figure 1.1) over $\mathcal{S}_B$ defined in Exercise 1.3. The diagram should show the events corresponding to "rolling a 3," "flipping tails," and "flipping tails and rolling a 3."

Exercise 1.5. Let $\mathcal{S}_B$ be defined as in Exercise 1.3, but with a loaded coin and loaded dice. A probability distribution is given in Table 1.15. What is the probability that the loaded coin lands "tails"? What is the conditional probability of rolling a 4, given that the coin lands tails? Which of the loaded dice has the highest chance of rolling 4 or more?

$t1$	$\frac{5}{18}$	$t6$	$\frac{1}{18}$
$t2$	$\frac{1}{9}$	$h1$	$\frac{1}{24}$
$t3$	$\frac{1}{9}$	$h2$	$\frac{1}{24}$
$t4$	$\frac{1}{18}$	$h3$	$\frac{1}{8}$
$t5$	$\frac{1}{18}$	$h4$	$\frac{1}{8}$

Table 1.15. Probabilities for $\mathcal{S}_B$ in Exercise 1.5.

Exercise 1.6. Prove that

$$P(\mathcal{A} \,|\, \mathcal{B} \cup \mathcal{C}) P(\mathcal{B} \,|\, \mathcal{C}) = P(\mathcal{A} \cap \mathcal{B} \,|\, \mathcal{C}).$$

Exercise 1.7. A farmer has a cow, which he suspects is pregnant. He administers a test to the urine of the cow to determine whether it is pregnant. There are four outcomes in this experiment:

1. The cow is pregnant and the test is positive.
2. The cow is pregnant, but the test is negative.
3. The cow is not pregnant, but the test is positive.
4. The cow is not pregnant, and the test is negative.

The prior probability of the event that the cow is pregnant is 0.05, the probability of the event that the test is positive, when the cow indeed is pregnant, is 0.98 and the probability that the test is negative, when the cow is not pregnant, is 0.999. The test turns out to be positive. What is the posterior probability of the cow being pregnant?

Exercise 1.8. Consider the following two experiments: One consists in throwing a red six-sided die, and one consists in throwing a blue six-sided die. We let R be a variable representing the roll of the red die, having a set of states $\{r1, r2, r3, r4, r5, r6\}$, and B be a variable representing the roll of the blue die (states $\{b1, b2, b3, b4, b5, b6\}$). Assume that the red die is fair so that $P(R = r1) = \cdots = P(R = r6) = \frac{1}{6}$, and that the variable for the blue die has probabilities $P(B = b1) = P(B = b2) = P(B = b3) = \frac{1}{12}$ and $P(B = b4) = P(B = b5) = P(B = b6) = \frac{1}{4}$. Give an example of a sample space for an experiment consisting of throwing both the red and the blue die. Using $P(R)$ and $P(B)$, what is the probability distribution for your sample space?

Exercise 1.9. Consider the sample space $\mathcal{S}_B$ from Exercise 1.3, with probability distribution as defined in Table 1.15. Recast the sample space as variables. What is the probability distribution for each variable?

Exercise 1.10. Prove the fundamental rule for variables:

$$P(A, B) = P(A \mid B)P(B) .$$

Exercise 1.11. Calculate $P(A)$, $P(B)$, $P(A \mid B)$, and $P(B \mid A)$ from the table for $P(A, B)$ (Table 1.16).

	b_1	b_2	b_3
a_1	0.05	0.10	0.05
a_2	0.15	0.00	0.25
a_3	0.10	0.20	0.10

Table 1.16. $P(A, B)$ for Exercise 1.11.

Exercise 1.12. Table 1.17 describes a test T for an event A. The number 0.01 is the frequency of *false negatives*, and the number 0.001 is the frequency of *false positives*.

(*i*) The police can order a blood test on drivers under the suspicion of having consumed too much alcohol. The test has the above characteristics. Experience says that 20% of the drivers under suspicion do in fact drive with too much alcohol in their blood. A suspicious driver has a positive blood test. What is the probability that the driver is guilty of driving under the influence of alcohol?

(*ii*) The police block a road, take blood samples of all drivers, and use the same test. It is estimated that one out of 1,000 drivers have too much alcohol in their blood. A driver has a positive test result. What is the probability that the driver is guilty of driving under the influence of alcohol?

	$A = yes$	$A = no$
$T = yes$	0.99	0.001
$T = no$	0.01	0.999

Table 1.17. Table for Exercise 1.12. Conditional probabilities $P(T \mid A)$ characterizing test T for A.

Exercise 1.13. In Table 1.18, a joint probability table for the binary variables A, B, and C is given.

- Calculate $P(B, C)$ and $P(B)$.
- Are A and C independent given B?

	b_1	b_2
a_1	$(0.006, 0.054)$	$(0.048, 0.432)$
a_2	$(0.014, 0.126)$	$(0.032, 0.288)$

Table 1.18. $P(A, B, C)$ for Exercise 1.13.

Exercise 1.14. Write a short algorithm that given an $n \times m$ potential $\phi(A, B)$ calculates $\sum_A \phi$. Use your algorithm on the joint probability table $P(A, B)$ in Table 1.2 and on the conditional probability table $P(A|B)$ in Table 1.1.

Exercise 1.15. Prove that the associative, commutative, and distributive laws hold for potentials.

Exercise 1.16. Let $\phi(x) = ax$ be a distribution on $[0, 1]$. Determine a. What are the mean and the variance of ϕ?

Exercise 1.17. Let $\phi(x) = a\sin(x)$ be a distribution on $[0, \pi]$. Determine a and the mean of ϕ.

Part I

Probabilistic Graphical Models

2

Causal and Bayesian Networks

In this chapter we introduce causal networks, which are the basic graphical feature for (almost) everything in this book. We give rules for reasoning about relevance in causal networks; is knowledge of A relevant for my belief about B? These sections deal with reasoning under uncertainty in general. Next, Bayesian networks are defined as causal networks with the strength of the causal links represented as conditional probabilities. Finally, the chain rule for Bayesian networks is presented. The chain rule is the property that makes Bayesian networks a very powerful tool for representing domains with inherent uncertainty. The sections on Bayesian networks assume knowledge of probability calculus as laid out in Sections 1.1–1.4.

2.1 Reasoning Under Uncertainty

2.1.1 Car Start Problem

The following is an example of the type of reasoning that humans do daily.

"In the morning, my car will not start. I can hear the starter turn, but nothing happens. There may be several reasons for my problem. I can hear the starter roll, so there must be power from the battery. Therefore, the most-probable causes are that the fuel has been stolen overnight or that the spark plugs are dirty. It may also be due to dirt in the carburetor, a loose connection in the ignition system, or something more serious. To find out, I first look at the fuel meter. It shows half full, so I decide to clean the spark plugs."

To have a computer do the same kind of reasoning, we need answers to questions such as, "What made me conclude that among the probable causes "stolen fuel", and "dirty spark plugs" are the two most-probable causes?" or "What made me decide to look at the fuel meter, and how can an observation concerning fuel make me conclude on the seemingly unrelated spark plugs?" To be more precise, we need ways of representing the problem and ways of performing inference in this representation such that a computer can simulate this kind of reasoning and perhaps do it better and faster than humans.

For propositional logic, Boolean logic is the representation framework, and various derived structures, such as truth tables and binary decision diagrams, have been invented together with efficient algorithms for inference.

In logical reasoning, we use four kinds of logical connectives: conjunction, disjunction, implication, and negation. In other words, simple logical statements are of the kind, "if it rains, then the lawn is wet," "both John and Mary have caught the flu," "either they stay at home or they go to the cinema," or "the lawn is not wet." From a set of logical statements, we can deduce new statements. From the two statements "if it rains, then the lawn is wet" and "the lawn is not wet," we can infer that it is not raining.

When we are dealing with uncertain events, it would be nice if we could use similar connectives with certainties rather than truth values attached, so we may extend the truth values of propositional logic to "certainties," which are numbers between 0 and 1. A certainty 0 means "certainly not true," and the higher the number, the higher the certainty. Certainty 1 means "certainly true."

We could then work with statements such as, "if I take a cup of coffee while on break, I will with certainty 0.5 stay awake during the next lecture" or "if I take a short walk during the break, I will with certainty 0.8 stay awake during the next lecture." Now, suppose I take a walk as well as have a cup of coffee. How certain can I be to stay awake? To answer this, I need a rule for how to *combine* certainties. In other words, I need a function that takes the two certainties 0.5 and 0.8 and returns a number, which should be the certainty resulting from combining the certainty from the two statements.

The same is needed for chaining: "if a then b with certainty x," and "if b then c with certainty y." I know a, so what is the certainty of c?

It has turned out that any function for combination and chaining will in some situations lead to wrong conclusions.

Another problem, which is also a problem for logical reasoning, is abduction: I have the rule "a woman has long hair with certainty 0.7." I see a long-haired person. What can I infer about the person's sex?

2.1.2 A Causal Perspective on the Car Start Problem

A way of structuring a situation for reasoning under uncertainty is to construct a graph representing causal relations between events.

Example 2.1 (A reduced Car Start Problem).

To simplify the situation, assume that we have the events $\{yes, no\}$ for *Fuel?*, $\{yes, no\}$ for *Clean Spark Plugs?*, $\{full, \frac{1}{2}, empty\}$ for *Fuel Meter*, and $\{yes, no\}$ for *Start?*. In other words, the events are clustered around variables, each with a set of outcomes, also called *states*. We know that the state of *Fuel?* and the state of *Clean Spark Plugs?* have a causal impact on the state of *Start?*. Also, the state of *Fuel?* has an impact on the state of *Fuel Meter Standing*. This is represented by the graph in Figure 2.1.

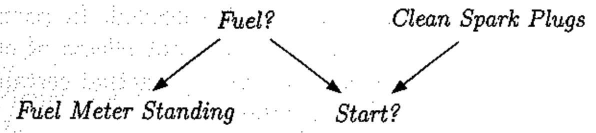


Fig. 2.1. A causal network for the reduced Car Start Problem.

If we add a direction from *no* to *yes* inside each variable (and from *empty* to *full*), we can also represent directions of the impact. For the present situation, we can say that all the impacts are positive (with the direction); that is, the more the certainty of the cause is moved in a positive direction, the more the certainty of the affected variable will also be moved in a positive direction. To indicate this, we can label the links with the sign "+" as is done in Figure 2.2.

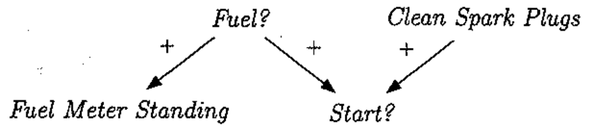


Fig. 2.2. A causal network for the reduced Car Start Problem with a sign indicating direction of impact.

We can use the graph in Figure 2.2 to perform some reasoning. Obviously, if I know that the spark plugs are not clean, then the certainty for no start will increase. However, my situation is the opposite. I realize that I have a start problem. As my certainty on *Start?* is moved in a negative direction, I find the possible causes (*Clean Spark Plugs?* and *Fuel?*) for such a move more certain; that is, the sign "+" is valid for both directions. Now, because the certainty on for *Fuel?* = *no* has increased, I will have a higher expectation that *Fuel Meter Standing* is in state *empty*.

The movement of the certainty for *Fuel Meter Standing* tells me that by reading the fuel meter I will get information related to the start problem. I read the fuel meter, it says $\frac{1}{2}$, and reasoning backward yields that the certainty on *Fuel?* is moved in a negative direction.

So far, the reasoning has been governed by simple rules that can easily be formalized. The conclusion is harder: "Lack of fuel does not seem to be the reason for my start problem, so most probably the spark plugs are not clean." Is there a formalized rule that allows this kind of reasoning on a causal network to be computerized? We will return to this problem in Section 2.2.

Note: The reasoning has focused on changes of certainty. In certainty calculus, if the actual certainty of a specific event must be calculated, then knowledge of certainties prior to any information is also needed. In particular, prior certainties are required for the events that are not effects of causes in the network. If, for example, my car cannot start, the actual certainty that the fuel has been stolen depends on my neighborhood.

2.2 Causal Networks and d-Separation

A causal network consists of a set of *variables* and a set of *directed links* (also called *arcs*) between variables. Mathematically, the structure is called a *directed graph*. When talking about the relations in a directed graph, we use the wording of family relations: if there is a link from A to B, we say that B is a *child* of A, and A is a *parent* of B.

The variables represent propositions (or sample spaces), see also Section 1.3. A variable can have any number of states (or outcomes). A variable may, for example, be the color of a car (states *blue*, *green*, *red*, *brown*), the number of children in a specific family (states 0, 1, 2, 3, 4, 5, 6, > 6), or a disease (states *bronchitis*, *tuberculosis*, *lung cancer*). Variables may have a countable or a continuous state set, but we consider only variables with a finite number of states (we shall return to the issue of continuous state spaces in Section 3.3.8).

In a causal network, a variable represents a set of possible states of affairs. A variable is in exactly one of its states; which one may be unknown to us.

As illustrated in Section 2.1.2, causal networks can be used to follow how a change of certainty in one variable may change the certainty for other variables. We present in this section a set of rules for that kind of reasoning. The rules are independent of the particular calculus for uncertainty.

Serial Connections

Consider the situation in Figure 2.3. Here A has an influence on B, which in turn has an influence on C. Obviously, evidence about A will influence the certainty of B, which then influences the certainty of C. Similarly, evidence about C will influence the certainty of A through B. On the other hand, if the state of B is known, then the channel is blocked, and A and C become independent; we say that A and C are d-separated given B. When the state of a variable is known, we say that the variable is *instantiated*.

We conclude that evidence may be transmitted through a serial connection unless the state of the variable in the connection is known.

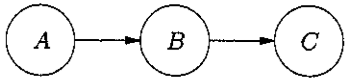


Fig. 2.3. Serial connection. When B is instantiated, it blocks communication between A and C.

Example 2.2. Figure 2.4 shows a causal model for the relations between *Rainfall* (*no*, *light*, *medium*, *heavy*), *Water level* (*low*, *medium*, *high*), and *Flooding* (*yes*, *no*). If I have not observed the water level, then knowing that there has been a flooding will increase my belief that the water level is high, which in turn will tell me something about the rainfall. The same line of reasoning holds in the other direction. On the other hand, if I already know the water level, then knowing that there has been flooding will not tell me anything new about rainfall.

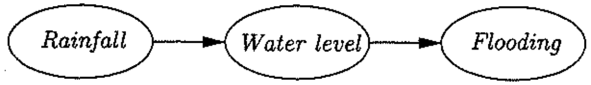


Fig. 2.4. A causal model for *Rainfall*, *Water level*, and *Flooding*.

Diverging Connections

The situation in Figure 2.5 is called a *diverging* connection. Influence can pass between all the children of A unless the state of A is known. That is, $B, C, \ldots, E$ are d-separated given A.

Evidence may be transmitted through a diverging connection unless it is instantiated.

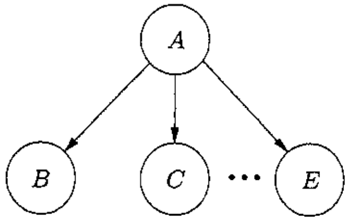


Fig. 2.5. Diverging connection. If A is instantiated, it blocks communication between its children.

Example 2.3. Figure 2.6 shows the causal relations between *Sex* (*male*, *female*), *length of hair* (*long*, *short*), and *stature* (<168 cm, $\geq$168 cm).

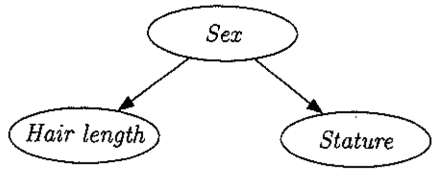


Fig. 2.6. Sex has an impact on length of hair as well as stature.

If we do not know the sex of a person, seeing the length of his/her hair will tell us more about the sex, and this in turn will focus our belief on his/her stature. On the other hand, if we know that the person is a man, then the length of his hair gives us no extra clue on his stature.

Converging Connections

A description of the situation in Figure 2.7 requires a little more care. If nothing is known about A except what may be inferred from knowledge of its parents $B, \dots, E$, then the parents are independent: evidence about one of them cannot influence the certainties of the others through A. Knowledge of one possible cause of an event does not tell us anything about the other possible causes. However, if anything is known about the consequences, then information on one possible cause may tell us something about the other causes. This is the *explaining away* effect illustrated in the car start problem: the car cannot start, and the potential causes include dirty spark plugs and an empty fuel tank. If we now get the information that there is fuel in the tank, then our certainty in the spark plugs being dirty will increase (since this will explain why the car cannot start). Conversely, if we get the information that there is no fuel on the car, then our certainty in the spark plugs being dirty will decrease (since the lack of fuel explains why the car cannot start). In Figure 2.8, two examples are shown. Observe that in the second example we observe only A indirectly through information about F; knowing the state of F tells us something about the state of E, which in turn tells us something about A.

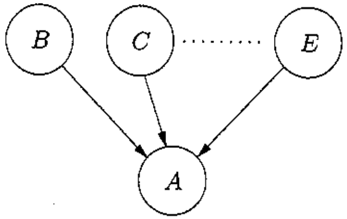


Fig. 2.7. Converging connection. If A changes certainty, it opens communication between its parents.

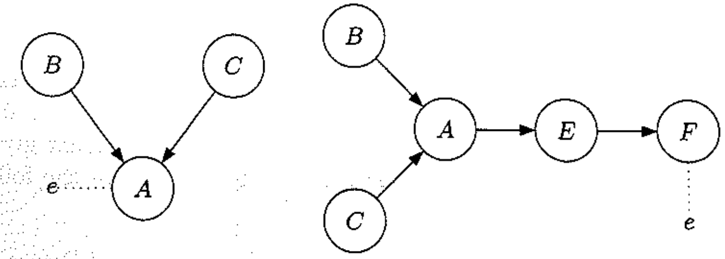


Fig. 2.8. Examples in which the parents of A are dependent. The dotted lines indicate insertion of evidence.

The conclusion is that evidence may be transmitted through a converging connection only if either the variable in the connection or one of its descendants has received evidence.

Remark: Evidence about a variable is a statement of the certainties of its states. If the variable is instantiated, we call it *hard* evidence; otherwise, it is called *soft*. In the example above, we can say that hard evidence about the variable F provides soft evidence about the variable A. Blocking in the case of serial and diverging connections requires hard evidence, whereas opening in the case of converging connections holds for all kinds of evidence.

Example 2.4. Figure 2.9 shows the causal relations among *Salmonella* infection, *flu*, *nausea*, and *pallor*.

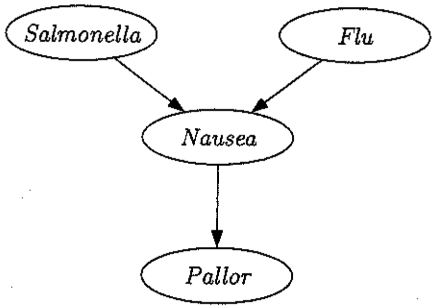


Fig. 2.9. Salmonella and flu may cause nausea, which in turn causes pallor.

If we know nothing of nausea or pallor, then the information on whether the person has a *Salmonella* infection will not tell us anything about flu. However, if we have noticed that the person is pale, then the information that he/she does not have a *Salmonella* infection will make us more ready to believe that he/she has the flu.

2.2.1 d-separation

The three preceding cases cover all ways in which evidence may be transmitted through a variable, and following the rules it is possible to decide for any pair of variables in a causal network whether they are independent given the evidence entered into the network. The rules are formulated in the following definition.

Definition 2.1 (d-separation). *Two distinct variables A and B in a causal network are* d-separated *("d" for "directed graph") if for all paths between A and B, there is an intermediate variable V (distinct from A and B) such that either*

- *the connection is serial or diverging and V is instantiated*
  or
- *the connection is converging, and neither V nor any of V's descendants have received evidence.*

If A and B are not d-separated, we call them d-connected.

Figure 2.10 gives an example of a larger network. The evidence entered at B and M represents instantiations. If evidence is entered at A, it may be transmitted to D. The variable B is blocked, so the evidence cannot pass through B to E. However, it may be passed to H and K. Since the child M of K has received evidence, evidence from H may pass to I and further to E, C, F, J, and L, so the path $A - D - H - K - I - E - C - F - J - L$ is a d-connecting path. Figure 2.11 gives two other examples.

Note that although A and B are d-connected, changes in the belief in A will not necessarily change the belief in B. To stress this difference, we will sometimes say that A and B are *structurally independent* if they are d-separated (see also Exercise 2.23).

In connection to d-separation, a special set of nodes for a node A is the so-called *Markov blanket* for A:

Definition 2.2. *The* Markov blanket *of a variable A is the set consisting of the parents of A, the children of A, and the variables sharing a child with A.*

The Markov blanket has the property that when instantiated, A is d-separated from the rest of the network (see Figure 2.12).

You may wonder why we have introduced d-separation as a definition rather than as a theorem. A theorem should be as follows.

Claim: If A and B are d-separated, then changes in the certainty of A have no impact on the certainty of B.

However, the claim cannot be established as a theorem without a more-precise description of the concept of "certainty." You can take d-separation as a property of human reasoning and require that any certainty calculus should comply with the claim.

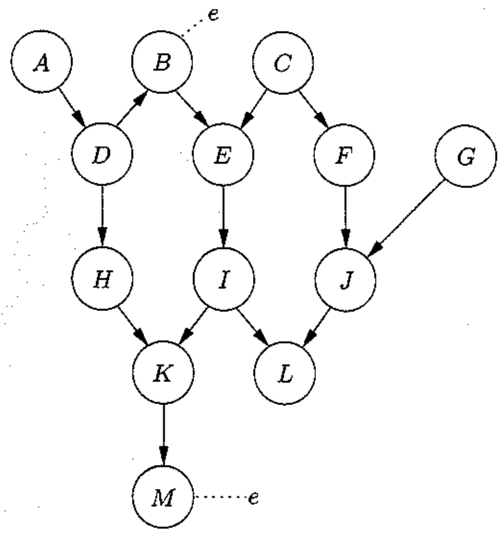
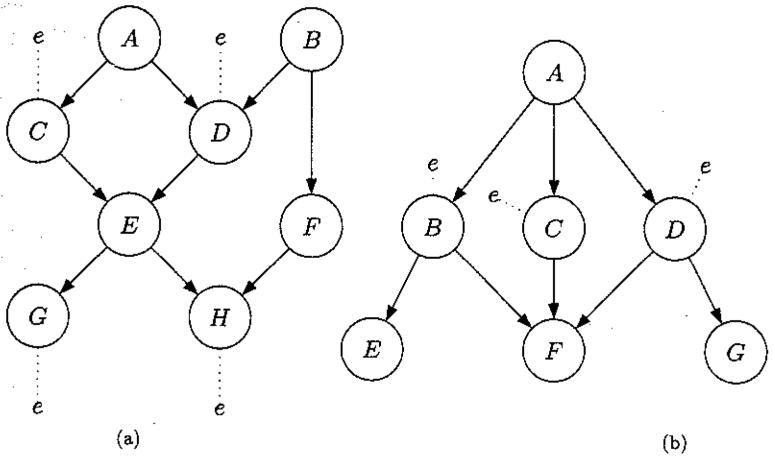


Fig. 2.10. A causal network with M and B instantiated. The node A is d-separated from G only.



(a)

(b)

Fig. 2.11. Causal networks with hard evidence entered (the variables are instantiated). (a) Although all neighbors of E are instantiated, it is d-connected to F, B, and A. (b) F is d-separated from the remaining uninstantiated variables.

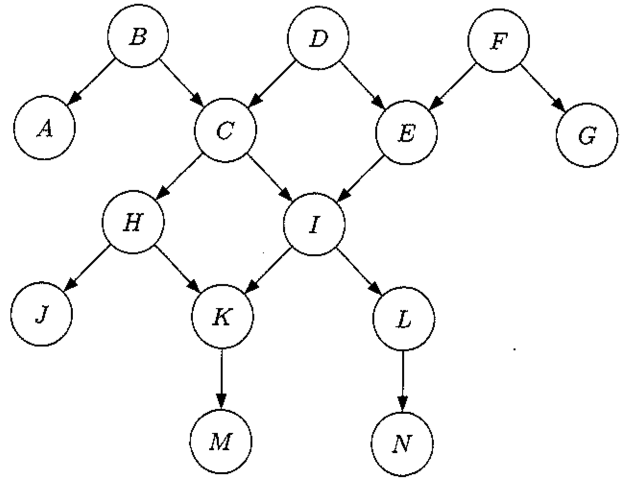


Fig. 2.12. The Markov blanket for I is $\{C, E, H, K, L\}$. Note that if only I's neighbors are instantiated, then J is not d-separated from I.

From the definition of d-separation we see that in order to test whether two variables, say A and B, are d-separated given hard evidence on a set of variables $\mathcal{C}$ you would have to check whether all paths connecting A and B are d-separating paths. An easier way of performing this test, without having to consider the various types of connections, is as follows: First you construct the so-called *ancestral graph* consisting of A, B, and $\mathcal{C}$ together with all nodes from which there is a directed path to either A, B, or $\mathcal{C}$ (see Figure 2.13(a)). Next, you insert an undirected link between each pair of nodes with a common child and then you make all links undirected. The resulting graph (see Figure 2.13(b)) is known as the *moral graph* for Figure 2.13(a). The moral graph can now be used to check whether A and B are d-separated given $\mathcal{C}$: if all paths connecting A and B intersect $\mathcal{C}$, then A and B are d-separated given $\mathcal{C}$.

The above procedure generalizes straightforwardly to the case in which we work with sets of variables rather than single variables: you just construct the ancestral graph using these sets of variables and perform the same steps as above: $\mathcal{A}$ and $\mathcal{B}$ are then d-separated given $\mathcal{C}$ if all paths connecting a variable in $\mathcal{A}$ with a variable in $\mathcal{B}$ intersect a variable in $\mathcal{C}$.

2.3 Bayesian Networks

2.3.1 Definition of Bayesian Networks

Causal relations also have a quantitative side, namely their *strength*. This can be expressed by attaching numbers to the links.

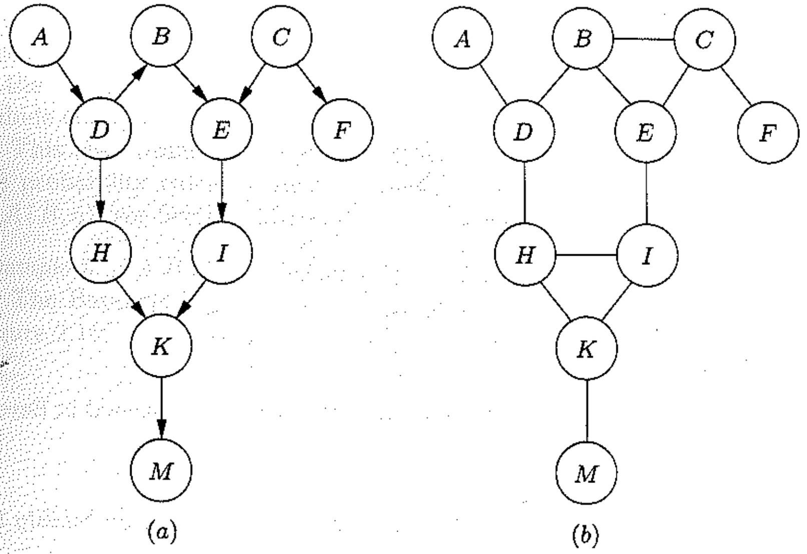


Fig. 2.13. To test whether A is d-separated from F given evidence on B and M in Figure 2.10, we first construct the ancestral graph for $\{A, B, F, M\}$ (figure (a)). Next we add an undirected link between pairs of nodes with a common child and then the direction is dropped on all links (figure (b)). In the resulting graph we have that the path $A - D - H - K - I - E - C - F$ does not intersect B and M, hence A and F are d-connected given B and M.

Let A be a parent of B. Using probability calculus, it would be natural to let $P(B \mid A)$ be the strength of the link. However, if C is also a parent of B, then the two conditional probabilities $P(B \mid A)$ and $P(B \mid C)$ alone do not give any clue about how the impacts from A and C interact. They may cooperate or counteract in various ways, so we need a specification of $P(B \mid A, C)$.

It may happen that the domain to be modeled contains causal feedback cycles (see Figure 2.14).

Feedback cycles are difficult to model quantitatively. For causal networks, no calculus has been developed that can cope with feedback cycles, but certain noncausal models have been proposed to deal with this issue. For Bayesian networks we require that the network does not contain cycles.

Definition 2.3. *A* Bayesian network *consists of the following:*

- *A set of* variables[1] *and a set of* directed edges *between variables.*
- *Each variable has a finite set of mutually exclusive states.*
- *The variables together with the directed edges form an* acyclic directed graph *(traditionally abbreviated DAG); a directed graph is* acyclic *if there is no directed path $A_1 \rightarrow \cdots \rightarrow A_n$ so that $A_1 = A_n$.*

[1] When we wish to emphasize that this kind of variable represents a sample space we call it a *chance variable*.

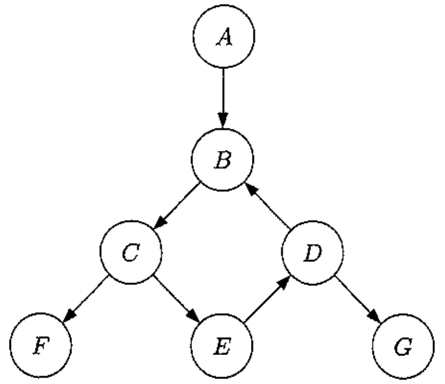


Fig. 2.14. A directed graph with a feedback cycle. This is not allowed in Bayesian networks.

- *To each variable A with parents $B_1, \ldots, B_n$, a conditional probability table $P(A \mid B_1, \ldots, B_n)$ is attached.*

Note that if A has no parents, then the table reduces to the unconditional probability table $P(A)$. For the DAG in Figure 2.15, the prior probabilities $P(A)$ and $P(B)$ must be specified. It has been claimed that prior probabilities are an unwanted introduction of bias to the model, and calculi have been invented in order to avoid it. However, as discussed in Section 2.1.2, prior probabilities are necessary not for mathematical reasons but because prior certainty assessments are an integral part of human reasoning about certainty (see also Exercise 1.12).

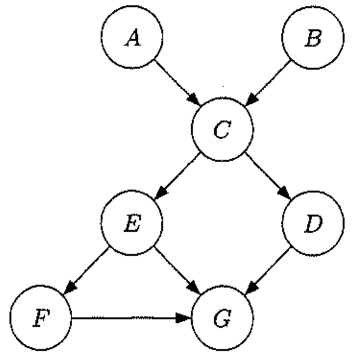


Fig. 2.15. A directed acyclic graph (DAG). The probabilities to specify are $P(A)$, $P(B)$, $P(C \mid A, B)$, $P(E \mid C)$, $P(D \mid C)$, $P(F \mid E)$, and $P(G \mid D, E, F)$.

The definition of Bayesian networks does not refer to causality, and there is no requirement that the links represent causal impact. That is, when building the structure of a Bayesian network model, we need not insist on having the links go in a causal direction. However, we then need to check the model's d-separation properties and ensure that they correspond to our perception of the world's conditional independence properties. The model should not include conditional independences that do not hold in the real world.

This also means that if A and B are d-separated given evidence e, then the probability calculus used for Bayesian networks must yield $P(A \mid e) = P(A \mid B, e)$ (see Section 2.3.2).

Example 2.5 (A Bayesian network for the Car Start Problem).

The Bayesian network for the reduced Car Start Problem is the one in Figure 2.16.

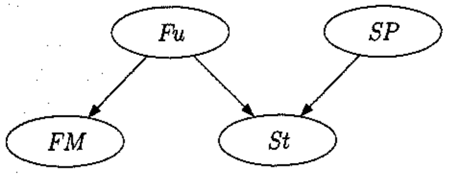


Fig. 2.16. The causal network for the reduced car start problem. We have used the abbreviations *Fu (Fuel?), SP (Clean Spark Plugs?), St (Start?),* and *FM (Fuel Meter Standing).*

For the quantitative modeling, we need the probability assessments $P(Fu)$, $P(SP)$, $P(St \mid Fu, SP), P(FM \mid Fu)$. To avoid having to deal with numbers that are too small, let $P(Fu) = (0.98, 0.02)$ and $P(SP) = (0.96, 0.04)$. The remaining tables are given in Table 2.1. Note that the table for $P(FM \mid Fu)$ reflects the fact that the fuel meter may be malfunctioning, and the table for $P(St \mid Fu, SP)$ leaves room for causes other than no fuel and dirty spark plugs by assigning $P(St = no \mid Fu = yes, SP = yes) > 0$.

2.3.2 The Chain Rule for Bayesian Networks

Let $\mathcal{U} = \{A_1, \ldots, A_n\}$ be a universe of variables. If we have access to the joint probability table $P(\mathcal{U}) = P(A_1, \ldots, A_n)$, then we can also calculate $P(A_i)$ as well as $P(A_i \mid e)$, where e is evidence about some of the variables in the Bayesian network (see, e.g., Section 1.3.1). However, $P(\mathcal{U})$ grows exponentially with the number of variables, and $\mathcal{U}$ need not be very large before the table becomes intractably large. Therefore, we look for a more compact *representation* of $P(\mathcal{U})$, i.e., a way of storing information from which $P(\mathcal{U})$ can be calculated if needed.

Let BN be a Bayesian network over $\mathcal{U}$, and let $P(\mathcal{U})$ be a probability distribution reflecting the properties specified by BN: (i) the conditional probabilities for a variable given its parents in $P(\mathcal{U})$ must be as specified in BN, and (ii) if the variables A and B are d-separated in BN given the set $\mathcal{C}$, then A and B are independent given $\mathcal{C}$ in $P(\mathcal{U})$.

	$Fu = yes$	$Fu = no$
$FM = full$	0.39	0.001
$FM = \frac{1}{2}$	0.60	0.001
$FM = empty$	0.01	0.998

$P(FM \mid Fu)$

	$Fu = yes$	$Fu = no$
$Sp = yes$	(0.99, 0.01)	(0,1)
$Sp = no$	(0.01, 0.99)	(0,1)

$P(St \mid Fu, Sp)$

Table 2.1. Conditional probabilities for the model in Figure 2.16. The numbers (x, y) in the lower table represent $(St = yes, St = no)$.

Based on these two properties, what other properties can be deduced about $P(\mathcal{U})$? If the universe consists of only one variable A, then BN specifies $P(A)$, and $P(\mathcal{U})$ is uniquely determined. We shall show that this holds in general.

For probability distributions over sets of variables, we have an equation called *the chain rule*. For Bayesian networks this equation has a special form. First we state the general chain rule:

Proposition 2.1 (The general chain rule). *Let $\mathcal{U} = \{A_1, \ldots, A_n\}$ be a set of variables. Then for any probability distribution $P(\mathcal{U})$ we have*

$$P(\mathcal{U}) = P(A_n \mid A_1, \ldots, A_{n-1})P(A_{n-1} \mid A_1, \ldots, A_{n-2}) \ldots P(A_2 \mid A_1)P(A_1).$$

Proof. Iterative use of the fundamental rule:

$$\begin{aligned} P(\mathcal{U}) &= P(A_n \mid A_1, \ldots, A_{n-1})P(A_1, \ldots, A_{n-1}), \\ P(A_1, \ldots, A_{n-1}) &= P(A_{n-1} \mid A_1, \ldots, A_{n-2})P(A_1, \ldots, A_{n-2}), \\ &\vdots \\ P(A_1, A_2) &= P(A_2 \mid A_1)P(A_1). \end{aligned}$$

□

Theorem 2.1 (The chain rule for Bayesian networks). *Let BN be a Bayesian network over $\mathcal{U} = \{A_1, \ldots, A_n\}$. Then BN specifies a unique joint probability distribution $P(\mathcal{U})$ given by the product of all conditional probability tables specified in BN:*

$$P(\mathcal{U}) = \prod_{i=1}^{n} P(A_i \mid \mathrm{pa}(A_i)),$$

where $\mathrm{pa}(A_i)$ are the parents of A_i in BN, and $P(\mathcal{U})$ reflects the properties of BN.

Proof. First we should show that $P(\mathcal{U})$ is indeed a probability distribution. That is, we need to show that Axioms 1–3 hold. This is left as an exercise (see Exercise 2.15).

Next we prove that the specification of BN is consistent, so that $P(\mathcal{U})$ reflects the properties of BN. It is not hard to prove that the probability distribution specified by the product in the chain rule reflects the conditional probabilities from BN (see Exercise 2.16). We also need to prove that the product reflects the d-separation properties. This is done through induction in the number of variables in BN.

When BN has one variable, it is obvious that the d-separation properties specified by BN hold for the product of all specified conditional probabilities.

Assume that for any Bayesian network with $n-1$ variables and a distribution $P(\mathcal{U})$ specified as the product of all conditional probabilities, it holds that if A and B are d-separated given $\mathcal{C}$, then $P(A \mid B, \mathcal{C}) = P(A \mid \mathcal{C})$. Let BN be a Bayesian network with n variables $\{A_1, \ldots, A_n\}$. Assume that A_n has no children and let BN' be the result of removing A_n from BN. Clearly BN' is a Bayesian network with the same conditional probability distributions as BN (except for A_n) and with the same d-separation properties over $\{A_1, \ldots, A_{n-1}\}$ as BN. Moreover,

$$\begin{aligned} P(\mathcal{U} \setminus \{A_n\}) = \sum_{A_n} P(\mathcal{U}) &= \sum_{A_n} \prod_{i=1}^{n} P(A_i \mid \mathrm{pa}(A_i)) \\ &= \prod_{i=1}^{n-1} P(A_i \mid \mathrm{pa}(A_i)) \sum_{A_n} P(A_n \mid \mathrm{pa}(A_n)) \\ &= \prod_{i=1}^{n-1} P(A_i \mid \mathrm{pa}(A_i)) 1 = \prod_{i=1}^{n-1} P(A_i \mid \mathrm{pa}(A_i)), \end{aligned}$$

and by the induction hypothesis $P(\mathcal{U} \setminus \{A_n\})$ reflects the properties of BN'. Now, if A and B are d-separated given $\mathcal{C}$ in BN, then they are also d-separated in BN', and therefore $P(A \mid B, \mathcal{C}) = P(A \mid \mathcal{C})$. To prove that it also holds for d-separation properties involving A_n, we consider the case in which $A_n \in \mathcal{C}$ and the case in which $A = A_n$. For the first case we have that since A_n participates only in a converging connection, it holds that if A and B are d-separated given $\mathcal{C}$, then they are also d-separated given $\mathcal{C} \setminus \{A_n\}$ and we get the situation above. For the second case, we first note that

$$P(A_n \mid B, \mathcal{C}) = \sum_{\mathrm{pa}(A_n)} P(A_n \mid B, \mathcal{C}, \mathrm{pa}(A_n))P(\mathrm{pa}(A_n) \mid B, \mathcal{C}).$$

Now, if A_n and B are d-separated given $\mathcal{C}$, then $\mathrm{pa}(A_n)$ and B are also d-separated given $\mathcal{C}$, and since A_n is not involved, we have $P(\mathrm{pa}(A_n) \mid B, \mathcal{C}) =$

$P(\mathrm{pa}(A_n) \mid \mathcal{C})$. So we need to prove only that $P(A_n \mid \mathcal{B}, \mathcal{C}, \mathrm{pa}(A_n)) = P(A_n \mid \mathrm{pa}(A_n))$. Using the fundamental rule and the chain rule, we get

$$\begin{aligned} P(A_n \mid \mathcal{B}, \mathcal{C}, \mathrm{pa}(A_n)) &= \frac{P(A_n, \mathcal{B}, \mathcal{C}, \mathrm{pa}(A_n))}{P(\mathcal{B}, \mathcal{C}, \mathrm{pa}(A_n))} = \frac{\sum_{\mathcal{U} \setminus \{A_n, \mathcal{B}, \mathcal{C}, \mathrm{pa}(A_n)\}} P(\mathcal{U})}{\sum_{\mathcal{U} \setminus \{\mathcal{B}, \mathcal{C}, \mathrm{pa}(A_n)\}} P(\mathcal{U})} \\ &= \frac{\sum_{\mathcal{U} \setminus \{A_n, \mathcal{B}, \mathcal{C}, \mathrm{pa}(A_n)\}} \prod_{i=1}^{n} P(A_i \mid \mathrm{pa}(A_i))}{\sum_{\mathcal{U} \setminus \{\mathcal{B}, \mathcal{C}, \mathrm{pa}(A_n)\}} \prod_{i=1}^{n} P(A_i \mid \mathrm{pa}(A_i))} \\ &= \frac{P(A_n \mid \mathrm{pa}(A_n)) \sum_{\mathcal{U} \setminus \{A_n, \mathcal{B}, \mathcal{C}, \mathrm{pa}(A_n)\}} \prod_{i=1}^{n-1} P(A_i \mid \mathrm{pa}(A_i))}{\sum_{\mathcal{U} \setminus \{A_n, \mathcal{B}, \mathcal{C}, \mathrm{pa}(A_n)\}} \prod_{i=1}^{n-1} P(A_i \mid \mathrm{pa}(A_i)) \sum_{A_n} P(A_n \mid \mathrm{pa}(A_n))} \\ &= \frac{P(A_n \mid \mathrm{pa}(A_n)) \sum_{\mathcal{U} \setminus \{A_n, \mathcal{B}, \mathcal{C}, \mathrm{pa}(A_n)\}} \prod_{i=1}^{n-1} P(A_i \mid \mathrm{pa}(A_i))}{\sum_{\mathcal{U} \setminus \{A_n, \mathcal{B}, \mathcal{C}, \mathrm{pa}(A_n)\}} \prod_{i=1}^{n-1} P(A_i \mid \mathrm{pa}(A_i)) \mathbf{1}} \\ &= P(A_n \mid \mathrm{pa}(A_n)). \end{aligned}$$

To prove uniqueness, let $\{A_1, \ldots, A_n\}$ be a topological ordering of the variables. Then, for each variable A_i with parents $\mathrm{pa}(A_i)$ we have that A_i is d-separated from $\{A_1, \ldots, A_{i-1}\} \setminus \mathrm{pa}(A_i)$ given $\mathrm{pa}(A_i)$ (see Exercise 2.11). This means that for any distribution P reflecting the specifications by BN we must have $P(A_i \mid A_1, \ldots, A_{i-1}) = P(A_i \mid \mathrm{pa}(A_i))$. Substituting this in the general chain rule yields that any distribution reflecting the specifications by BN must be the product of the conditional probabilities specified in BN. □

The chain rule yields that a Bayesian network is a compact representation of a joint probability distribution. The following example illustrates how to exploit that for reasoning under uncertainty.

Example 2.6 (The Car Start Problem revisited).

In this example, we apply the rules of probability calculus to the Car Start Problem. This is done to illustrate that probability calculus can be used to perform the reasoning in the example, in particular, explaining away. In Chapter 4, we give general algorithms for probability updating in Bayesian networks. We will use the Bayesian network from Example 2.5 to perform the reasoning in Section 2.1.1.

We will use the joint probability table for the reasoning. The joint probability table is calculated from the chain rule for Bayesian networks,

$$P(Fu, FM, SP, St) = P(Fu)P(SP)P(FM \mid Fu)P(St \mid Fu, SP).$$

The result is given in Tables 2.2 and 2.3.

The evidence $St = no$ tells us that we are in the context of Table 2.3. By marginalizing FM and Fu out of Table 2.3 (summing each row), we get

$$P(SP, St = no) = (0.02864, 0.03965).$$

	$FM = full$	$FM = \frac{1}{2}$	$FM = empty$
$Sp = yes$	$(0.363, 0)$	$(0.559, 0)$	$(0.0093, 0)$
$Sp = no$	$(0.00015, 0)$	$(0.00024, 0)$	$(3.9 \cdot 10^{-6}, 0)$

Table 2.2. The joint probability table for $P(Fu, FM, SP, St = yes)$.

	$FM = full$	$FM = \frac{1}{2}$	$FM = empty$
$Sp = yes$	$(0.00367, 1.9 \cdot 10^{-5})$	$(0.00564, 1.9 \cdot 10^{-5})$	$(9.4 \cdot 10^{-5}, 0.0192)$
$Sp = no$	$(0.01514, 8 \cdot 10^{-7})$	$(0.0233, 8 \cdot 10^{-7})$	$(0.000388, 0.000798)$

Table 2.3. The joint probability table for $P(Fu, FM, SP, St = no)$. The numbers (x, y) in the table represent $(Fu = yes, Fu = no)$.

We get the conditional probability $P(SP \mid St = no)$ by dividing by $P(St = no)$. This is easy, since $P(St = no) = P(SP = yes, St = no) + P(SP = no, St = no) = 0.02864 + 0.03965 = 0.06829$, and we get

$$P(SP \mid St = no) = \left(\frac{0.02864}{0.06829}, \frac{0.03965}{0.06829}\right) = (0.42, 0.58).$$

Another way of saying this is that the distribution we end up with will be a set of numbers that sum to 1. If they do not, normalize by dividing by the sum.

In the same way, we get $P(Fu \mid St = no) = (0.71, 0.29)$.

Next, we get the information that $FM = \frac{1}{2}$, and the context for calculation is limited to the part with $FM = \frac{1}{2}$ and $St = no$. The numbers are given in Table 2.4.

	$Fu = yes$	$Fu = no$
$Sp = yes$	0.00564	$1.9 \cdot 10^{-5}$
$Sp = no$	0.0233	$8 \cdot 10^{-7}$

Table 2.4. $P(Fu, SP, St = no, FM = \frac{1}{2})$.

By marginalizing Sp out and normalizing, we get $P(Fu|St = no, FM = \frac{1}{2}) = (0.999, 0.001)$, and by marginalizing Fu out and normalizing we get $P(SP|St = no, FM = \frac{1}{2}) = (0.196, 0.804)$. The probability of $SP = yes$ increased by observing $FM = \frac{1}{2}$, so the calculus did catch the explaining away effect.

2.3.3 Inserting Evidence

Bayesian networks are used for calculating new probabilities when you get new information. The information so far has been of the type "$A = a$," where A is a variable and a is a state of A. Let A have n states with $P(A) = (x_1, \ldots, x_n)$, and assume that we get the information e that A can be only in state i or j. This statement expresses that all states except i and j are impossible, and we have the probability distribution $P(A,e) = (0, \ldots, 0, x_i, 0, \ldots, 0, x_j, 0, \ldots, 0)$. Note that $P(e)$, the prior probability of e, is obtained by marginalizing A out of $P(A,e)$. Note also that $P(A,e)$ is the result of multiplying $P(A)$ by $(0, \ldots, 0, 1, 0, \ldots, 0, 1, 0, \ldots, 0)$, where the 1's are at the i'th and j'th places.

Definition 2.4. *Let A be a variable with n states. A* finding *on A is an n-dimensional table of zeros and ones.*

To distinguish between the statement e, "A is in either state i or j," and the corresponding 0/1-finding vector, we sometimes use the boldface notation $\mathbf{e}$ for the finding. Semantically, a finding is a statement that certain states of A are impossible.

Now, assume that you have a joint probability table, $P(\mathcal{U})$, and let $\mathbf{e}$ be the preceding finding. The joint probability table $P(\mathcal{U}, e)$ is the table obtained from $P(\mathcal{U})$ by replacing all entries with A not in state i or j by the value zero and leaving the other entries unchanged. This is the same as multiplying $P(\mathcal{U})$ by $\mathbf{e}$,

$$P(\mathcal{U}, e) = P(\mathcal{U}) \cdot \mathbf{e}.$$

Note that $P(e) = \sum_U P(\mathcal{U}, e) = \sum_U (P(\mathcal{U}) \cdot \mathbf{e})$. Using the chain rule for Bayesian networks, we have the following theorem.

Theorem 2.2. *Let BN be a Bayesian network over the universe $\mathcal{U}$, and let $\mathbf{e}_1, \ldots, \mathbf{e}_m$ be findings. Then*

$$P(\mathcal{U}, e) = \prod_{A \in \mathcal{U}} P(A \,|\, pa(A)) \cdot \prod_{i=1}^{m} \mathbf{e}_i,$$

and for $A \in \mathcal{U}$ we have

$$P(A \,|\, e) = \frac{\sum_{\mathcal{U} \setminus \{A\}} P(\mathcal{U}, e)}{P(e)}.$$

Some types of evidence cannot be represented as findings. You may, for example, receive a statement from someone that the chance of A being in state a_1 is twice as high as for a_2. This type of evidence is called *likelihood evidence*. It is possible to treat this kind of evidence in Bayesian networks. The preceding statement is then represented by the distribution $(0.67, 0.33)$, and Theorem 2.2 still holds. However, because it is unclear what it means that a likelihood statement is true, $P(e)$ cannot be interpreted as the probability of the evidence, and $P(\mathcal{U}, e)$ therefore has an unclear semantics. We will not deal further with likelihood evidence.

2.3.4 Calculating Probabilities in Practice

As described in Section 2.3.3 and illustrated in Example 2.6, probability updating in Bayesian networks can be performed using the chain rule to calculate $P(\mathcal{U})$, the joint probability table of the universe. However, $\mathcal{U}$ need not be large before $P(\mathcal{U})$ becomes intractably large. In this section, we illustrate how the calculations can be performed without having to deal with the full joint table. In Chapter 4, we give a detailed treatment of algorithms for probability updating.

Consider the Bayesian network in Figure 2.17, and assume that all variables have ten states. Assume that we have the evidence $e = \{D = d, F = f\}$, and we wish to calculate $P(A \,|\, e)$.

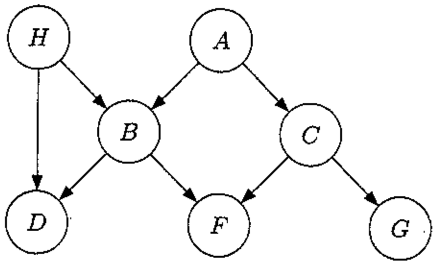


Fig. 2.17. A Bayesian network.

From the chain rule we have

$$\begin{aligned} P(\mathcal{U}, e) &= P(A, B, C, d, f, G, H) \\ &= P(A)P(H)P(B \,|\, A, H)P(C \,|\, A)P(d \,|\, B, H)P(f \,|\, B, C)P(G \,|\, C), \end{aligned}$$

where for example $P(d \,|\, B, H)$ denotes the table over B and H resulting from fixing the D-entry to the state d. We say that the conditional probability table has been *instantiated* to $D = d$. Notice that we need not calculate the full table $P(\mathcal{U})$ with 10^7 entries. If we wait until evidence is entered, we will in this case need to work with a table with only 10^5 entries. Later, we see that we need not work with tables larger than 1000 entries.

To calculate $P(A, e)$, we marginalize the variables B, C, G, and H out of $P(A, B, C, d, f, G, H)$. The order in which we marginalize does not affect the result (Section 1.4), so let us start with G; that is, we wish to calculate

$$\begin{aligned} &\sum_G P(A, B, C, d, f, G, H) \\ &\quad = \sum_G P(A)P(H)P(B \,|\, A, H)P(C \,|\, A)P(d \,|\, B, H)P(f \,|\, B, C)P(G \,|\, C). \end{aligned}$$

In the right-hand product, only the last table contains G in its domain, and due to the distributive law (Section 1.4) we have

$$\begin{aligned}&\sum_G P(A,B,C,d,f,G,H)\\ &\quad = P(A)P(H)P(B\,|\,A,H)P(C\,|\,A)P(d\,|\,B,H)P(f\,|\,B,C)\sum_G P(G\,|\,C),\end{aligned}$$

and we need only calculate $\sum_G P(G\,|\,C)$. Actually, for each state c of C, we have $\sum_G P(G\,|\,c) = 1$; hence no calculations are necessary. We therefore get

$$\begin{aligned}P(A,B,C,d,f,H) &= \sum_G P(A,B,C,d,f,G,H)\\ &= P(A)P(H)P(B\,|\,A,H)P(C\,|\,A)P(d\,|\,B,H)P(f\,|\,B,C).\end{aligned}$$

Next, we marginalize H out. Using the distributive law again, we get

$$\begin{aligned}&\sum_H P(A,B,C,d,f,H)\\ &\quad = P(A)P(C\,|\,A)P(f\,|\,B,C)\sum_H P(H)P(B\,|\,A,H)P(d\,|\,B,H).\end{aligned}$$

We multiply the three tables $P(H), P(B\,|\,A,H)$, and $P(d\,|\,B,H)$, and we marginalize H out of the product. The result is a table $T(d,B,A)$, and we have

$$P(A,B,C,d,f) = P(A)P(C\,|\,A)P(f\,|\,B,C)T(d,B,A).$$

Finally, we calculate this product and marginalize B and C out of it.

Notice that we never work with a table of more than three variables (the table produced by multiplying $P(H)$, $P(B\,|\,A,H)$, and $P(d\,|\,B,H)$) compared to the five variables in $P(A,B,C,d,f,G,H)$.

The method we just used is called *variable elimination* and can be described in the following way: we start with a set $\mathcal{T}$ of tables, and whenever we wish to marginalize a variable X, we take from $\mathcal{T}$ all tables with X in their domains, calculate the product of them, marginalize X out of it, and place the resulting table in $\mathcal{T}$.

2.4 Graphical Models – Formal Languages for Model Specification

From a mathematical point of view, the basic property of Bayesian networks is the chain rule: a Bayesian network is a compact representation of the joint probability table over its universe. In this respect, a Bayesian network is one type of compact representation among many others. However, there is more to it than this: From a knowledge engineering point of view, a Bayesian network is a type of *graphical model*. The structure of the network is formulated in a graphical communication language for which the language features have a very simple semantics, namely causality. This does not mean that "causality" is an easy concept. It may be very difficult to experience causality, and philosophically the concept is not fully understood. However, most often humans can communicate sensibly about causal relations in a knowledge domain. Furthermore, the graphical specification also specifies the requirements for the quantitative part of the model (the conditional probabilities). In Chapter 3, we extend the modeling language, and in Part II we present other types of graphical models.

As mentioned, graphical models are communication languages. They consist of a qualitative part, where features from graph theory are used, and a quantitative part consisting of *potentials*, which are real-valued functions over sets of nodes from the graph; in Bayesian networks the potentials are conditional probability tables. The graphical part specifies the kind of potentials and their domains.

Graphical models can be used for interpersonal communication: The graphical specification is easy for humans to read, and it helps focus attention, for example in a group working jointly on building a model. For interpersonal communication, the semantics of the various graph-theoretic features must be rather welldefined if misunderstandings are to be avoided.

The next step in the use of graphical models has to do with communication to a computer. You wish to communicate a graphical model to a computer, and the computer should be able to process the model and give answers to various queries. In order to achieve this, the specification language must be formally defined with a well-defined syntax and semantics.

The first concern in constructing a graphical modeling language is to ensure that it is sufficiently welldefined so that it can be communicated to a computer. This covers the graphical part as well as the specification of potentials. The next concern is the scope of the language: what is the range of domains and tasks that you will be able to model with this language? The final concern is tractability: do you have algorithms such that in reasonable time the computer can process a model and query to provide answers?

The Bayesian network is a sufficiently welldefined language, and behind the graphical specification in the user interface, the computer systems for processing Bayesian networks have an alphanumeric specification language, which for some systems is open to the user. Actually, the language for Bayesian networks is a context-free language with a single context-sensitive aspect (no directed cycles).

The scope of the Bayesian network language is hard to define, but the examples in the next chapter show that it has a very broad scope.

Tractability is not a yes or no issue. As described in Chapter 4, there are algorithms for probability updating in Bayesian networks, but basically probability updating is NP-hard. This means that some models have an updating time exponential in the number of nodes.

On the other hand, the running times of the algorithms can be easily calculated without actually running them. In Chapter 4 and Part II, we treat complexity issues for the various graphical languages presented.

2.5 Summary

d-Separation in Causal Networks

Two distinct variables A and B in a causal network are d-separated if for all paths between A and B, there is an intermediate variable V (distinct from A and B) such that either

- the connection is serial or diverging, and V is instantiated, or
- the connection is converging, and neither V nor any of V's descendants have received evidence.

Definition of Bayesian Networks

A Bayesian network consists of the following:

- There is a set of *variables* and a set of *directed edges* between variables.
- Each variable has a finite set of mutually exclusive states.
- The variables together with the directed edges form an *acyclic directed graph* (DAG).
- To each variable A with parents $B_1, \ldots, B_n$ there is attached a conditional probability table $P(A \mid B_1, \ldots, B_n)$.

The Chain Rule for Bayesian Networks

Let BN be a Bayesian network over $\mathcal{U} = \{A_1, \ldots, A_n\}$. Then BN specifies a unique joint probability distribution $P(\mathcal{U})$ given by the product of all conditional probability tables specified in BN:

$$P(\mathcal{U}) = \prod_{i=1}^{n} P(A_i \mid \mathrm{pa}(A_i)),$$

where $\mathrm{pa}(A_i)$ are the parents of A_i in BN, and $P(\mathcal{U})$ reflects the properties of BN.

Admittance of d-Separation in Bayesian Networks

If A and B are d-separated in a Bayesian network with evidence e entered, then $P(A \mid B, e) = P(A \mid e)$.

Inserting Evidence

Let $\mathbf{e}_1, \ldots, \mathbf{e}_m$ be findings, and then

$$P(\mathcal{U}, e) = \prod_{i=1}^{n} P(A_i \mid \mathrm{pa}(A_i)) \prod_{j=1}^{m} \mathbf{e}_j$$

and

$$P(A \mid e) = \frac{\sum_{\mathcal{U} \setminus \{A\}} P(\mathcal{U}, e)}{P(e)}.$$

2.6 Bibliographical Notes

The connection between causation and conditional independence was studied by Spohn (1980), and later investigated with special focus on Bayesian networks in (Pearl, 2000). The concepts of causal network, d-connection, and the definition in Section 2.2.1 are due to Pearl (1986) and Verma (1987). A proof that Bayesian networks admit d-separation can be found in (Pearl, 1988) or in (Lauritzen, 1996). Geiger and Pearl (1988) proved that d-separation is the correct criterion for directed graphical models, in the sense that for any DAG, a probability distribution can be found for which the d-separation criterion is sound and complete. Meek (1995) furthermore proved that for a given DAG, the set of discrete probability distributions for which the d-separation criterion is not complete has measure zero. That is, given a random Bayesian network, there is almost no chance that it contains conditionally independent variables that cannot be read off the graph by d-separation. The method for discovering d-separation properties using ancestral graphs was first presented in (Lauritzen *et al.*, 1990).

Bayesian networks have a long history in statistics, and can be traced back at least to the work in (Minsky, 1963). In the first half of the 1980s they were introduced to the field of expert systems through work by Pearl (1982) and Spiegelhalter and Knill-Jones (1984). Some of the first real-world applications of Bayesian networks were Munin (Andreassen *et al.*, 1989, 1992) and Pathfinder (Heckerman *et al.*, 1992). The basis for the inference method presented in Section 2.3.4 originates from (D'Ambrosio, 1991) and was modified to the presented variable elimination in (Dechter, 1996). The fact that inference is NP-hard was proved in (Cooper, 1987).

2.7 Exercises

Exercise 2.1. To illustrate that simple rules cannot cope with uncertainty reasoning, consider the following two cases:

(i) I have an urn with a red ball and a white ball in it. If I add a red ball and shake it, what is the certainty of drawing a red ball in one draw? If I add a white ball instead, what is the certainty of drawing a red ball? If I combine the two actions, what is the certainty of drawing a red ball?
(ii) When shooting, I am more certain to hit the target if I close the left eye. I am also more certain to hit the target if I close the right eye. What is the combined certainty if I do both?

Exercise 2.2. Construct a causal network and follow the reasoning in the following story. Mr. Holmes is working in his office when he receives a phone call from his neighbor, who tells him that Holmes' burglar alarm has gone off. Convinced that a burglar has broken into his house, Holmes rushes to his car and heads for home. On his way, he listens to the radio, and in the news it is reported that there has been a small earthquake in the area. Knowing that earthquakes have a tendency to turn on burglar alarms, he returns to work.

Exercise 2.3. Consider the Car Start Problem in Section 2.1.1 with the causal network in Figure 2.1, and the following twist on the story: "I distinctly remember visiting the pump last night, so the fuel meter should be reading *full.* Since this is not the case, either there must be a leak in the tank, someone has stolen gasoline during the night, or the fuel meter is malfunctioning. Sniffing the air I smell no gasoline, so I conclude that a thief has been visiting last night or that the fuel meter is malfunctioning." Alter the causal network in Figure 2.1 to incorporate the above twist on the story.

Exercise 2.4. In the graphs in Figures 2.18 and 2.19, determine which variables are d-separated from A.

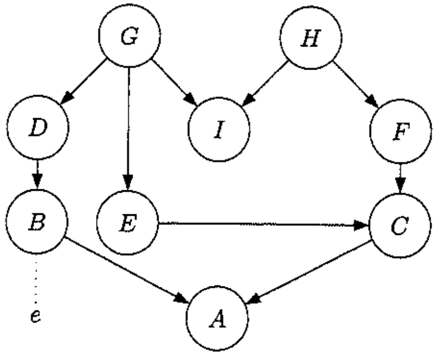


Fig. 2.18. Figure for Exercise 2.4.

Exercise 2.5. For each pair of variables in the causal network in Figure 2.1, state whether the variables can be d-separated, and if so which set(s) of variables that allow this.

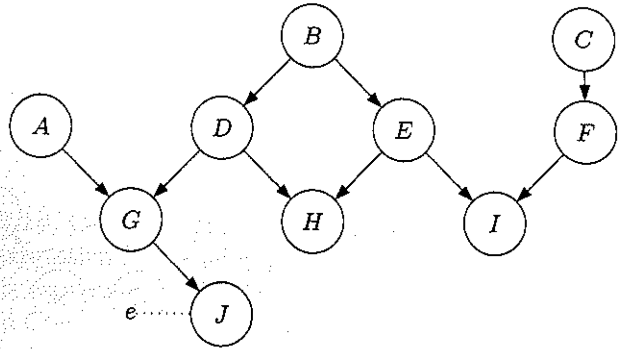


Fig. 2.19. Figure for Exercise 2.4.

Exercise 2.6. Consider the network in Figure 2.20. What are the minimal set(s) of variables required to d-separate C and E (that is, sets of variables for which no proper subset d-separates C and E)? What are the minimal set(s) of variables required to d-separate A and B? What are the maximal set(s) of variables that d-separate C and E (that is, sets of variables for which no proper superset d-separates C and E)? What are the maximal set(s) of variables that d-separate A and B?

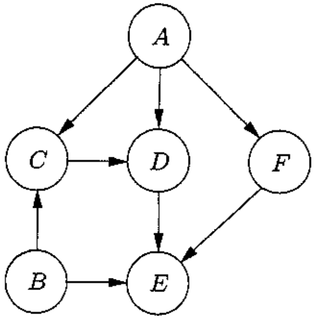


Fig. 2.20. A causal network for Exercise 2.6.

Exercise 2.7. Consider the network in Figure 2.20. What is the Markov blanket of each variable?

Exercise 2.8. Let A be a variable in a DAG. Assume that all variables in A's Markov blanket are instantiated. Show that A is d-separated from the remaining uninstantiated variables.

Exercise 2.9. Apply the procedure using the ancestral graph given in Section 2.2.1 to determine whether A is d-separated from C given B in the network in Figure 2.19.

Exercise 2.10. Let D_1 and D_2 be DAGs over the same variables. The graph D_1 is an *I-submap* of D_2 if all d-separation properties of D_1 also hold for D_2. If D_2 is also an I-submap of D_1, they are said to be *I-equivalent*. Which of the four DAGs in Figure 2.21 are I-equivalent?

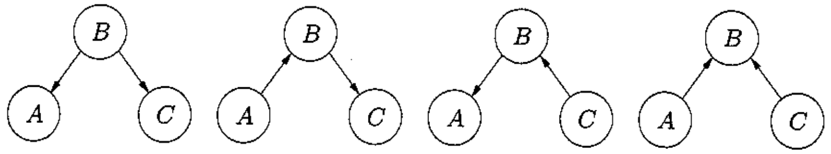


Fig. 2.21. Figure for Exercise 2.10.

Exercise 2.11. Let $\{A_1, \ldots, A_n\}$ be a topological ordering of the variables in a Bayesian network, and consider variable A_i with parents $\mathrm{pa}(A_i)$. Prove that A_i is d-separated from $\{A_1, \ldots, A_{i-1}\} \setminus \mathrm{pa}(A_i)$ given $pa(A_i)$.

Exercise 2.12. Consider the network in Figure 2.20. Which conditional probability tables must be specified to turn the graph into a Bayesian network?

Exercise 2.13. In Figure 2.22 the structure of a simple Bayesian network is shown. The accompanying conditional probability tables are shown in Tables 2.5 and 2.6, and the prior probabilities for A are 0.9 and 0.1. Are A and C d-separated given B? Are A and C conditionally independent given B?

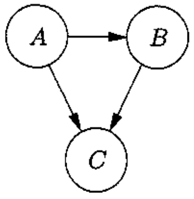


Fig. 2.22. A simple Bayesian network for Exercise 2.13.

	$A = a_1$	$A = a_2$
$B = b_1$	0.3	0.6
$B = b_2$	0.7	0.4

Table 2.5. $P(B \mid A)$.

	$A = a_1$	$A = a_2$
$B = b_1$	(0.1 ; 0.9)	(0.1 ; 0.9)
$B = b_2$	(0.2 ; 0.8)	(0.2 ; 0.8)

Table 2.6. $P(C \mid A, B)$.

Exercise 2.14. Consider the network in Figure 2.20. Using the chain rule, establish an expression for the joint distribution over the universe $\{A, B, C, D, E, F\}$. Use this expression to show that B and D are conditionally independent given A and C.

Exercise 2.15. Prove that the probability distribution $P(\mathcal{U})$ defined by the chain rule for Bayesian networks is indeed a probability distribution.

Exercise 2.16. Prove that the probability distribution $P(\mathcal{U})$ defined by the chain rule for a Bayesian network BN reflects the conditional probabilities specified in BN.

Exercise 2.17. Consider the Bayesian network from Exercise 2.13 and the finding $e = (0, 1)$ over A. What is $P(B, C, e)$?

Exercise 2.18. What steps would be taken if variable elimination were used to calculate the probability table $P(F \mid C = c_1)$ for the network in Figure 2.20? Assuming that each variable has ten states, what is the maximum size of a table during the procedure?

Exercise 2.19. Consider the DAG (a) in Exercise 2.10.

- Show that $P(B \mid A, C) = P(B \mid A)$.
- We have $P(A) = (0.1, 0.9)$ and the conditional probability tables in Table 2.7. Calculate $P(A, B, C)$.

	a_1	a_2
b_1	0.2	0.3
b_2	0.8	0.7

$P(B \mid A)$

	a_1	a_2
c_1	0.5	0.6
c_2	0.5	0.4

$P(C \mid A)$

Table 2.7. Conditional probability tables for Exercise 2.19.

Exercise 2.20. ^{E} Install an editor for Bayesian networks (a reference to a list of systems can be found in the preface).

Exercise 2.21. ^{E} Construct a Bayesian network for Exercise 1.12.

Exercise 2.22. ^{E} Construct a Bayesian network to follow the reasoning from Exercise 2.2. Use your own estimates of probabilities for the network.

Exercise 2.23. ^{E} Consider the Bayesian network in Figure 2.23 with conditional probabilities given in Table 2.8. Use your system to investigate whether A and C are independent.

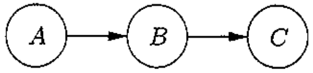


Fig. 2.23. Figure for Exercise 2.23.

	$A = yes$	$A = no$
b_1	0.6	0.2
b_2	0.1	0.5
b_3	0.2	0.1
b_4	0.1	0.2

$P(B \mid A)$

	b_1	b_2	b_3	b_4
$C = yes$	0.8	0.8	0.2	0.2
$C = no$	0.2	0.2	0.8	0.8

$P(C \mid B)$

Table 2.8. Tables for Exercise 2.23.

Exercise 2.24. ^{E} Use your system and Section 2.5 to perform the reasoning in Section 2.1.2.

3

Building Models

The framework of Bayesian networks is a very efficient language for building models of domains with inherent uncertainty. However, as can be seen from the calculations in Section 2.6, it is a tedious job to perform evidence transmission even for very simple Bayesian networks. Fortunately, software tools that can do the calculational job for us are available. In the rest of this book, we assume that the reader has access to such a system (some URLs are given in the preface). Therefore, we can start by concentrating on how to use Bayesian networks in model building and defer a presentation of methods for probability updating to Chapter 4.

In Section 3.1, we examine through examples the considerations you may go through when determining the structure of a Bayesian network model. Section 3.2 gives examples of estimation of conditional probabilities. The examples cover theoretically well-founded probabilities as well as probabilities taken from databases and purely subjective estimates. Section 3.3 introduces various modeling tricks to use when the quantity of numbers to acquire is overwhelming. Finally, Section 3.4 considers other types of queries that can be answered by Bayesian networks besides standard probability updating.

3.1 Catching the Structure

The first thing to have in mind when organizing a Bayesian network model is that its purpose is to give estimates of certainties for events that are not directly observable (or observable only at an unacceptable cost), and the primary task in model building is to identify these events. We call them *hypothesis events*. The hypothesis events detected are then grouped into sets of mutually exclusive and exhaustive events to form *hypothesis variable*.

The next thing to have in mind is that in order to come up with a certainty estimate, we should provide some information channels, and the task is to identify the types of achievable information that may reveal something about the hypothesis variables. These types of information are grouped into *information variables*, and a typical piece of information is a statement that a certain variable is in a particular state, but softer statements are also allowed.

Having identified the variables for the model, the next thing will be to establish the directed links for a causal network.

3.1.1 Milk Test

> Milk from a cow may be infected. To detect whether the milk is infected, you have a test, which may give either a *positive* or a *negative* test result. The test is not perfect. It may give a positive result on clean milk as well as a negative result on infected milk.

We have two hypothesis events: *milk infected* and *milk not infected*, and because they are mutually exclusive and exhaustive, they are grouped into the variable *Infected?* with the states *yes* and *no*. A possible information source is the test results, which can be either *positive* or *negative*. For this, we establish the variable *Test* with states *pos* and *neg*.

The causal direction between the two variables is from *Infected?* to *Test* (see Figure 3.1).

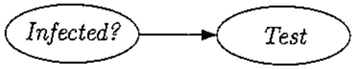


Fig. 3.1. The Bayesian network for the milk test.

Warning: Certainly, no sensible person will claim that a positive test result may infect the milk. However, our reasoning is often performed in the diagnostic direction, and in more complex situations you may therefore be tempted to wrongly direct the link from "symptom" to "disease."

From one day to another, the state of the milk can change. Cows with infected milk will heal over time, and a clean cow has a risk of having infected milk the next day. Now, imagine that the farmer performs the test each day. After a week, he has not only the current test result but also the six previous test results. For each day, we have a model like the one in Figure 3.1. These seven models should be connected such that past knowledge can be used for the current conclusion. A natural way would be to let the state of the milk yesterday have an impact on the state today. This yields the model in Figure 3.2.

The model in Figure 3.2 contains a set of hidden assumptions, which can be read from the d-separation properties.

First, the model assumes the *Markov property*: if we know the present, then the past has no influence on the future. In the language of d-separation, the assumption is that, for example, Inf_{i-1} is d-separated from Inf_{i+1} given Inf_i. If we know that the milk on day four is infected, then this can be used to forecast the probability that the milk will be infected on day five. This forecast will not be improved by knowing that the milk was not infected on day three. For various diseases, such an assumption will not be valid. Some diseases have a natural span of time. For example, if I have the flu today but was healthy yesterday, then I will most probably have the flu the day after tomorrow. On the other hand, if I have had the flu for four days, then there is a good chance that I will be cured the day after tomorrow. If the Markov property of Figure 3.2 does not reflect reality, the model should be changed. For example, it may be argued that you also need to go an extra day back, and the model will be as in Figure 3.3.

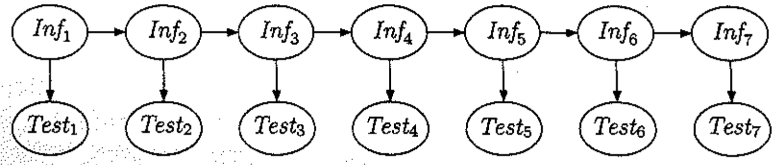


Fig. 3.2. A seven-day model for the milk test.

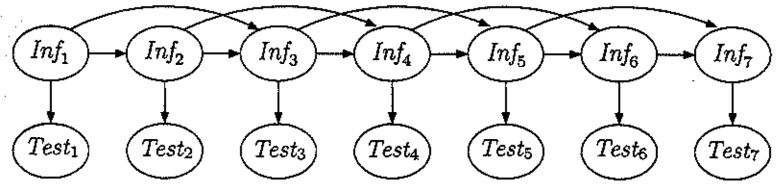


Fig. 3.3. A seven-day model with a two-day memory of infection.

Notice that although we in practice will never know the state of the infection nodes, it makes a difference whether the memory links are included. In the reasoning, we cannot exploit knowledge of the exact state of the previous infection node, but we may use a probability distribution based on a test result.

The second hidden assumption has to do with the test. Any two test nodes are d-separated given any infection node on the path. This means that the fault probability of the test is independent of whether it was previously correct. In other words, the fact that the test was wrong yesterday has no influence on whether the test will be correct today. If this does not reflect the behavior of the test, you may, for example, include its performance yesterday in the model. This is done in Figure 3.4.

A minor digression on modeling of tests: It is good to have as a rule that no test is perfect. Unless you explicitly know otherwise, a test should always

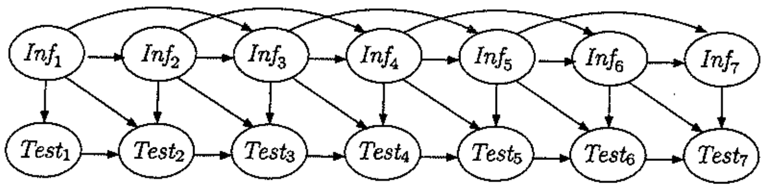


Fig. 3.4. A seven-day model with two-day memory for infection and a one-day memory of correctness of test.

be given a positive probability of false positives as well as false negatives. This is not all, though. You should also take the mechanism for false test results into account. Consider for example an HIV test with a probability of false positives of 10^{-5}, and assume that a person has received a positive test result. Now, you may have the option of repeating the test, but will this be of any help? It will depend on the mechanisms that cause the test to give a wrong result. If a test is positive because this particular person's blood is composed so that it will produce a positive test result regardless of a positive HIV infection, then a repeated test will not provide new information. If, on the other hand, the experiment is such that it now and then goes wrong, then a repeated test may be worthwhile and it will be advisable to repeat the test before the "verdict" is passed (in case the second test result is negative, a third test may be advisable). Models for these two types of failure mechanisms are shown in Figure 3.5.

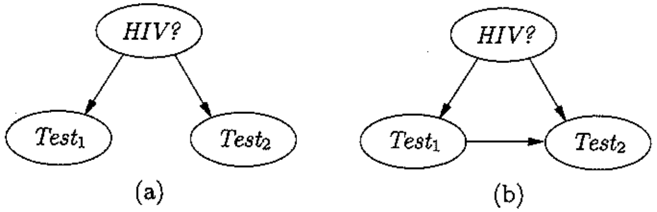


Fig. 3.5. Model (a) illustrates the scenario in which a repeated test may provide new information, and model (b) shows the situation in which repeating a test always produces the same result.

3.1.2 Cold or Angina?

I wake up in the morning with a sore throat. It may be the beginning of a cold or I may suffer from angina (inflammation of the throat). If it is severe angina, I will not go to work. To gain more insight, I can take my temperature, and I can look down my throat for yellow spots.

Here we have five hypothesis events *Cold?* {*no*, *yes*} and *Angina?* {*no*, *mild*, *severe*}. The hypothesis events must be organized into a set of variables with mutually exclusive and exhaustive states. We may use the variables indicated previously, but we may also use only one variable *Sick?* with states {*no*, *cold*, *mild angina*, *severe angina*}. In the latter case, suffering from both cold and angina is excluded as a possibility. We choose to use the two variables *Cold?* and *Angina?*.

The information variables are *Sore Throat?* {*no*, *yes*}, *See Spots?* {*no*, *yes*}, and *Fever?* {*no*, *low*, *high*}. The variable *Fever?* causes a problem because it really is continuous. In Section 3.3.8, we give methods on how to deal with continuous variables.

Now it is time to consider the causal structure between the variables. We need not worry about how information is transmitted through the network. The only thing to worry about is which variables have a direct causal impact on other variables.

In this example, we have that *Cold?* has a causal impact on *Sore Throat?* and *Fever?* while *Angina?* has an impact on all information variables. The model is given in Figure 3.6.

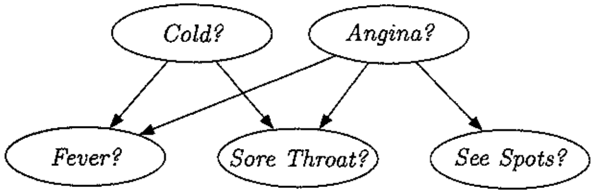


Fig. 3.6. A model for *Cold?* or *Angina?*.

The next thing to check is whether the conditional independences laid down in the model correspond to reality. For example, the model in Figure 3.6 yields that if we know the state of *Angina?*, then seeing spots will not have an impact on the expectation either for *Fever?* or for *Sore Throat?*. If we do not agree, we may introduce a link from *See Spots?* to, for example, *Fever?*. For now, we accept the conditional independences given by the model.

3.1.3 Insemination

Six weeks after insemination of a cow, you can perform two tests to determine whether the cow is pregnant: a blood test and a urine test.

Following the method from Section 3.1.1, we construct a model as in Figure 3.7. The variable *Pr* {*yes*,*no*} represents a possible pregnancy, and *BT* {*pos*,*neg*} and *UT* {*pos*,*neg*} represent the results of the blood test and the urine test, respectively.

Next, we will analyze the conditional independences stated by the model. We ask the expert whether it is correct that the outcomes of the two tests

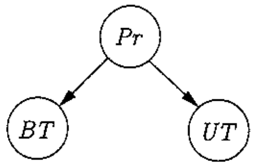


Fig. 3.7. A model for pregnancy.

are independent given *Pr*. More specifically, assume that we know the cow is pregnant. From this, we infer some expectations for the test results. Now, if we get a negative test result from the blood test, will this change our expectation for the urine test? The experts say that it will, and we must conclude that the model is not a proper reflection of reality.

There are several ways to change the model. You might, for example, introduce a link between the two test nodes, but there is no natural direction. To find out what to do, you must study the process more carefully, and it turns out that what the two tests actually do is to trace indications of hormonal changes in the cow. A more-refined model will involve a variable *Ho*, reflecting whether hormonal changes have taken place in the cow, and the model will be as in Figure 3.8.

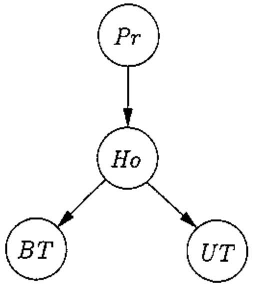


Fig. 3.8. A more correct model for pregnancy. Both the blood test (*BT*) and the urine test (*UT*) measure the hormonal state (*Ho*).

For the model in Figure 3.8, it does not hold that *BT* and *UT* are independent given *Pr*. The model states that *BT* and *UT* are independent given *Ho* (which should be checked). If the model in Figure 3.7 is used for diagnosing a possible pregnancy, a negative outcome of both the blood test and the urine test will be counted as two independent pieces of evidence and therefore overestimate the probability for the insemination to have failed (see Exercise 3.8).

In the model in Figure 3.8, we have introduced the variable *Ho*, which is neither a hypothesis variable nor an information variable. Such variables are called *mediating variables*. Mediating variables are often introduced when two

variables are not (conditionally) independent as opposed to the situation in the current model. Some standard situations are illustrated in Figure 3.9.

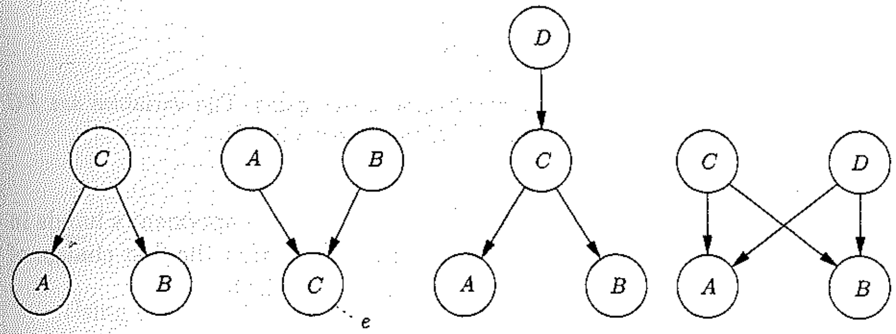


Fig. 3.9. Examples in which an intermediate variable *C* "resolves" undirected dependencies. In examples (*a*) and (*b*), *A* and *B* are not independent, whereas *A* and *B* are not independent given *D* in examples (*c*) and (*d*).

3.1.4 A Simplified Poker Game

> In this poker game, each player receives three cards and is allowed two rounds of changing cards. In the first round, you may discard any number of cards from your hand and get replacements from the pack of cards. In the second round, you may discard at most two cards. After the two rounds of card changing, I am interested in an estimate of my opponent's hand.

The hypothesis events are the various types of hands in the game. They may be classified in the following way (in increasing rank): nothing special, 1 ace, 2 of the same value, 2 aces, flush (3 of a suit), straight (3 of consecutive value), 3 of the same value, straight flush. Ambiguities are resolved according to rank. This is, of course, a simplification, but it is often necessary to do so in modeling. The hypothesis events are collected into one hypothesis variable *OH* (opponent's hand) with the preceding classes as states.

The only information to acquire is the number of cards the player discards in the two rounds. Therefore, the information variables are *FC* (first change) with states *0, 1, 2, 3* and *SC* (second change) with states *0, 1, 2*. By saying this, we are making an approximation again. The information on the cards you have seen is relevant for your opponent's hand. If, for example, you have seen three aces, then he cannot have two aces.

A causal structure for the information variables and the hypothesis variable could be as in Figure 3.10. However, this structure will leave us with no clue as to how to specify the probabilities.

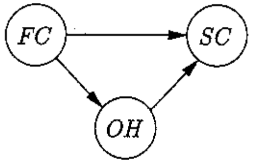


Fig. 3.10. An oversimplified structure for the poker game. The variables are *FC* (first change), *SC* (second change), and *OH* (opponent's hand).

What we need are mediating variables describing the opponent's hands in the process: the initial hand *OH0* and the hand *OH1* after the first change of cards. The causal structure will then be as in Figure 3.11.

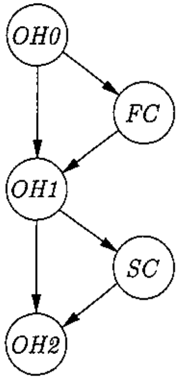


Fig. 3.11. A structure for the poker game. The two mediating variables *OH0* and *OH1* are introduced. *OH2* is the variable for my opponent's final hand.

To determine the states of *OH0* and *OH1*, we must produce a classification that is relevant for determining the states of the children (*FC* and *OH1*, say). We may let *OH0* and *OH1* have the states *nothing special, 1 ace, 2 of consecutive value, 2 of a suit, 2 of the same value, 2 of a suit and 2 of consecutive value, 2 of a suit and 2 of the same value, 2 of consecutive value and 2 of the same value, flush, straight, 3 of the same value, straight flush*.

We defer further discussion of the classification to the section on specifying the probabilities (Section 3.2).

3.1.5 Naive Bayes Models

In the previous sections we saw examples of Bayesian networks that were designed to capture the independence properties in the domains being modeled. However, the first Bayesian diagnostic systems were actually constructed

based on much simpler models, namely so-called *naive Bayes models*. In a naive Bayes model the information variables are assumed to be independent given the hypothesis variable (see Figure 3.12).

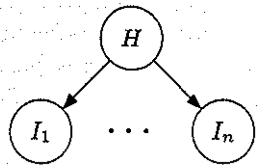


Fig. 3.12. A naive Bayes model.

Using this assumption, the conditional probability distribution for the hypothesis variable given the information variables is very easy to calculate, and the overall process (from model specification to probability updating) can be summarized as follows:

- Let the possible diseases be collected into one hypothesis variable H with prior probability $P(H)$.
- For all information variables I, acquire the conditional probability distribution $P(I \mid H)$ (the *likelihood* of H given I).
- For any set of observations $f_1, \ldots, f_n$ on the variables $I_1, \ldots, I_n$, calculate the product $P(f_1, \ldots, f_n \mid H) = P(f_1 \mid H) \cdot P(f_2 \mid H) \cdots P(f_n \mid H)$. This product is also called the likelihood for H given $f_1, \ldots, f_n$. The posterior probability for H is then calculated as

$$\begin{aligned} P(H \mid f_1, \ldots, f_n) &= \mu P(H) P(f_1, \ldots, f_n \mid H) \\ &= \mu P(H) \prod_{i=1}^{n} P(f_i \mid H), \end{aligned} \tag{3.1}$$

where $\mu = 1/P(f_1, \ldots, f_n)$ is a normalization constant.

What is particularly attractive with the calculation in equation (3.1) is that the time complexity is linear in the number of information variables, and that each term in the product involves only two numbers (assuming that the hypothesis variable is binary), one for $P(f_i \mid H = y)$ and one for $P(f_i \mid H = n)$. On the other hand, as we also saw from the insemination example, the independence assumption need not hold, and if the model is used anyway, the conclusions may be misleading. However, in certain application areas (such as diagnosis) the naive Bayes model has been shown to provide very good performance, even when the independence assumption is violated. This is partly due to the fact that for many diagnostic problems we are interested only in identifying the most probable disease. In other words, if the conditional independence assumption does not change which state has the highest probability, then the naive Bayes model can be used without affecting the performance of the system. We shall return to these models in Section 8.1.

3.1.6 Causality

In the examples presented in the previous section, there was no problem in establishing the links and their directions. However, you cannot expect this part of the modeling always to go smoothly.

First, causal relations are not always obvious – recall the debates on whether smoking causes lung cancer or whether a person's sex has an impact on his/her ability in the technical sciences. Furthermore, causality is not a well-understood concept. Is a causal relation a property of the real world, or rather, is it a concept in our minds helping us to organize our perception of the world? For now, we make only one point about this issue, namely that in some situations you may be able to infer information about causality based on actions that change the state of the world. For example, assume that you are confronted with two correlated variables A and B, but you cannot determine a direction. If you observe the state of A, you will change your belief of B and vice versa. A good test then is to imagine that some outside agent *fixes* the state of A. If this does not make you change your belief of B, then A is not a cause of B. On the other hand, if this imagined test indicates no causal arrow in any direction, then you should look for an event that has a causal impact on both A and B. If C is such a candidate, then check whether A and B become independent given C (see Figure 3.9). We shall briefly return to the issue of discovering causal relations in Section 7.1, where we discuss methods for learning Bayesian networks from data.

3.2 Determining the Conditional Probabilities

The numbers (conditional probabilities) that you need to specify for a Bayesian network are called the *parameters* of the network. The basis for the conditional probabilities can have an epistemological status ranging from well-founded theory over frequencies in a database to subjective estimates. We will give examples of each type.

3.2.1 Milk Test

For the milk test in Figure 3.1, we need $P(\mathit{Infected?})$ and $P(\mathit{Test} \mid \mathit{Infected?})$. The retailer of the test should provide $P(\mathit{Test} \mid \mathit{Infected?})$. Any producer of such kinds of tests is supposed to have performed a series of tests yielding the relevant numbers, namely the frequency of *false positives*, $P(\mathit{Test} = \mathit{pos} \mid \mathit{Infected?} = \mathit{no})$, and the frequency of *false negatives*, $P(\mathit{Test} = \mathit{neg} \mid \mathit{Infected?} = \mathit{yes})$. Let both numbers be 0.01.

The numbers provided by the retailer are not sufficient for the user of the test. In the case of a positive test result, the milk may still be clean, and to come up with a probability we need the prior probabilities $P(\mathit{Infected?})$.

An estimate of the prior probability would in this case be the daily frequency λ of infected milk for each cow at the particular farm. Estimating λ may be a bit tricky because the farmer may have no experience with actually testing the milk from each specific cow with a perfect test. Assume that this particular farm has 50 cows, and that the milk from all cows is poured into a container and transported to the dairy, which tests the milk with a very precise test. The farmer's experience is that on average the dairy reports his milk to be infected once a month.

Now we must make various assumptions. The first assumption could be that the daily λ is the same for all cows. The next assumption could be that outbreaks of infected milk for the cows in the farm are independent. This yields a coin-tossing model with $P(\mathit{Infected?} = \mathit{yes}) = \lambda$. The information we have is that if we toss fifty coins at the same time, the frequency of at least one of them coming up with $\mathit{Infected?} = \mathit{yes}$ is 1 out of 30. That is, in 29 days out of 30, none of the cows are infected and the probability that all the cows are clean on a given day is therefore 29/30. Moreover, from the assumption of the outbreaks being independent we also have that the probability of all 50 cows being clean on a given day is $(1-\lambda)^{50}$:

$$P(\mathit{Inf}_1, \ldots, \mathit{Inf}_{50}) = (1-\lambda_1)\cdots(1-\lambda_{50}) = (1-\lambda)^{50}.$$

Combining all this, we now have

$$(1-\lambda)^{50} = \frac{29}{30},$$

which yields the estimate

$$\lambda = 1 - \left(\frac{29}{30}\right)^{0.02} \approx 0.0007.$$

This completes the model, and next you can use a computer system to calculate posterior probabilities. The interesting question for this situation is, if we get a positive test result, what is the probability that the milk is infected? This is left as an exercise (see Exercise 3.5).

For the seven-day model in Figure 3.2, we also need $P(\mathit{Inf}_{i+1} \mid \mathit{Inf}_i)$. There are two numbers to estimate: the risk of becoming infected and the chance of being cured. These numbers must be based on experience. For the sake of the example, let the risk of becoming infected be 0.0002 and the chance of being cured 0.3. This gives the numbers in Table 3.1.

For the seven-day model with a two-day memory of infection (Figure 3.3), we need $P(\mathit{Inf}_{i+1} \mid \mathit{Inf}_i, \mathit{Inf}_{i-1})$. If we assume that the risk of being infected is the same as before, that the infection always lasts at least two days, and that after this the chance of being cured is 0.4 each of the following days, then the numbers are as in Table 3.2 (see Exercise 3.10).

For the seven-day model with two-day memory of infection as well as correctness of test (Figure 3.4), we furthermore need $P(\mathit{Test}_{i+1} \mid \mathit{Inf}_i, \mathit{Inf}_{i+1},$

		Inf_i	
		yes	*no*
Inf_{i+1}	*yes*	0.7	0.0002
	no	0.3	0.9998

Table 3.1. $P(Inf_{i+1} \mid Inf_i)$.

		Inf_{i-1}	
		yes	*no*
Inf_i	*yes*	0.6	1
	no	0.0002	0.0002

Table 3.2. $P(Inf_{i+1} = yes \mid Inf_i, Inf_{i-1})$.

$Test_i$). If we assume that a correct test has a 99.9% chance of being correct next time, and an incorrect test has a 90% risk of also being incorrect next time, we can calculate all required numbers for the four-dimensional table. However, by introducing mediating variables, Cor_i, the specification of numbers could be easier, and the tables would be smaller. Figure 3.13 shows how the model could be simplified.

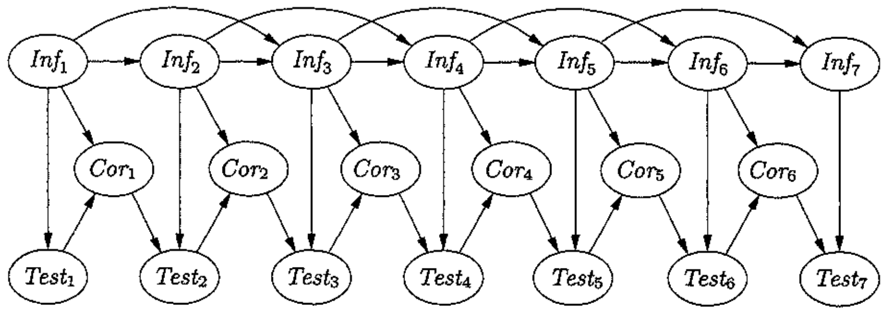


Fig. 3.13. A seven-day model with a two-day memory for infection and a one-day memory of correctness of test.

With the preceding assumptions, the required tables are as in Table 3.3.

3.2.2 Stud Farm

The stallion Brian has sired Dorothy on the mare Ann and sired Eric on the mare Cecily. Dorothy and Fred are the parents of Henry, and Eric has sired Irene on Gwenn. Ann is the mother of both Fred and Gwenn, but their fathers are in no way related. The colt John with

		Inf_i	
		yes	*no*
$Test_i$	*pos*	1	0
	neg	0	1

		Cor_{i-1}	
		yes	*no*
Inf_i	*yes*	0.999	0.1
	no	0.001	0.9

Table 3.3. The conditional probability distributions $P(Cor_i = yes \mid Inf_i, Test_i)$ and $P(Test_i = pos \mid Inf_i, Cor_{i-1})$.

> the parents Henry and Irene has been born recently; unfortunately, it turns out that John suffers from a life-threatening hereditary disease carried by a recessive gene. The disease is so serious that John is displaced instantly, and since the stud farm wants the gene out of production, Henry and Irene are taken out of breeding. What are the probabilities for the remaining horses to be carriers of the unwanted gene?

The genealogical structure for the horses is given in Figure 3.14.

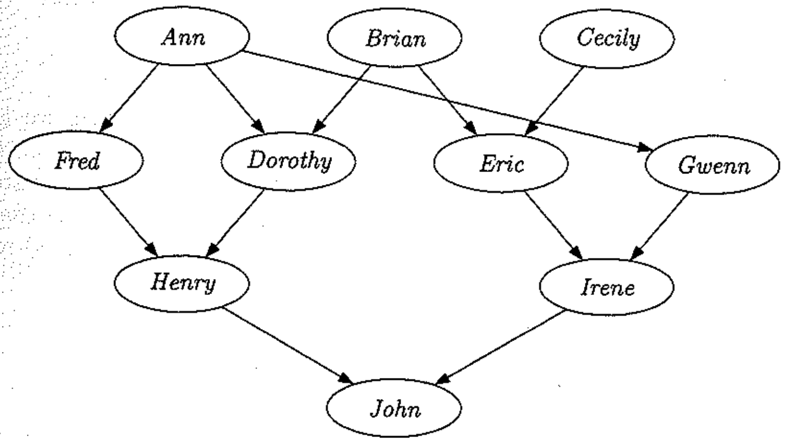


Fig. 3.14. Genealogical structure for the horses in the stud farm.

The only information variable is John. Before the information on John is acquired, he may have three genotypes: he may be sick (aa), a carrier (aA), or he may be pure (AA). The hypothesis events are the genotypes of all other horses in the stud farm.

The conditional probabilities for inheritance are both empirically and theoretically wellstudied, and the probabilities are as shown in Table 3.4.

The inheritance tables could be as in Table 3.4. However, for all horses except John, we have additional knowledge. Since they are in production, they cannot be of type aa. A way to incorporate this would be to build a

	aa	aA	AA
aa	(1, 0, 0)	(0.5, 0.5, 0)	(0, 1, 0)
aA	(0.5, 0.5, 0)	(0.25, 0.5, 0.25)	(0, 0.5, 0.5)
AA	(0, 1, 0)	(0, 0.5, 0.5)	(0, 0, 1)

Table 3.4. $P(\mathit{Child} \mid \mathit{Father}, \mathit{Mother})$ for genetic inheritance. The numbers (α, β, γ) are the child's probabilities for (aa, aA, AA).

Bayesian network in which all inheritance is modeled in the same way and afterward enter the findings that all horses but John are not aa. It is also possible to calculate the conditional probabilities directly. If we first consider inheritance from parents that may be only of genotype AA or aA, we get Table 3.5.

	aA	AA
aA	(0.25, 0.5, 0.25)	(0, 0.5, 0.5)
AA	(0, 0.5, 0.5)	(0, 0, 1)

Table 3.5. $P(\mathit{Child} \mid \mathit{Father}, \mathit{Mother})$ when the parents are not sick.

The table for John is as in Table 3.5. For the other horses, we know that aa is impossible. This is taken care of by removing the state aa from the distribution and normalizing the remaining distribution. For example, $P(\mathit{Child} \mid aA, aA) = (0.25, 0.5, 0.25)$, but since aa is impossible, we get the distribution $(0, 0.5, 0.25)$, which is normalized to $(0, 0.67, 0.33)$. The final result is shown in Table 3.6.

	aA	AA
aA	(0.67, 0.33)	(0.5, 0.5)
AA	(0.5, 0.5)	(0, 1)

Table 3.6. $P(\mathit{Child} \mid \mathit{Father}, \mathit{Mother})$ with aa removed.

In order to deal with Fred and Gwenn, we introduce the two unknown fathers I and K as mediating variables and assume that they are not sick. For the horses at the top of the network, we specify prior probabilities. This will be an estimate of the frequency of the unwanted gene, and there is no theoretical way to derive it. Let us assume that the frequency is such that the prior belief of a horse being a carrier is 0.01.

In Figure 3.15, the final model with initial probabilities is shown; Figure 3.16 gives the posterior probabilities given that John is aa; and in Figure 3.17 you can see the posterior probabilities with the prior beliefs at the top changed to 0.0001. Note that the sensitivity to the prior beliefs is very small for the horses whose posterior probability for *carrier* is much greater than 0, for instance in the cases of Ann and Brian.

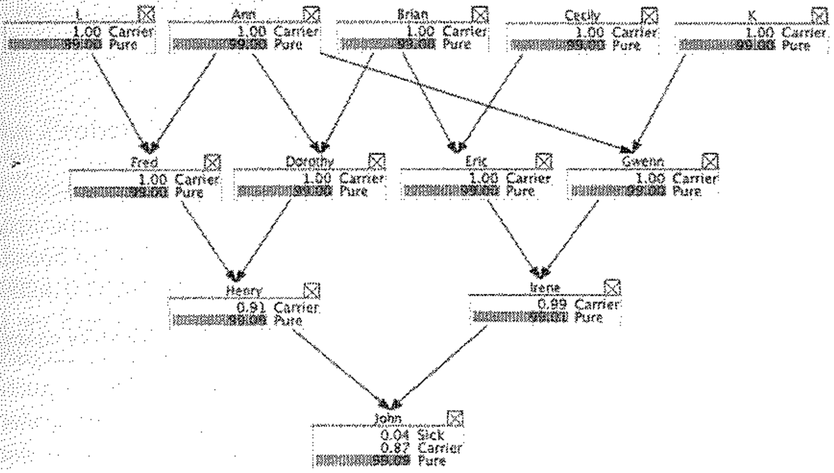


Fig. 3.15. The stud farm model with initial probabilities.

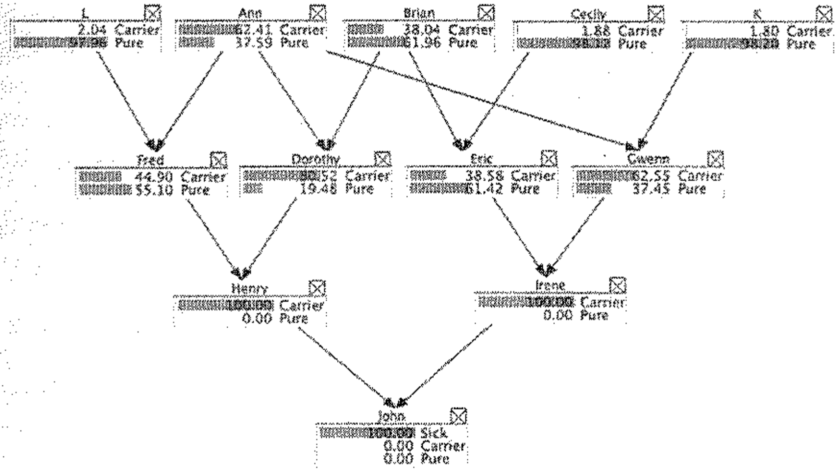


Fig. 3.16. Stud farm probabilities given that John is sick.

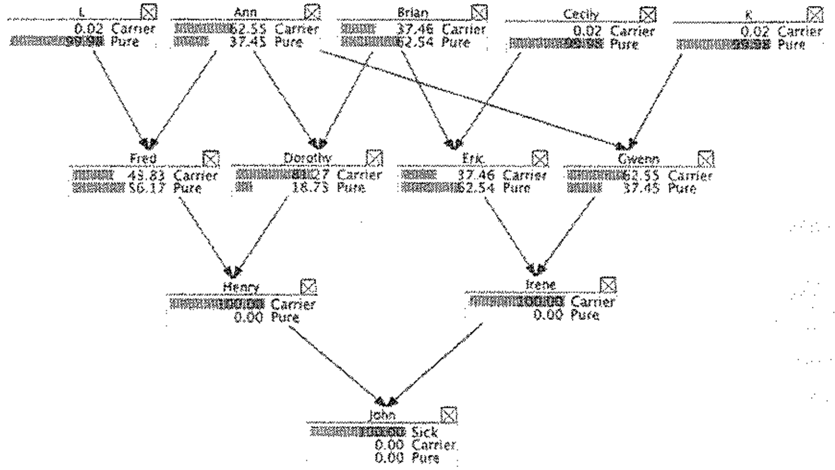


Fig. 3.17. Stud farm probabilities with prior probabilities for top variables changed to (0.0001, 0.9999).

3.2.3 Poker Game

In the stud farm example, the conditional probabilities were established mainly through theoretical considerations. This should also be attempted for the model of the poker game developed in Section 3.1.4, but it cannot be carried through entirely.

Consider for example $P(FC|\, OH0)$. It is not possible to give probabilities that are valid for any opponent. It is heavily dependent on the opponent's insight, psychology, and game strategy. We will assume the following strategy:

- If nothing special (*no*), then change 3.
- If 1 ace (*1 a*), then keep the ace.
- If 2 of consecutive value (*2 cons*), 2 of a suit (*2 s*), or 2 of the same value (*2 v*), then discard the third card.
- If 2 of a suit and 2 of consecutive value, then keep 2 of a suit (this strategy could be substituted by a random strategy for keeping either 2 of a suit or 2 of consecutive value).
- If 2 of a suit and 2 of the same value or 2 of consecutive value and 2 of the same value, then keep the 2 of the same value.
- If flush (*fl*), straight (*st*), 3 of the same value (*3 v*), or straight flush (*sfl*), then keep it.

Based on the preceding strategy, a logical link between *FC* and *OH0* is established. Note that the strategy makes the states for combined hands redundant. They play no role, and therefore we remove them.

The strategy for $P(SC|\, OH1)$ is the same except that in the case of *no*, only 2 cards are discarded.

These strategies seem to be the most rational. However, deterministic strategies in games do not always work, since they give your opponent valuable information about your hand. A good strategy should therefore be random rather than deterministic. Sometimes you may, for example, change nothing although you have a weak hand. Some people call it bluff, but it is really a way of increasing your opponent's uncertainty no matter what you do.

The remaining probabilities to specify are $P(OH0), P(OH1\,|\,OH0, FC)$, and $P(OH2\,|\,OH1, SC)$.

The Probability Distribution $P(OH0)$

The states are (*no, 1 a, 2 cons, 2 s, 2 v, fl, st, 3 v, sfl*), and through various (approximated) combinatorial calculations, the prior probability distribution is found to be $P(OH0) = (0.1569, 0.0765, 0.0635, 0.4447, 0.1694, 0.0494, 0.0353, 0.0024, 0.0024)$. For example, in order to determine the probability $P(OH0 = st)$ we first calculate the number of different ways in which we can obtain a straight: by disregarding permutations of the three cards, we get $52 \cdot 4 \cdot 4$ by letting *ka2* be a straight. However, since we do not want to include straight flushes, we subtract the number of ways (52) in which we can obtain a straight flush (again disregarding permutations), and finally we divide by the number of ways to draw three cards out of 52 cards (the latter is equal to the binomial coefficient $\binom{52}{3}$):

$$P(OH0 = st) = \frac{52 \cdot 4 \cdot 4 - 52}{\binom{52}{3}} = 0.0353.$$

The Probability Distribution $P(OH1\,|\,OH0, FC)$

Due to the logical links between *OH0* and *FC*, it is sufficient to consider only nine out of the possible 36 parent configurations, namely (*no, 3*), (*1 a, 2*), (*2 cons, 1*), (*2 s, 1*), (*2 v, 1*), (*fl, 0*), (*st, 0*), (*3 v, 0*), (*sfl, 0*). The last four are obvious. In Table 3.7, the results of the approximate combinatorial calculations are given.

The probabilities for the remaining parent configurations may be whatever is convenient, so put, for example, $P(OH1|\ 3\ v,\ 1) = (1, 0, \ldots, 0)$.

The Probability Distribution $P(OH2|\, OH1,\ SC)$

First, a table $P(OH2'\,|\,OH1, SC)$ similar (but not identical in the numbers) to Table 3.7 can be calculated. However, the states of $OH2'$ are not the ones we are interested in. We are interested in the *value* of the hand, and a state such as *2 cons* is of no value unless one of them is an ace. Therefore, the probabilities for the states of $OH2'$ are transformed to probabilities for *OH2*. For the transformation, the following rules are used:

		(OH0, FC)				
		(no, 3)	(1 a, 2)	(2 cons, 1)	(2 s, 1)	(2 v, 1)
	no	0.1583	0	0	0	0
	1 a	0.0534	0.1814	0	0	0
	2 cons	0.0635	0.0681	0.3470	0	0
	2 s	0.4659	0.4796	0.3674	0.6224	0
OH1	2 v	0.1694	0.1738	0.1224	0.1224	0.9592
	fl	0.0494	0.0536	0	0.2143	0
	st	0.0353	0.0383	0.1632	0.0307	0
	3 v	0.0024	0.0026	0	0	0.0408
	sfl	0.0024	0.0026	0	0.0102	0

Table 3.7. $P(OH1 \mid OH0,\ FC)$ for the nonobvious parent configurations.

$$1\ a = 1\ a + \frac{1}{6}(2\ cons + 2\ s),$$

$$no = no + \frac{5}{6}(2\ cons + 2\ s).$$

The probabilities of *2 a* are calculated specifically. The resulting probabilities are given in Table 3.8.

		(OH1, Sc)				
		(no, 2)	(1 a, 2)	(2 cons, 1)	(2 s, 1)	(2 v, 1)
	no	0.5613	0	0.5903	0.5121	0
	1 a	0.1570	0.2425	0.1181	0.1024	0
	2 v	0.1757	0.0667	0.1154	0.1154	0.8838
OH2	2 a	0.0055	0.1145	0.0096	0.0096	0.0736
	fl	0.0559	0.0559	0	0.2188	0
	st	0.0392	0.0392	0.1666	0.0313	0
	3 v	0.0027	0.0027	0	0	0.0426
	sfl	0.0027	0.0027	0	0.0104	0

Table 3.8. $P(OH2 \mid OH1,\ SC)$ for the nonobvious configurations.

Using a model such as the one in Figure 3.11 and with the conditional probability tables specified in this section, we have established a model for assisting a (novice) poker player. However, if my opponent knows that I use the system, he can change cards in such a way that affects my estimate of his hand.

3.2.4 Transmission of Symbol Strings

A language L over 2 symbols $(\mathbf{a}, \mathbf{b})$ is transmitted through a channel. Each word is surrounded by the delimiter symbol c. In the transmission some characters may be corrupted by noise and be confused with others.

A five-letter word is transmitted. Give a model that can determine the probabilities for the transmitted symbols given the received symbols.

There are five hypothesis variables $T_1, \ldots, T_5$ with states a, b and five information variables $R_1, \ldots, R_5$ with states a, b, c. There is a causal relation from T_i to R_i. Furthermore, there may also be a relation from T_i to $T_{i+1} (i = 1, \ldots, 4)$ encoding that certain pairs of symbols are more likely to occur than others. You could also consider more-involved relations from pairs of symbols to symbols, but for now we refrain from doing that. The structure is given in Figure 3.18.

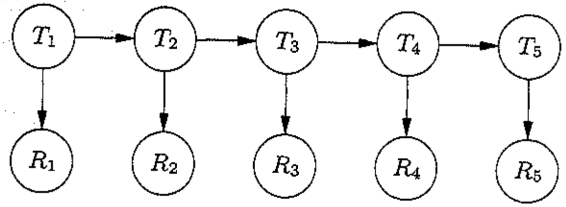


Fig. 3.18. A model for symbol transmission. T_i are the symbols transmitted; R_i are the symbols received.

The conditional probabilities can be established through experience. The probabilities $P(R_i \mid T_i)$ will be based on statistics describing the frequencies of confusion. Let Table 3.9 be the result.

	$T = a$	$T = b$
$R = a$	0.80	0.15
$R = b$	0.10	0.80
$R = c$	0.10	0.05

Table 3.9. $P(R \mid T)$ under transmission.

You may obtain the probabilities $P(T_{i+1} \mid T_i)$ by investigating the five-letter words in L. What is the frequency of the first letter? What is the frequency of the second letter given that the first letter is **a**? You continue to do this for each letter. You can refine this frequency analysis by also taking the frequencies of the words into consideration. Let Table 3.10 be the result of a frequency analysis.

You can calculate the required probabilities from Table 3.10 using the fundamental rule. The prior probabilities for T_1 are $(0.5, 0.5)$, and $P(T_2, T_1)$ is

First 2 letters	aaa	aab	aba	abb	baa	bab	bba	bbb
aa	0.017	0.021	0.019	0.019	0.045	0.068	0.045	0.068
ab	0.033	0.040	0.037	0.038	0.011	0.016	0.010	0.015
ba	0.011	0.014	0.010	0.010	0.031	0.046	0.031	0.045
bb	0.050	0.060	0.056	0.057	0.016	0.023	0.015	0.023

Table 3.10. Frequencies of five-letter words in L. The word **abaab**, for example, has frequency 0.040.

achieved by adding the elements in each row. Table 3.11 gives two conditional probabilities.

	a	b
a	0.6	0.4
b	0.4	0.6

$P(T_2 \mid T_1)$

	a	b
a	0.24	0.74
b	0.76	0.26

$P(T_3 \mid T_2)$

Table 3.11. Two conditional probabilities for five-letter words in L.

An alternative model would be to have a hypothesis variable, *Word*, with 32 states and with Table 3.10 as prior probabilities (see Figure 3.19).

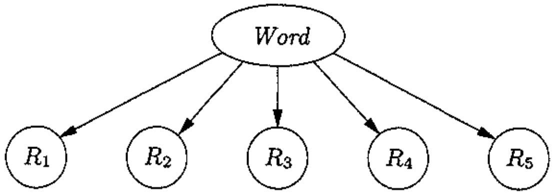


Fig. 3.19. An alternative model for symbol transmission. *Word* is the set of possible transmitted words.

This is manageable because of the small number of five-letter words over $\{\mathbf{a}, \mathbf{b}\}$; but if the alphabet had 24 symbols, and if six-letter words were considered, the number of states in *Word* would become intractably large. On the other hand, the model of Figure 3.18 may be too simple to catch the dependencies in Table 3.10, so the task really is to analyze the table in order to find the simplest structure describing it. There are methods for doing this, and we return to this topic in Chapter 7.

3.2.5 Cold or Angina?

The estimation of the conditional probabilities for the example introduced in Section 3.1.2 has a very subjective flavor based on my own experience with colds and anginas. I estimate the following probabilities: $P(Cold?)$, $P(Angina?)$, $P(See\ Spots? \mid Angina?)$, $P(Fever? \mid Cold?, Angina?)$, $P(Sore\ Th roat? \mid Cold?, Angina?)$.

Because in the morning I do not recall having been chilly yesterday, the prior probabilities $P(Cold?)$ and $P(Angina?)$ are my subjective recollections of how often I wake up in the morning with a cold or with an angina. Because cold is more frequent than angina, I put $P(Cold?) = (0.97, 0.03)$ and $P(Angina?) = (0.993, 0.005, 0.002)$; the order of the states are taken from Section 3.1.2.

Without angina or with mild angina, I will not see spots. With severe angina, I would expect to see spots, but I may not. I put $P(See\ Spots? \mid An gina? = severe) = (0.1, 0.9)$.

The Probability Distribution $P(Sore\ Throat? \mid Cold?, Angina?)$

If I suffer from neither a cold nor angina, I have a background probability of 0.05 of having a sore throat in the morning; this background probability covers everything other than cold and angina that may result in a sore throat. A cold as well as angina may give me a sore throat. If I only have a cold, the probability of a sore throat is 0.4. If I have mild angina, the probability of a sore throat is 0.7, and in the case of severe angina, I will certainly have a sore throat. What if I have both a cold and mild angina? I do not have sufficient experience to come up with a reliable estimate. Instead, I can use the two conditional probabilities from before: out of 100 mornings, I will wake up five mornings with a "background produced" sore throat. Out of the remaining 95 mornings, the cold yields a sore throat in 40% of them, that is, 38 mornings. Out of the remaining 57 mornings, mild angina will cause a sore throat in 70% of them: 39.9 mornings. In total, if I have both mild angina and a cold, I will have a sore throat in 82.9 mornings out of 100. The number 82.9 indicates an unjustified precision, and for psychological reasons we set the probability to 0.85. In Section 3.3.2 on "noisy-or," we give a systematic treatment of this method of estimating probabilities. The full table for $P(SoreThroat? \mid Cold?, Angina?)$ is given in Table 3.12. It is left as an exercise to complete the model.

	$Angina? = no$	$Angina? = mild$	$Angina? = severe$
$Cold? = no$	0.05	0.7	1
$Cold? = yes$	0.4	0.85	1

Table 3.12. $P(Sore\ Throat? = yes \mid Cold?, Angina?)$.

3.2.6 Why Causal Networks?

As mentioned previously, the structure of a Bayesian network need not reflect cause–effect relations. The only requirement is that the d-separation properties of the network hold for the domain modeled. There are, however, good reasons to strive for causal networks. The model in Figure 3.20 can be used to illustrate some of the points. We have a disease *Dis* and two tests, *Ts* and *Tt*.

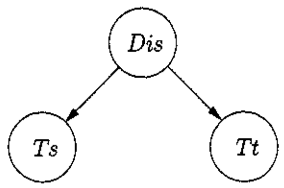


Fig. 3.20. A model for a disease with two tests.

When diagnosing, you usually reason opposite to the directions of the arrows in Figure 3.20, and trained physicians are usually inclined to provide conditional probabilities in the diagnostic direction. A model reflecting this might look like the one in Figure 3.21 a).

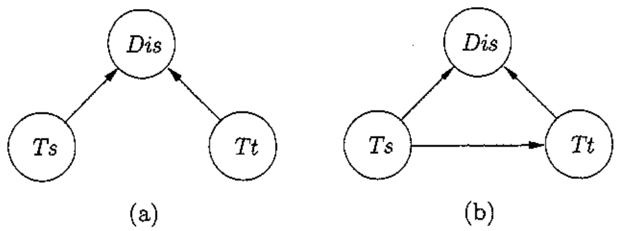


Fig. 3.21. Diagnostic models for the situation in Figure 3.20: (a) with a wrong independence, (b) with no (conditional) independence.

The model in Figure 3.21(a) is not correct. According to this model, *Ts* and *Tt* are independent (which is not the case in Figure 3.20), and there is no way to correct it by specifying the potentials in a sophisticated manner. To correct the model, you must add some extra structure making *Ts* and *Tt* dependent. You may, for example, introduce a link from *Ts* to *Tt*, as is done in Figure 3.21(b). Therefore, to get a correct model, it is not sufficient to acquire $P(Dis \mid Ts, Tt)$ together with the "priors" $P(Ts)$ and $P(Tt)$. This also illustrates another point, namely that a correct model of a causal domain is minimal with respect to links. In other words, if for some reason you wish to represent a causal relation with a link directed opposite to the causal direction, then the total number of links can not decrease, and most likely it will increase.

The model in Figure 3.20 has another advantage over the models in Figure 3.21, namely that the conditional probabilities $P(Ts \mid Dis)$ and $P(Tt \mid Dis)$ are more stable than the conditional probabilities specified for the models in Figure 3.21. The conditional probabilities for Figure 3.20 reflect general properties of the relation between diseases and tests, and they are the ones that a manufacturer of tests can publish, whereas the conditional probabilities for Figure 3.21 are a mixture of disease–test relations and prior frequencies of the disease.

It may happen that it is not possible to acquire the conditional probabilities for a correct model, but instead, other types of conditional probabilities are available. Assume, for example, that for the model in Figure 3.20, we can acquire only the potentials $P(Dis \mid Ts)$, $P(Dis \mid Tt)$, $P(Ts)$, and $P(Tt)$. Using Bayes' rule on $P(Dis \mid Ts)$ and $P(Ts)$, we get $P(Dis)$ and $P(Ts \mid Dis)$. The same can be done with $P(Dis \mid Tt)$ and $P(Tt)$. If the two calculations of $P(Dis)$ give the same result, we have the required potentials. If, on the other hand, the two calculations disagree, there is no safe way to solve the conflict. It can happen in many different situations that you have a set of potentials, but the model requires another set and there is no safe way of inferring the needed potentials. It is a lively area of research to construct engineering methods for getting the best out of what you have.

In Chapter 9, we deal with *interventions*. They provide another good reason for constructing causal models. An intervention is an action that has an impact on the state of certain variables. The impact of an intervention will spread in the causal direction, but not opposite to the causal direction. If the model does not reflect causal directions, it cannot be used to simulate the impact of interventions.

3.3 Modeling Methods

Much skepticism of Bayesian networks stems from the question of where the numbers come from. As shown in the previous section, they come from many different sources. If you are building a model over a domain in which experts actually *do* take decisions based on estimates, why should you not be able to make your Bayesian network estimate at least as well as the experts? You can, for example, use the technique described in Section 1.1 to acquire the probabilities from the experts. The acquisition of numbers is, of course, not without problems, and in this section we give some methods that can help you in this job. Also, we provide some modeling tricks.

3.3.1 Undirected Relations

It may happen that the model must contain dependence relations among variables A, B, C, say, but it is neither desirable nor possible to attach directions to them.[1] The relation may, for example, be a description of possible configurations. This difficulty may be overcome by using conditional dependence as described in Section 2.2.1 (converging influence).

Let $R(A, B, C)$ describe the relation using the values 0 and 1; $R(A, B, C) = 1$ for all valid configurations of A, B, and C. Add a new variable D with two states y and n and let A, B, and C be parents of D (see Figure 3.22). Assign D the deterministic conditional probability table given as $P(D = y \mid A, B, C) = R(A, B, C)$ (and $P(D = n \mid A, B, C) = 1 - R(A, B, C)$) and enter the evidence $D = y$. The variable D is called a *constraint variable*, and by entering $D = y$ we are basically forcing the relation/constraint to hold.

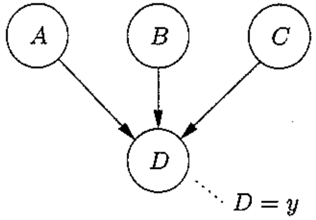


Fig. 3.22. A way to introduce undirected relations among A, B, and C.

Example 3.1. If we want to model that A, B, and C are always in the same state, then we can assign D the conditional probability table given in Table 3.13 (assuming that A, B, and C are binary).

		$C = y$		$C = n$	
		$B = y$	$B = n$	$B = y$	$B = n$
A	y	1	0	0	0
	n	0	0	0	1

Table 3.13. The conditional probability distribution $P(D = y \mid A, B, C)$ for the constraint variable D modeling that A, B, and C are always in the same state.

Example 3.2. I have washed two pairs of socks in the washing machine. The washing has been rather hard on them, so they are now difficult to distinguish. However, it is important for me to pair them correctly. To classify the socks, I have pattern and color. A classification model may be like the one in Figure 3.23. The variables S_i have states t_1 and t_2 for the two types, the variables P_i have two pattern types, and the variables C_i have two color types. The constraint that there are exactly two socks of each type is described in Table 3.14.

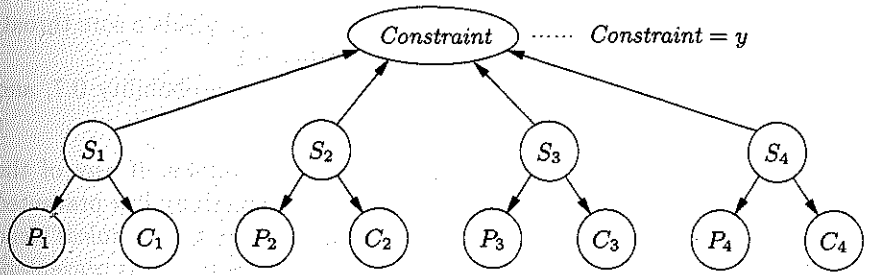


Fig. 3.23. A model for classifying pairs of socks.

S_1	t_1	t_1	t_1	t_1	t_1	t_1	t_1	t_1	t_2	t_2	t_2	t_2	t_2	t_2	t_2	t_2
S_2	t_1	t_1	t_1	t_1	t_2	t_2	t_2	t_2	t_1	t_1	t_1	t_1	t_2	t_2	t_2	t_2
S_3	t_1	t_1	t_2	t_2	t_1	t_1	t_2	t_2	t_1	t_1	t_2	t_2	t_1	t_1	t_2	t_2
S_4	t_1	t_2	t_1	t_2	t_1	t_2	t_1	t_2	t_1	t_2	t_1	t_2	t_1	t_2	t_1	t_2
P	0	0	0	1	0	1	1	0	0	1	1	0	1	0	0	0

Table 3.14. The table for $P = P(\text{Constraint} = y \mid S_1, S_2, S_3, S_4)$; t_1 and t_2 are the two states of S_1, S_2, S_3, S_4.

The situation is more subtle if the relation $R(A, B, C)$ is of probabilistic nature. If A, B, and C have no parents, $R(A, B, C)$ can be a joint probability table. On the other hand, if A has a parent, then $R(A, B, C)$ may be considered as representing a feedback cycle. We shall not deal with this problem but refer the reader to the literature on chain graphs.

3.3.2 Noisy-Or

When a variable A has several parents, you must specify $P(A \mid \mathbf{c})$ for each configuration $\mathbf{c}$ of the parents. If you take the distributions from a database, the number of cases for each configuration may become too small. Also, the configurations may be too specific for any expert. You may also be in the situation that you have reasonable estimates of $P(A \mid B)$ and $P(A \mid C)$, but you require $P(A \mid B, C)$. Then, you should look for assumptions that reduce the number of distributions to specify.

Consider in Section 3.2.5 the conditional probability table for $P(\mathit{Sore\ Throat?} \mid \mathit{Cold?}, \mathit{Angina?})$. It was possible to get estimates of $P(\mathit{Sore\ Throat?} \mid$

[1] In that case, the model is called a *chain graph*. A chain graph is an acyclic graph with both directed and nondirected links, where *acyclic* means that all cycles consist of only nondirected links.

Cold?) and $P(Sore\ \ Throat? \,|\, Angina?)$, but is there a general way to describe how they then combine into $P(Sore\ Throat? \,|\, Cold?,\ Angina?)$? The following is a way of describing it.

There are three events causing me to have a sore throat in the morning:

- the "background event," which in 5% of the mornings yields a sore throat;
- *cold*, which causes a sore throat with probability 0.4;
- *angina*, which when *mild* causes a sore throat with probability 0.7, and when it is *severe* it certainly causes a sore throat.

The preceding uncertainty can be interpreted as follows. If any of the causes are present, then I have a sore throat unless something has prevented it. In other words, if I have *mild* angina, then I have a sore throat unless some other circumstances prevent it, and there is a 30% chance that it is prevented. In the same way, there is a 60% chance that some inhibitor prevents me from having a sore throat although I have a cold, and the background event is prevented with probability 0.95.

Now, if we assume that the preventing factors are independent, then the combined probabilities are easy to calculate as one minus the product of the appropriate probabilities for the inhibitors (note that the background event is always a fact). The probabilities are given in Table 3.15.

	Angina? = *no*	*Angina?* = *mild*	*Angina?* = *severe*
Cold? = *no*	0.05	$1 - 0.95 \cdot 0.3$	1
Cold? = *yes*	$1 - 0.95 \cdot 0.6$	$1 - 0.95 \cdot 0.3 \cdot 0.6$	1

Table 3.15. Calculation of $P(Sore\ Throat? = yes \,|\, Cold?,\ Angina?)$. Note that some numbers are slightly different from the corresponding numbers in Table 3.12.

Another way to view the calculations above is to make the independence assumptions explicit in the model. Consider the model shown in Figure 3.24(a) and introduce an intermediate node ST_C between *Sore Throat* (ST) and *Cold?* (C) as well as an intermediate node ST_A between *Sore Throat?* and *Angina?* (A). The node ST_C captures the effect that *Cold?* has on *Sore Throat?* (i.e., it represents a "cold-induced" sore throat), whereas ST_A represent an "angina-induced" sore throat. In order to model the "background event" we introduce two additional nodes B and ST_B, where B represent the "background event," and ST_B plays the same role as ST_C and ST_A above. The three nodes ST_A, ST_B, and ST_C also represent the inhibitors, and they are assigned the conditional probability tables shown in Table 3.16; the numbers have been deduced from the itemized list above. Finally, since we will have a sore throat no matter whether it is induced by cold, angina, or something else, we assign ST a conditional probability distribution that corresponds to a logical-or. The resulting model is shown in Figure 3.24(b), where the variables ST_A, ST_B, and ST_C are independent, reflecting the assumption that the inhibitors are independent. Moreover, if we marginalize out the variables ST_A, ST_B, and ST_C, we end up with the conditional probability table in Table 3.15 (see also Exercise 3.20).

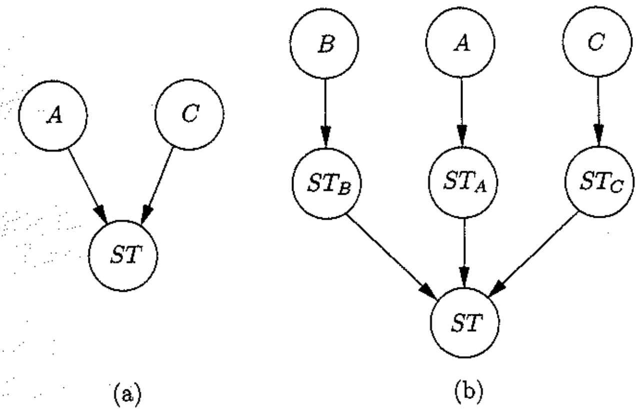


Fig. 3.24. Figure (a) shows the model structure for $P(ST \,|\, C, A)$, and figure (b) shows the model structure that explicitly represent the independence assumption about the inhibitors.

		A		
		no	*mild*	*severe*
ST_A	*yes*	0	$1 - 0.3$	1
	no	1	0.3	0

$P(ST_A \,|\, A)$

		B
		yes
ST_B	*yes*	$1 - 0.95$
	no	0.95

$P(ST_B \,|\, B)$

		C	
		no	*yes*
ST_C	*yes*	0	$1 - 0.6$
	no	1	0.6

$P(ST_C \,|\, C)$

Table 3.16. The conditional probability tables $P(ST_A \,|\, A)$, $P(ST_B \,|\, B)$, and $P(ST_C \,|\, C)$.

The preceding construction is an example of the simplifying assumption called a *noisy-or*. In what follows we put this assumption into a more general context, albeit only with binary variables.

Let $A_1, \ldots, A_n$ be binary variables listing all the causes of the binary variable B. Each event $A_i = y$ causes $B = y$ unless an *inhibitor* prevents it, and the probability for that is q_i (see Figure 3.25).

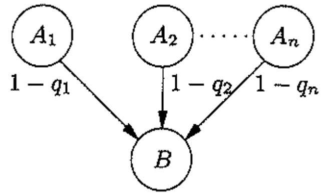


Fig. 3.25. The general situation for noisy-or. Here q_i is the probability that the impact of A_i is inhibited.

In other words, $P(B = n \mid A_i = y) = q_i$. We assume that *all inhibitors are independent.* Then $P(B = n \mid A_1, A_2, \ldots, A_n) = \prod_{j \in Y} q_j$, where Y is the set of indices for variables in the state y. For example,

$$\begin{aligned} P(B = y \mid A_1 = y, A_2 = y, A_3 = \cdots = A_n = n) \\ &= 1 - P(B = n \mid A_1 = y, A_2 = y, A_3 = \cdots = A_n = n) \\ &= 1 - q_1 \cdot q_2. \end{aligned}$$

By assuming "noisy-or," the number of probabilities to estimate grows linearly with the number of parents.

Note 1. We require $P(B = y \mid A_1 = \cdots = A_n = n)$ to be 0. This may seem to restrict the applicability of the approach. However, as in the preceding example, if $P(B = y) > 0$ when none of the causal events in the model are on, then introduce a background event that is always on.

Note 2. The complementary construction to noisy-or is called *noisy-and.* A set of causes should all be "on" in order to have an effect. However, the causes have random inhibitors, which are mutually independent.

Note 3. As in Figure 3.24(b), noisy-or can be modeled directly without performing the calculations (see Figure 3.26). This highlights the assumptions behind the noisy-or gate. If a cause is on, then its effect may be prevented by an inhibitor, and the probabilities for the inhibitors to be present are independent.

Note 4. The noisy-or model has been generalized to variables having more than two states, and in this form it is called a *noisy-max*; in this model we assume that the states of B are ordered.

3.3.3 Divorcing

Let $A_1, \ldots, A_n$ be a list of variables all of which are causes of B. If you wish to specify $P(B \mid A_1, \ldots, A_n)$, you might have a very large knowledge acquisition task ahead of you. Either you need to ask the experts on the distribution of B given very specific parent configurations or, if the table must be extracted from a database, you need a very large set of cases. The following example illustrates the problem.

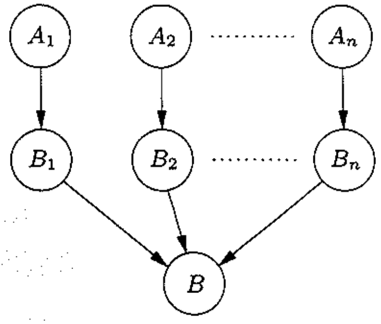


Fig. 3.26. Direct modeling of a noisy-or gate. Here $P(B_i \mid A_i)$ is the original $P(B \mid A_i)$, and $P(B \mid B_1, \ldots, B_n)$ is logical or.

Example 3.3 (Granting a loan). A bank will decide on a mortgage loan for a customer who wishes to purchase a house. The customer is asked to fill in a form giving information on various financial and personal matters together with various key information on the house. The answers are used to estimate the probability that the bank will get its money back.

The information can be the following: type of job, yearly income, other financial commitments, number and types of cars in the family, number of previous addresses during the last five years, number of children in the family, number of divorces, size and age of the house, price of the house, and type of environment.

In principle, each slot in the form represents a variable with a causal impact on the variable *Money back?*. If we assume that each parent variable has five states, we have already listed a parent space with $5^{11} \approx 5{,}000{,}000$ configurations. For each configuration, we request a distribution for A. No person can estimate that number of distributions, nor can he or she estimate a distribution for a divorced businesswoman with a yearly income of \$50,000, having loans of \$70,000 already, one car, three previous addresses, two children, wanting to purchase a twenty-year-old house of 150 m^2 at the price of \$200,000 in a farming area. Also, if the distributions are to be taken from a database, the bank will need at least 50,000,000 cases that may not be more than 10 years old.

To handle this kind of task, we *divorce* the parents. The set of parents $A_1, \ldots, A_i$ for B is divorced from the parents $A_{i+1}, \ldots, A_n$ by introducing a mediating variable C, making C a child of $A_1, \ldots, A_i$ and a parent of B (see Figure 3.27).

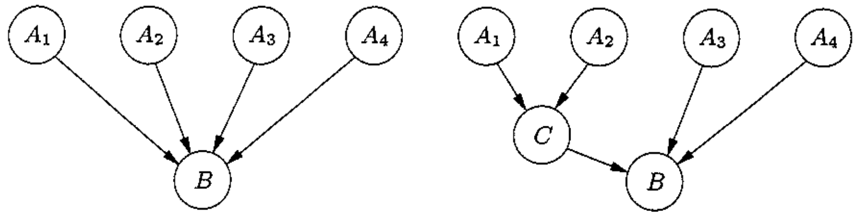


Fig. 3.27. Parents A_1 and A_2 are divorced from A_3 and A_4 by introducing the variable C.

The assumption behind divorcing is the following (with reference to Figure 3.27).

The set of configurations (A_1, A_2) can be partitioned into the sets $c_1, \dots, c_m$ such that whenever two configurations (a_1, a_2) and (a_1', a_2') are elements in the same c_i, then $P(B \mid a_1', a_2', A_3, A_4) = P(B \mid a_1, a_2, A_3, A_4)$. The divorcing variable then has $c_1, \dots, c_m$ as states.

In the example of granting a loan, it is impossible to perform an analysis as before, and you will group the variables based on another type of insight into the domain. For example, the variables about the house can be grouped and given a common child variable describing how safe the mortgage will be, the financial variables may be grouped for a variable describing the applicant's financial abilities; and the remaining variables may describe the applicant's stability.

In connection to the example of granting a loan, it should be noted that if we only want to perform a classification, then we need not build a Bayesian network. Other techniques such as statistical classifiers and classification trees (see Section 8.4) may be more adequate. However, if we also wish to calculate decision recommendations, we will need the posterior probabilities provided by a Bayesian network. We will deal further with this in Chapter 9.

3.3.4 Noisy Functional Dependence

There are ways of directing the divorcing. "Noisy-or" and "noisy-and" are examples of a general method called *noisy functional dependence.*

Example 3.4 (Headache). Headache (*Ha*) may be caused by fever (*Fe*), hangover (*Ho*), fibrositis (*Fb*), brain tumor (*Bt*), and other causes (*Ot*), and you may choose to soothe it with aspirin (*As*) (we ignore the effect aspirin has on fever). Let *Ha* have the states *no*, *mild*, *moderate*, *severe*. The various causes support each other in the effect. If, for example, $Ho = y$ or $Fb = y$ is present, then it may yield a *mild Ha*, but if both are present, then the *Ha* would be *moderate*. Furthermore, if also $As = y$, then *Ha* may drop to *no* or *mild*. Although the various parents of *Ha* combine in a rather involved manner, we still have the feeling that the impacts of the causes are independent. This kind of independence can be described as follows: if the headache is at level l, and we add an extra cause for headache, then the result is a headache at level q independent of how the initial state has been caused.

Assume that we can estimate conditional probabilities of type $P(Ha \mid C)$, and we want to combine the effects of the various causes. For this, we can imagine that we attach a number to the states of *Ha*: *no* $\mapsto 0$, *mild* $\mapsto 1$, *moderate* $\mapsto 2$, *severe* $\mapsto 4$, and the "adding up" of the effects consists in adding the numbers. A model could be similar to the one in Figure 3.28.

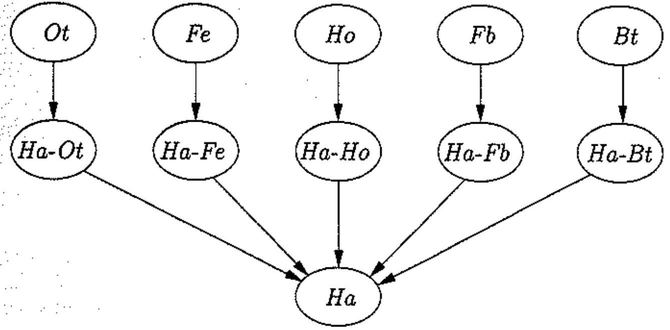


Fig. 3.28. A model for causes of headache. The bottom node adds up the effects.

The hidden assumption behind this method of adding up is that the effect from any cause is independent of the current state of headache, and it is faithfully reflected in the numbers attached to the headache states. To make it explicit in the model, we can give each headache node a child with numbers as states, these nodes are given a common child that adds the numbers, and a new node translates the numbers to *Ha* states (see Figure 3.29).

Now, for $P(\textit{Nu-Ha} \mid \textit{Nu-Ot}, \textit{Nu-Fe}, \textit{Nu-Ho}, \textit{Nu-Fb}, \textit{Nu-Bt})$ we can perform divorcing, we can add one number at a time (see Figures 3.30 and 3.31), or we can represent the function in any other kind of compact way.

The effect of aspirin can be included in two different ways. Either it subtracts a number from the sum or it has a direct effect on the headache state.

3.3.5 Expert Disagreements

It may happen that we are in a situation in which the experts disagree on the conditional probabilities for a model. Consider the model in Figure 3.32, and assume that we have three experts who agree on $P(B)$ and $P(C \mid A)$, but they disagree on $P(A)$ and $P(D \mid B, C)$. For the three experts, we have $P(A = y) = (0.1, 0.3, 0.4)$, and the table for $P(D \mid B, C)$ can be seen in Table 3.17.

If you have equal confidence in the three experts, you can take the mean of the three numbers. If your confidence in the experts varies, you may incorporate this and calculate a weighted average. For example, you may give the first

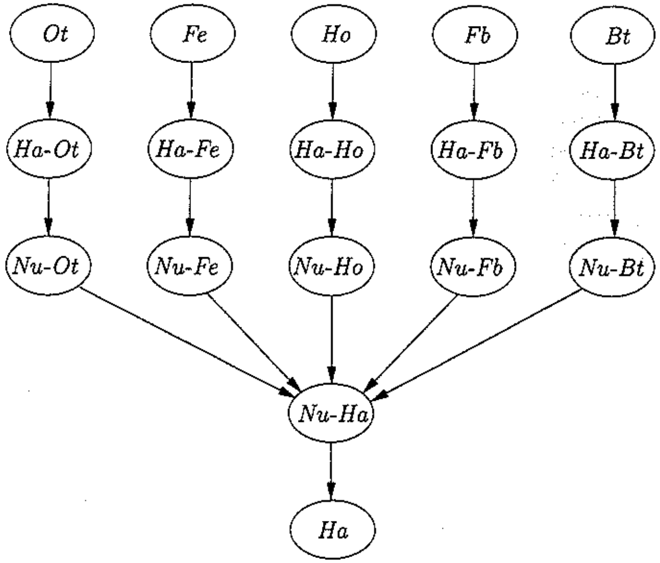


Fig. 3.29. A model that adds the headache states by transforming to numbers, adding, and transforming back to headache states again.

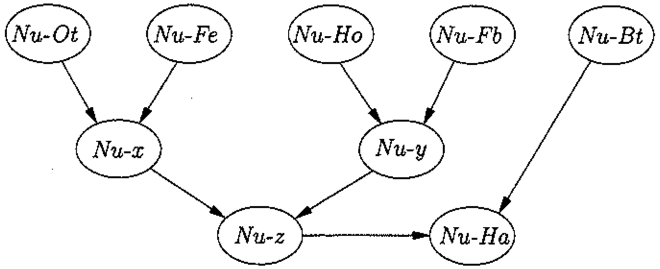


Fig. 3.30. The adder represented through divorcing.

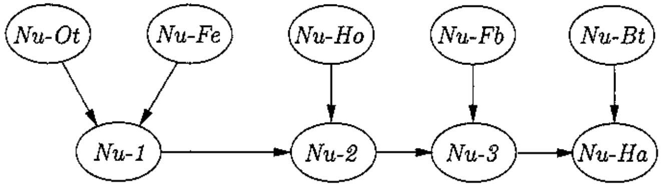


Fig. 3.31. The adder represented through adding one number at a time.

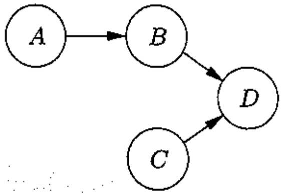


Fig. 3.32. A model with expert disagreements. All variables are binary.

		B	
		y	n
C	y	(0.4, 0.4, 0.6)	(0.7, 0.9, 0.7)
	n	(0.6, 0.4, 0.5)	(0.9, 0.7, 0.9)

Table 3.17. $P(D = y \mid B, C)$ for the three different experts s_1, s_2, s_3.

two experts a confidence weight 1 and the third expert a confidence weight 2. Because the total confidence weight is 4, you get a confidence distribution (0.25, 0.25, 0.5), and for A you have $P(A = y) = 0.25 \cdot 0.1 + 0.25 \cdot 0.3 + 0.5 \cdot 0.4 = 0.3$. The probability $P(D \mid B, C)$ is shown in Table 3.18.

		B	
		y	n
C	y	0.5	0.75
	n	0.5	0.85

Table 3.18. $P(D = y \mid B, C)$ weighted with confidence distribution (0.25, 0.25, 0.5).

The experts can be represented explicitly in the model by introducing a variable S with states s_1, s_2, and s_3. The variable S has a link to the nodes, about whose tables the three experts disagree (see Figure 3.33).

The variable S is given the confidence distribution (0.25, 0.25, 0.5) as before, and the child variables have a conditional probability table for each expert. The table $P(D = y \mid B, C, S)$ is as in Table 3.17.

By modeling the different expert opinions explicitly, you have prepared the model for *adaptation*. Whenever you have a case with evidence e entered into the model, you will get $P(S \mid e)$, which is an updated indication of which expert to believe. That is, you get a new confidence distribution that can be used for the next case, see also Section 6.3.

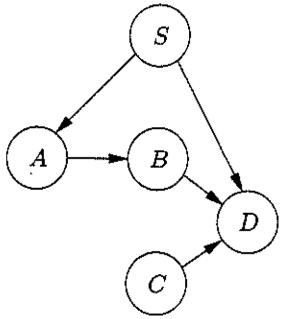


Fig. 3.33. The model from Figure 3.32 with the experts represented explicitly by the node S.

3.3.6 Object-Oriented Bayesian Networks

Complex Bayesian network models often include copies of almost-identical network fragments. Consider, for example, the Bayesian network shown in Figure 3.34, and assume that X_1 and X_2 have the same state space ($\text{sp}(X_1) = \text{sp}(X_2)$), and that the conditional probability tables associated with the nodes labeled A are identical; similarly for the nodes labeled B, C, D, and E. Given these two assumptions we see that the network contains four identical copies of the same *network fragment* defined by the five nodes A, B, C, D, E.

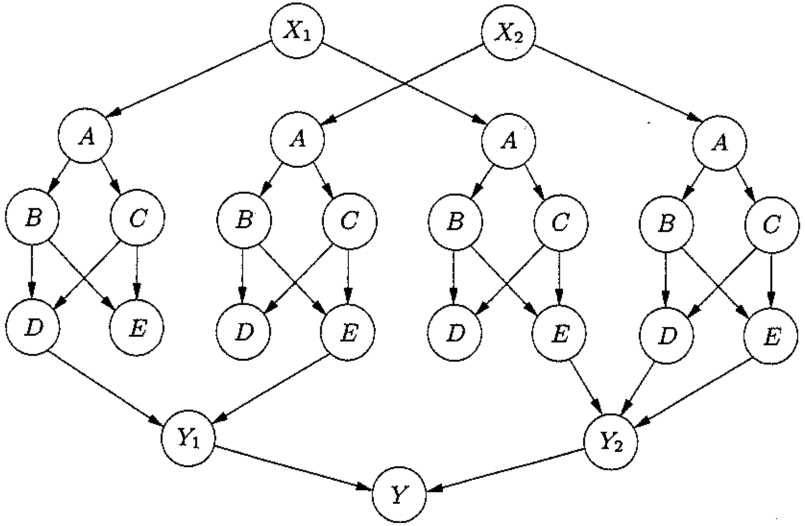


Fig. 3.34. A Bayesian network containing repetitive substructures.

The occurrence of such repetitive structures can be exploited during model construction. For example, instead of explicitly specifying the same network fragment multiple times, we could instead construct a generic network fragment that can be *instantiated* the required number of times. By borrowing terminology from the object-oriented programming paradigm, we call such a generic network fragment a *class*, and each network fragment that is produced by instantiating the class is called an *object*. Figure 3.35 shows a class description (called Class-name) for the duplicated network fragment in Figure 3.34. In order for the class to support the specification of the conditional probability distribution for A, the class includes an artificial node X (drawn as a dashed node) having the same state space as X_1 and X_2. Note that this node does *not* correspond to an actual variable, but should rather be seen as a "placeholder" that simply allows us to specify the probability distribution for A. The shaded nodes in Figure 3.35 indicate the part of the class/object that is accessible outside the object; they may be parents of nodes outside the object. Nodes that are neither dashed nor shaded are encapsulated within the object, and they may therefore be considered invisible to the rest of the model.

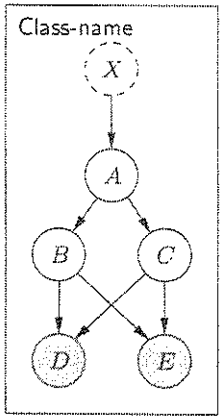


Fig. 3.35. A class model for the duplicated network fragment in Figure 3.34. Class-name is the name of the class.

Given such a class description, we can make an equivalent representation of the model in Figure 3.34 by instantiating the class four times and connecting X_1, X_2, Y_1, and Y_2 to the objects (labeled Inst. 1, Inst. 2, Inst. 3, Inst. 4) as appropriate. The resulting model is shown in Figure 3.36 and is called an *object-oriented Bayesian network model* (OOBN). The dashed arcs indicate which node X is a placeholder for in the various objects.

As implied by the discussion above, an object (or a class) can be seen as a function that given a certain input provides a probability distribution over a

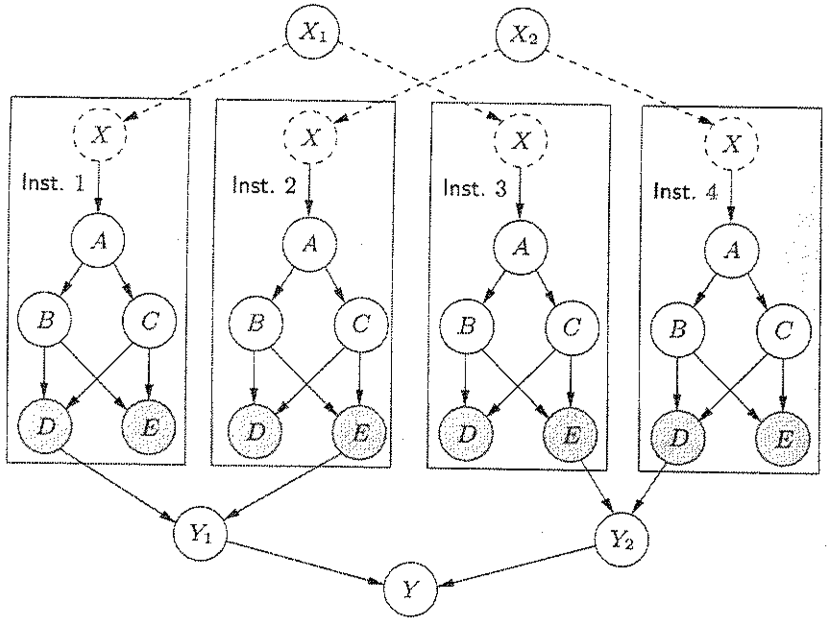


Fig. 3.36. An object-oriented Bayesian network representation of Figure 3.34.

set of variables. For example, the class shown in Figure 3.35 specifies a probability distribution over D and E given a state for X. Based on this perspective, we can partition the elements in an object into three sets: *input attributes*, *output attributes*, and *encapsulated attributes*. In the example above, X is an input attribute, D and E are output attributes, and A, B, and C are encapsulated attributes. Following standard programming terminology, the input attributes in the class description can be seen as the *formal parameters* of the corresponding function, whereas the *actual parameters* passed to an object are identified as the parents of the input attributes in the surrounding model. Thus, X can be considered a formal parameter, and X_1 is the actual parameter passed to the left-most object in Figure 3.36. In general, we also allow encapsulated attributes and output attributes to be objects themselves. However, input attributes must correspond to variables, since they serve as the parameters passed to the object. Note that the simplest type of class/object consists of a single variable, where the input attributes correspond to the parents of that variable.

The specification of encapsulated attributes is closely related to the concept of *information hiding* in the object-oriented programming paradigm. By taking this idea one step further, we obtain a straightforward mechanism for simplifying the visual representation of a model by abstracting away irrelevant details. For example, by abstracting away the encapsulated attributes in Figure 3.36 we obtain the OOBN shown in Figure 3.37. In general, when objects are encapsulated within other objects this approach provides us a method for obtaining a hierarchical representation of the model; each level corresponds to a particular level of abstraction revealing the encapsulated attributes for the current layer of objects.

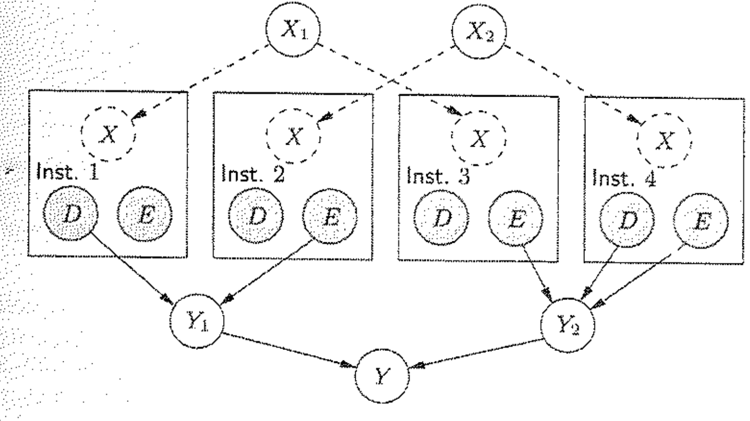


Fig. 3.37. An object-oriented Bayesian network model corresponding to the model shown in Figure 3.36. The encapsulated attributes have been hidden to simplify the representation.

Top-Down Construction of OOBNs

The input attributes and the output attributes are also referred to as the *interface* of the object, since instantiating these nodes will d-separate the internal part of the object (the encapsulated attributes) from the rest of the network (the proof is left as an exercise. This property supports a top-down model construction process: you may start constructing the model at a high level of abstraction by including only the interfaces of the objects without specifying their internal details. Later you can change the abstraction level and start specifying/refining the internal class description.

For example, assume that you should construct a Bayesian network model for the safety characteristics of a car. We know that the type of car and its maintenance level influence both the general steering characteristics of the car as well as its braking capabilities. In turn, these two aspects influence the steering safety and the braking power of the car.

We also know that the steering safety and the braking power are influenced by the grip of the car, and the grip is mainly determined by the tire type and the tire mileage. However, it may happen that at the time of model specification we do not know (or do not want to specify) the relationship between the grip of the car and tire type and mileage. See Figure 3.38 for a

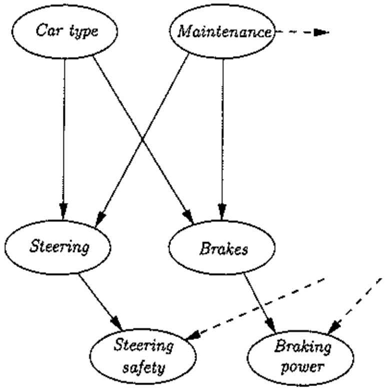


Fig. 3.38. A partial Bayesian network model for the safety characteristics of a car. The dashed arrows indicate unspecified parent and child relations.

partial Bayesian network representation. We could instead construct a class representing the grip of the car with a rudimentary internal structure and simply include the interface of the class in the model. An example is shown in Figure 3.39. Figure 3.40 shows two possible specifications of a class modeling the tire grip. The leftmost class could serve as an initial approximation to the more detailed specification shown at the right-hand side of Figure 3.40.

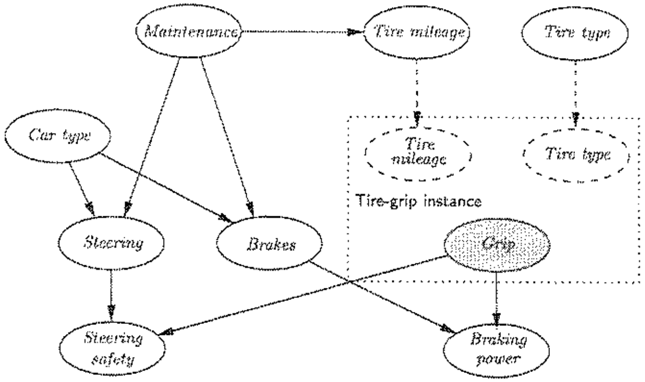


Fig. 3.39. An object-oriented Bayesian network model of the driving characteristics of a car.

Subclassing and Inheritance

A powerful property of object-oriented modeling is the use of subclassing (or inheritance) between classes. When a class C' is a *subclass* of another class

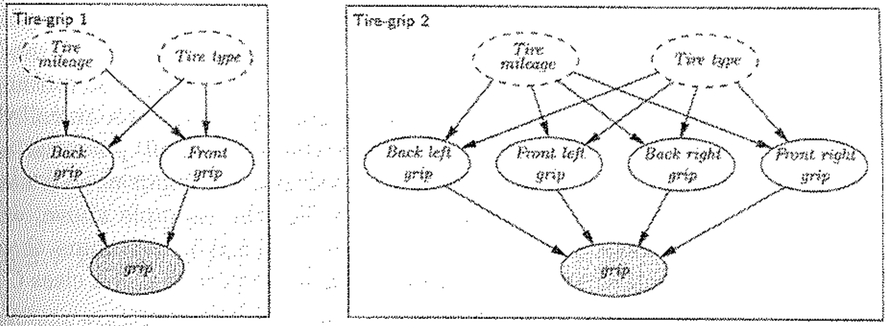


Fig. 3.40. Two possible refinements of the interface for the grip class illustrated in Figure 3.39. In the rightmost refinement, we model the grip on each of the tires.

C (also called the *superclass* for C'), then an instance of C can always be substituted with an instance of class C'. For example, consider again the two classes in Figure 3.40. We wish for the class Tire grip 2 to be viewed as a subclass of Tire-grip 1, which means that any instance of Tire-grip 1 can be substituted with an instance of Tire-grip 2. This example is quite obvious, since the two classes have the same interface connecting them to the rest of the model. However, suppose now that we should refine our grip model so that it also covers the car type; we assume that for a car with front-wheel drive there is a tendency for the front tires to be more worn than for a car with rear-wheel drive (conversely for cars with rear-wheel drive). One way to include these considerations into the model is to construct a class as in Figure 3.41.

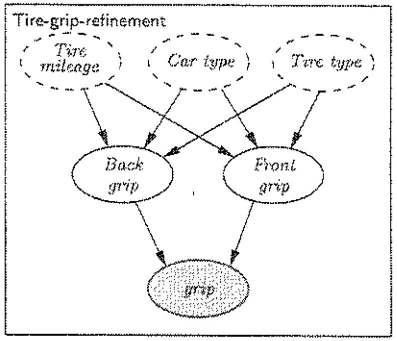


Fig. 3.41. The class Tire-grip-refinement taking the car type into account.

We would now like to be able to replace the instance in Figure 3.39 with an instance of class Tire-grip-refinement. However, this raises a technical question:

If we simply replace the instance in Figure 3.39 without connecting the input node *Car type* to an actual node in the model, then both *Back Grip* and *Front Grip* would have a parent with an unspecified probability distribution (see Figure 3.42). In order to avoid this problem, we associate a so-called *default potential* with each input node in the class; a default potential is simply a probability distribution that will be used when an input node is not connected to a node in the surrounding model. For the example above, we could specify the default potential $P(\textit{Car type}) = (0.5, 0.5)$, assuming that the node is binary. Based on these considerations we require that if a class C' should be a *subclass* of another class C, then it should hold that:

- the set of input variables for C is a subset of the input variables for C', and
- the set of output variables for C is a subset of the output variables for C'.

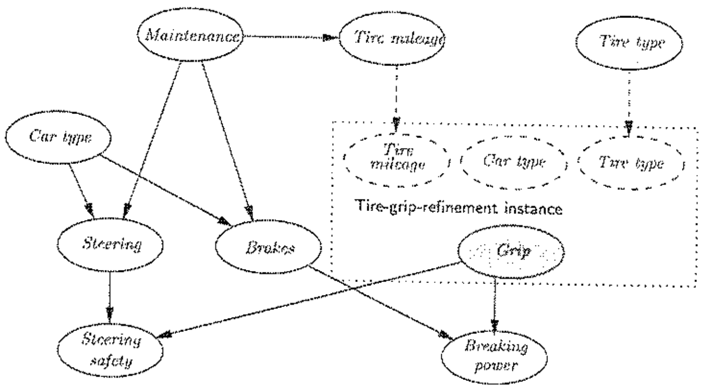


Fig. 3.42. An object-oriented Bayesian network model of the driving characteristics of a car. The input node *Car type* is associated with the default potential $P(\textit{Car type}) = (0.5, 0.5)$.

We can construct additional subclasses of Tire-Grip representing different aspects of the grip of the car. The classes can be organized in a hierarchy according to their subclass/superclass relationship. In turn we can view this class hierarchy as a model repository that facilitates a quick top-down model construction, and for more general settings, we can construct generic repositories of classes representing common modeling problems.

When we subsequently use the object-oriented Bayesian network model for answering queries (i.e., doing belief updating), we first observe that an object-oriented Bayesian network can be seen as a standard Bayesian network with some extra features for simplifying the model specification. This also implies that inference in an OOBN can be performed by first transforming the model into a standard Bayesian network, and then applying any inference algorithm on the produced network (see Chapter 4). Transforming an OOBN into a BN is basically a matter of recursively merging each input node with its parent in the surrounding model. Methods have also been developed whereby you keep the OOBN structure and respect the privacy of the encapsulated attributes. The inference method transmits probability distributions only over the interface nodes between the objects.

3.3.7 Dynamic Bayesian Networks

When working with domains that evolve over time, you can introduce a discrete time stamp and have a model for each unit of time. We call such a local model a *time slice*. Consider, for example, the model for infected milk in Figure 3.43.

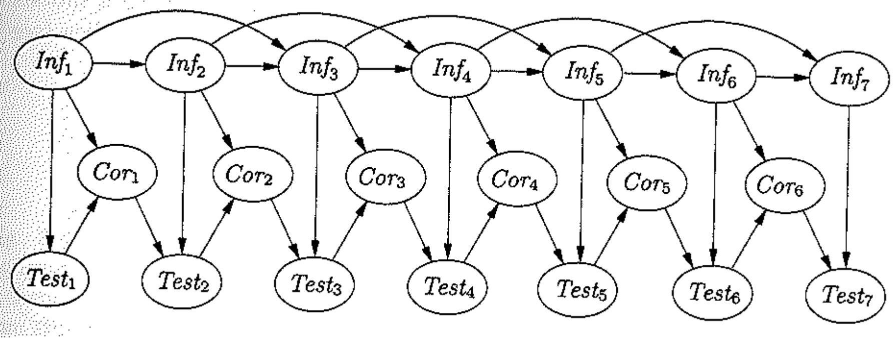


Fig. 3.43. A seven-day model with a two-day memory for infection as well as correctness of test.

For each time slice i, you have three variables Inf_i, $Test_i$, and Cor_i. The three variables are connected in a time slice, as shown in Figure 3.44.

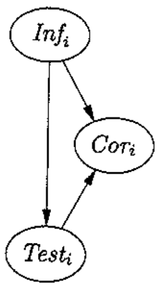


Fig. 3.44. A time slice for infected milk.

The time slices are connected through *temporal links* to constitute a full model. If the structures of the time slices are identical, and if the temporal links are the same, we say that the model is a *repetitive temporal model*. If the conditional probabilities are also identical, we call the model a *dynamic Bayesian network model*.

The model for transmission of symbols in Section 3.2.4 can be considered a temporal repetitive model, but it is not a dynamic Bayesian network because the conditional probabilities are not identical. On the other hand, the seven day model in Figure 3.2 is a dynamic Bayesian network.

A special category of time-stamped model is that of the *hidden Markov models*. They are strictly repetitive models with an extra assumption (the Markov property): the past has no impact on the future given the present. The model in Figure 3.2 is an example of a hidden Markov model, but in Figure 3.43 influence from Inf_{i-1} may flow to Inf_{i+1} regardless of our knowledge of time slice i. The latter model can, however, be transformed to a hidden Markov model by introducing a copy Inf_i^* of Inf_{i-1} in the ith time slice (see Figure 3.45).

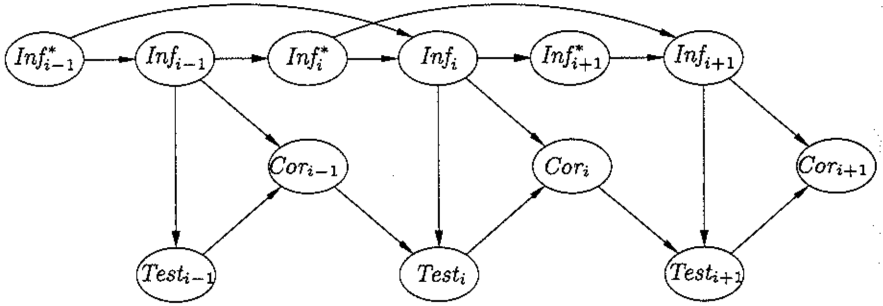


Fig. 3.45. The model of Figure 3.43 transformed into a hidden Markov model.

The reason for the term *hidden* Markov model is that under the surface (the test results) there is a hidden activity that cannot be observed (the infections).

A *Kalman filter* is a hidden Markov model in which exactly one variable has relatives outside the time slice. The model in Figure 3.2 is a Kalman filter. A *Markov chain* is a Kalman filter consisting of exactly one variable in each time slice. Note that a hidden Markov model can be transformed to a Markov chain by taking the cross product of all variables in each time slice.

In modeling domains that are evolving over time, there is a distinction between *finite-horizon* and *infinite-horizon* domains. The infected milk problem is an infinite-horizon domain, and a typical finite-horizon domain is a cornfield from sowing to harvest.

Specifying a repetitive temporal model can be eased by introducing a couple of new features to the specification language. Apart from the structure of a time slice, you must specify the number of time slices and the temporal links. The number of slices can be written in a special box, and you can introduce a special kind of arrow to specify temporal links. A number attached to a temporal link can specify the number of time steps to jump (if no number is specified, the link goes from slice i to slice $i + 1$). In Figure 3.46, we have used an extended specification language for the model in Figure 3.43.

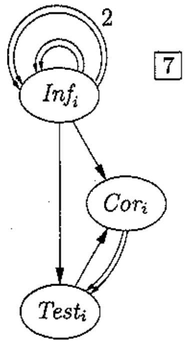


Fig. 3.46. A compact specification of the model in Figure 3.43 (an extension of Figure 3.44). The ⇉ indicates a temporal link. The number "2" attached to one of them specifies that it jumps two time steps (no number attached means a jump from slice i to slice $i + 1$).

Dynamic Bayesian networks are easily modeled through the object-oriented approach: the output variables are the variables with a child in later time slices, and the input variables are parents from earlier time slices. In Figure 3.46 the output variables are Inf_i and Cor_i, and the input variables are Inf_{i-1}, Inf_{i-2}, and Cor_{i-1}.

So from a modeling point of view, it is quite straightforward to work with time-stamped models. However, they will often yield calculational problems (see Exercise 3.25 and Chapter 4).

3.3.8 How to Deal with Continuous Variables

Consider the *Cold or Angina?* example from Section 3.1.2, in which the variable *Fever?* was given a discrete state space with three states (chosen a bit arbitrarily). A more natural way of representing fever would be to use a continuous variable (typically drawn using a double circle as in Figure 3.47(a)).

With a continuous variable we can no longer encode the uncertainty using a conditional probability table. Instead we will have to specify a *density function* for each combination of states for the parent variables for *Fever?*. A typical

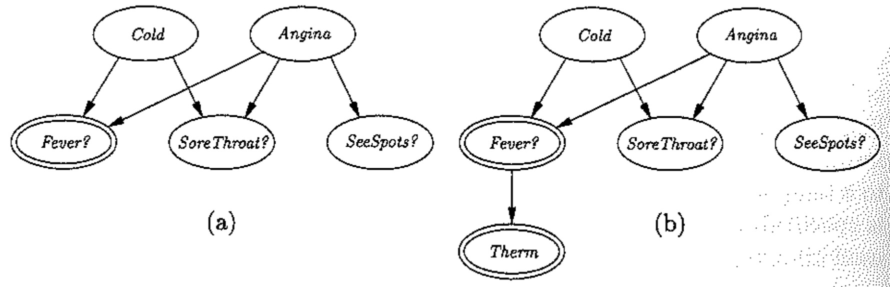


Fig. 3.47. Figure (a) shows the cold and angina model in which *Fever?* is represented by a continuous variable (drawn as a double circle). In Figure (b) the model is extended with another continuous variable *Therm* that models the accuracy of the thermometer.

density function is the normal distribution (or Gaussian distribution), which is defined by a mean μ and a variance σ^2 (see Figure 3.48 for examples):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right).$$

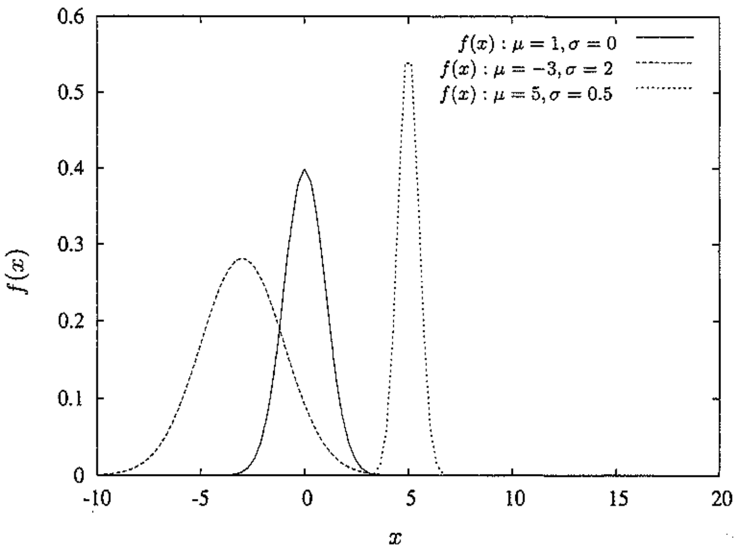


Fig. 3.48. Example of normal distributions with different values for the mean and the variance.

For the example above, we should therefore specify a μ and a σ^2 for each state combination of the variables *Cold* and *Angina* (the resulting function is also called a *conditional Gaussian distribution*). A possible specification could be as in Table 3.19.

		Cold?	
		no	*yes*
Angina?	*no*	(37°C, 0.25)	(37.5°C, 0.75)
	mild	(38°C, 0.5)	(38.5°C, 1)
	severe	(39°C, 0.75)	(39.5°C, 1.25)

Table 3.19. Means and variances for the *Fever?* variable.

The model in Figure 3.47(a) can be extended to also represent the accuracy of the thermometer. Specifically, the thermometer that I use is rather old with an accuracy corresponding to a variance of 0.25. In addition to this it has a peculiar tendency of showing 1°C plus 5% more than the actual temperature. This situation is modeled in Figure 3.47(b). The continuous variable *Therm* represents the thermometer, and it is assigned a conditional Gaussian distribution, where the variance is set to 0.25 and the mean is specified as a linear function of *Fever?*:

$$\mu_{Therm} = 1.0 + 1.05 \cdot x_{Fever?}.$$

Given this model, we can now answer queries such as $P(Cold \mid Therm = 39.2°C, SoreThroat? = yes, SeeSpots? = no)$ and $f(Fever \mid Therm = 39.2°C, SoreThroat? = yes, SeeSpots? = no)$; the latter density is a linear combination of conditional Gaussian distributions. For example, if we use the probabilities specified in Section 3.2.5 together with the conditional Gaussian distributions described above we get $P(Cold \mid Therm = 39.2°C, SoreThroat? = yes, SeeSpots? = no) = (0.13(y), 0, 87(n))$, and for $f(Fever \mid Therm = 39.2°C, SoreThroat? = yes, SeeSpots? = no)$ we get a mean and a variance of 36.67°C and 0.127, respectively. We will not present the methods for calculating posterior probabilities in networks with continuous variables.

Bayesian networks containing both discrete and continuous variables are also called *hybrid Bayesian networks*. Unfortunately, in order to perform exact probability updating in these types of networks we need to put some rather severe constraints on the networks. In general, we require that:

- Each continuous variable be assigned a (linear) conditional Gaussian distribution. That is, for each configuration $\mathbf{c}$ of the discrete parents, the variance $\sigma^2_{\mathbf{c}}$ is a constant (independent of the continuous parents) and the mean $\mu_{\mathbf{c}}$ is a linear function of the continuous parents $Y_1, \ldots, Y_m$:

$$\mu_{\mathbf{c}} = a_{\mathbf{c}} + \sum_{i=1}^{m} a^i_{\mathbf{c}} y_i.$$

- No discrete variable have continuous parents.

Note that if a continuous variable does not have any parents, then it is assigned an unconditional normal distribution.

Whether these two constraints can be met is strongly dependent on the domain being modeled. For example, you may argue that it is inappropriate to assign a conditional Gaussian distribution to the *Fever?* variable, since the distribution is defined over the entire real line and it will therefore also assign a nonzero probability mass to impossible temperature intervals. On the other hand, when specifying probabilities you are almost always making some kinds of approximations, and the question is then whether the specified Gaussian distribution is within an acceptable distance from what you deem the "correct" distribution. If it is not, you have to look for other ways of specifying the probabilities (an example of this is given below). The second constraint is more serious, since it puts restrictions on the structure of the domains that can be modeled. For instance, if we were to extend the model with a child, *Headache?* (having states *yes* and *no*), of *Fever?*, then the structural constraint would be violated.

If it is not possible to meet the two constraints above, then one possibility would be to approximate by discretizing the continuous variables. Assume that we have the specification in Table 3.19, and we should now specify intervals for a finite set of states. For the three states *no*, *low*, and *high*, it would be natural to use knowledge of fever. In other situations, you would try to determine intervals such that for each parent configuration most of the probability mass is concentrated in a few intervals. This may not be possible, and it will often be a delicate matter to establish a good set of intervals. In the current situation, we define low fever to be in the interval (37.5°C, 38.5°C). Consequently, *no* is $(-\infty, 37.5°\text{C})$ and *high* is $(38.5°\text{C}, \infty)$. Next, you use Table 3.19 to calculate the probability mass for each interval. The result is given in Table 3.20.

		Cold?	
		no	yes
Angina?	*no*	(0.834, 0.165, 0.01)	(0.5, 0.376, 0.124)
	mild	(0.24, 0.52, 0.24)	(0.159, 0.341, 0.5)
	severe	(0.042, 0.24, 0.718)	(0.037, 0.149, 0.814)

Table 3.20. The result of sampling Table 3.19 to the intervals for *no*, *low*, and *high*.

3.3.9 Interventions

You may wish to incorporate actions that change the state of some variables. You may, for example, wish to model the result of cleaning the spark plugs in the car start problem. If you use the model in Figure 2.16 directly and enter your cleaning of the spark plugs by entering *SP* = *yes*, you get incorrect results. The problem is that you may no longer have a start problem, and the state of *St* may be changed due to your action. The problem is called *persistence*. You may extend the model in Figure 2.16 with a variable *Clean?*, but then you also must introduce new nodes for the variables that may change state. Because you have a causal model, the nonpersistent nodes are the descendants of the nodes affected by the intervention (see Fig. 3.49). The variable *Clean?* has a special status in the model. It is not meaningful to give it prior probabilities, and the descendants of the nodes have no meaning before a decision on *Clean?* has been taken. Therefore, it is customary to give this kind of node a rectangular shape.

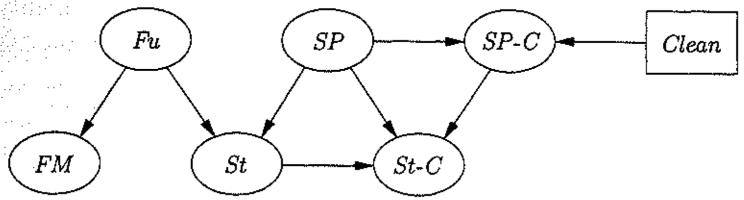


Fig. 3.49. A network modeling the effect of cleaning the spark plugs.

The conditional probabilities for new nodes are natural. If *Clean?* = *no*, then *SP-C* is in the same state as *SP*, and if *Clean?* = *yes* and *SP* = *yes*, then the probability that *SP-C* = *no* is the probability that you can clean the spark plugs properly. For *St-C*, you still have a start problem unless it was due to dirty spark plugs and they have been properly cleaned.

3.4 Special Features

A Bayesian network model is primarily used for belief updating. However, you may request other kinds of information from a model. This section outlines some types of requests. Chapter 5 gives a more detailed presentation. To illustrate the features in this section, we use the sore throat example from Section 3.1.2 (see Figure 3.50). However, we change the potentials slightly: when I suffer from mild angina, I will see yellow spots with probability 0.01, and it also happens with probability 0.001 that I have severe angina without a sore throat, provided that I do not have a cold. The rest of the potentials can be found in Sections 3.2.5 and Section 3.3.8.

We use the evidence $e = \{Fever? = no, SoreThroat? = no, See\ Spots? = yes\}$ (do not ask why I looked down my throat).

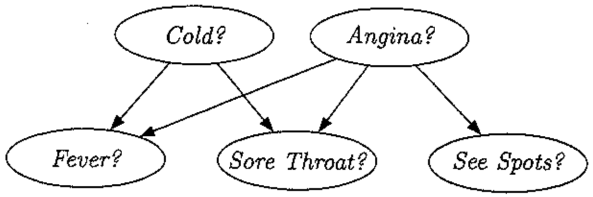


Fig. 3.50. The sore throat model.

3.4.1 Joint Probability Tables

Because it is not unusual to suffer from both cold and angina, it may be of interest to use the model in Figure 3.50 to calculate the joint probability table $P(\mathit{Angina?},\ \mathit{Cold?} \mid e)$. This can be done by use of the fundamental rule

$$P(\mathit{Angina?},\ \mathit{Cold?} \mid e) = P(\mathit{Angina?} \mid \mathit{Cold?}, e) P(\mathit{Cold?} \mid e).$$

Read $P(\mathit{Cold?} \mid e)$ from the system; then enter and propagate first *Cold?* = *yes* and then *Cold?* = *no* to get $P(\mathit{Angina?} \mid \mathit{Cold?},\ e)$.

This method is conceptually easy, but if you request the joint table for many variables, it is computationally very time-consuming. Other methods are presented in Chapter 5.

3.4.2 Most-Probable Explanation

Instead of requesting the full joint probability table, I may request the most-probable configuration of *Cold?* and *Angina?*. This can be achieved much faster than by calculating $P(\mathit{Cold?}, \mathit{Angina?} \mid e)$ and picking the state with highest probability.

In general, you have a set of instantiated variables and you request the most-probable configuration of the remaining variables. This is also called the *most-probable explanation*, MPE. MPE can be calculated similarly to probability updating (see Section 2.3.4 and Chapter 4). The only difference is that instead of marginalizing by summing out, you take the maximum. The distributive law for max reads $\max(ab, ac) = a \max(b, c)$. In the general form, it says

$$\text{If } A \notin \operatorname{dom}(\phi_1)\,,\ \text{then} \max_A \phi_1\phi_2 = \phi_1 \max_A \phi_2.$$

Most Bayesian network systems have a special feature for calculating MPE.

3.4.3 Data Conflict

Although the evidence e yields posterior probabilities for *Cold?* as well as for *Angina?*, it is more likely that I have misinterpreted what I saw in the throat.

In other words, in the light of neither fever nor sore throat, it is very likely that the evidence *See Spots?* = *yes* is faulty. It would be nice if the system by itself could raise a flag indicating that the evidence does not seem coherent.

To investigate coherence of the evidence, a *conflict measure* is defined. The idea behind the measure is that correct findings from a coherent case covered by the model support each other, and therefore we will expect them to be positively correlated. For example, if e_1 and e_2 are two pieces of evidence, then we would expect $P(e_1 \mid e_2) > P(e_1)$ and therefore $P(e_1, e_2) = P(e_1 \mid e_2)P(e_2) > P(e_1)P(e_2)$. Let $e = \{e_1, \ldots, e_m\}$ be a set of findings. Based on the intuition above, the conflict measure on e is defined as

$$\operatorname{conf}(e) = \log_2 \frac{P(e_1) \cdots P(e_m)}{P(e)}.$$

The conflict measure is easy to calculate because $P(e)$ is communicated by the system (see Example 3.9) and $P(e_i)$ can be read from the model in its initial state. If conf(e) is positive, the findings are not positively correlated, and we can take this as an indication that the evidence is conflicting. To be quite accurate, a high conflict measure is an indication that there is discrepancy between model and evidence. This may be due to flawed findings, it may be because we are faced with a very rare case, or the situation may not be covered by the model. This is discussed in more detail in Section 5.5.

3.4.4 Sensitivity Analysis

Sensitivity analysis refers to analyzing how sensitive the conclusions (the probabilities of the hypothesis variables) are to minor changes. The changes may be variations of the parameters of the model or may be changes of the evidence (*SE analysis*). In general, sensitivity analysis is rather technical and in this section we only give some hints. It is treated in more detail in Chapter 5.

Consider the angina example. The conclusion is $P(\mathit{Angina?} \mid e) = (0, 0.98, 0.02)$. SE analysis consists in answering questions such as, "what are the crucial findings?", "what if one of the findings was changed or removed?" or "what set of findings would be sufficient for the conclusion?" If we consider the conclusion to be that I suffer from mild angina, we see that the finding *See Spots?* = *yes* is not sufficient in itself because it indicates severe angina, nor is any of the other findings. Instead, *See Spots?* = *yes* together with *SoreThroat* = *no* is sufficient, and with these two findings fixed, the conclusion is insensitive to any finding on *Fever?*.

Now consider the parameters $t = P(\mathit{SoreThroat?} = \mathit{no} \mid \mathit{Angina?} = \mathit{severe}, \mathit{Cold?} = \mathit{no})$ and $s = P(\mathit{See\ Spots} = \mathit{yes} \mid \mathit{Angina?} = \mathit{mild})$. The initial values of t and s are 0.001 and 0.01, respectively. What we might look for is a functional expression for $P(\mathit{Angina?} = \mathit{mild} \mid e)(t)$ and $P(\mathit{Angina?} = \mathit{mild} \mid e)(s)$. This is called one-way sensitivity analysis. We might also request two-way sensitivity analysis by establishing $P(\mathit{Angina?} = \mathit{mild} \mid e)(t, s)$.

It follows from a general theorem that $P(e)(t)$ as well as $P(Angina? = mild, e)(t)$ are linear expressions (see Section 5.7), and hence $P(Angina? = mild \,|\, e)(t)$ is a quotient of two linear expressions. From the initial propagation, we can acquire $P(e)(0.001)$ and $P(Angina? = mild \,|\, e)(0.001)$. By changing t to 0.002 and propagating, we get $P(e)(0.002)$ and $P(Angina? = mild \,|\, e)(0.002)$. These four values are sufficient for determining the four constants in the functional expression for $P(Angina? = mild \,|\, e)(t)$.

3.5 Summary

Types of Variables in Building a Bayesian Network Model

Hypothesis variables: Variables with a state that is asked for. They are, however, either impossible or too costly to observe directly.
Information variables: Variables that can be observed.
Mediating variables: Variables introduced for a special purpose. It may be to properly reflect the independence properties in the domain, to facilitate the acquisition of conditional probabilities, to reduce the number of distributions to acquire for the network, or for other purposes.

Warning: It is tempting to introduce mediating variables in order to have a more refined model of the domain; however, if they do not serve any other purpose you should get rid of them. They jeopardize performance.

Acquiring Conditional Probabilities

Theoretically well founded probabilities as well as frequencies and purely subjective estimates can be used in the same network.

If the number of distributions is too large for a reasonable estimation, a simplifying assumption can reduce it.

Noisy-or: Let B have the parents $A_1, \ldots, A_n$ (all variables binary). Suppose that $A_i = y$ causes $B = y$ unless it is inhibited by an inhibitor Q_i that is active with probability q_i. Assume that the inhibitors are independent. Then,

$$P(B = n \,|\, a_1, \ldots, a_n) = \prod_{j \in Y} q_j,$$

where Y is the set of indices for the states y.

Divorcing: Let B have the parents $A_1, \ldots, A_n$. Assume that the set of configurations of $(A_1, \ldots, A_i)$ can be partitioned into the sets $c_1, \ldots, c_m$ such that whenever two configurations a_1^* and a_2^* of $(A_1, \ldots, A_i)$ are elements in the same c_j, then

$$P(B \,|\, a_1^*, A_{i+1}, \ldots, A_n) = P(B \,|\, a_2^*, A_{i+1}, \ldots, A_n).$$

Then, $A_1, \ldots, A_i$ can be divorced from $A_{i+1}, \ldots, A_n$ by introducing a mediating variable C with states $c_1, \ldots, c_m$, making C a child of $A_1, \ldots, A_i$ and a parent of B.

Other Tricks

Undirected relations – in particular, logical constraints – can be modeled by introducing a dummy child of the constrained variables and letting its potential reflect the relation.

For a specification language for *repeating structures*, see Figure 3.51.

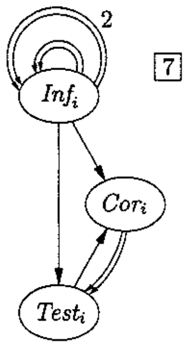


Fig. 3.51. A compact specification of a repeating structure with 7 slices. The ⇉ indicates a temporal link. The number "2" attached to one of them specifies that it jumps two time steps (no number attached means a jump from slice i to slice $i+1$).

Expert disagreements on potentials can be represented in a model by introducing a node representing the experts.

Continuous variables can be represented in the model if:

- they do not have any discrete children, and
- they are assigned a linear conditional Gaussian distribution.

If these two conditions cannot be met, an alternative is to transform them into variables with a finite number of states.

3.6 Bibliographical Notes

Naive Bayes was used by de Dombal *et al.* (1972) and can be traced back at least to Minsky (1963). Noisy-or was first described by Pearl (1986); divorcing was used in MUNIN (Andreassen *et al.*, 1989). Exercise 3.27 is based on (Cooper, 1990). Chain graphs are treated in depth in (Lauritzen, 1996). Dynamic Bayesian networks are described in (Kjærulff, 1992). The compact representation of repetitive structures was suggested by Bangsø and Wuillemin (2000). Andreassen (1992) discusses various ways of transforming conditional Gaussian variables into finite variables. A method not described in this chapter is *similarity networks* (Heckerman, 1990). The method helps in eliciting the conditional probabilities. Other elicitation methods can be found in (Druzdzel and van der Gaag, 1995). Object oriented Bayesian networks were introduced in (Koller and Pfeffer, 1997); the version presented here is the one from (Bangsø and Wuillemin, 2000). References for the special features in Section 3.4 are given in Section 5.9.

3.7 Exercises

Exercise 3.1. Peter is currently taking three courses on the topics of probability theory, linguistics, and algorithmics. At the end of the term he has to take an exam in two of the courses, but he has yet to be told which ones. Previously he has passed a mathematics and an English course, with good grades in the mathematics course and outstanding grades in the English course. At the moment, the workload from all three courses combined is getting too big, so Peter is considering dropping one of the courses, but he is unsure how this will affect his chances of getting good grades in the remaining ones. What are reasonable variables of interest in assessing Peter's situation? How do they group into information, hypothesis, and mediating variables?

Exercise 3.2. Assume that three mornings in a row I wonder whether my sore throat is due to cold or angina. Construct a model.

Exercise 3.3. Construct a model extending the model in Section 3.1.3 with a scanning test.

Exercise 3.4. Consider the following variables relating to a single household consisting of a couple and possibly some children:

- *Illness at the moment*, with states *severe illness*, *minor illness*, and *no illness*.
- *History of illness*, with states *cases of severe illness*, *often minor illnesses*, and *rarely minor illness*.
- *Number of children*, with states *none*, *one*, *two*, *three*, and *four and up*,
- *Working parents*, with states *both*, *father*, *mother*, and *none*.
- *Religion*, with states *Christianity*, *Judaism*, *Islam*, *Buddhism*, *Atheism*, and *other*.
- *Household income*, with states *$0–$50000*, *$50000–$100000*, and *$100000–and up*.
- *Fish-eating habits*, with states *often fish* and *rarely fish*.
- *Fiber-eating habits*, with states *lots of fiber* and *not much fiber*.
- *Drinking habits*, with states *never alcohol*, *wine once in a while*, *often wine*, and *wine every day*.

Try to construct a Bayesian network incorporating the above variables according to your perception of the world. What are the d-separation properties of the network you constructed?

Exercise 3.5. ^{E} Construct a model for a single milk test (Section 3.2.1). What is the probability of infected milk given a positive test result?

Exercise 3.6. ^{E} Ground meat purchased in the supermarket may be infected. On average, it happens once out of 600 times. A test with results *positive* and *negative* can be used. If the meat is *clean*, the test result will be *negative* in 499 out of 500 cases, and if the meat is *infected*, the test result will be *positive* in 499 out of 500 cases.

Construct a Bayesian network and use a software system to calculate the probability of *infected* for meat with a positive test result.

Exercise 3.7. ^{E} Complete the Bayesian network for Cold or angina? and perform a self-diagnosis.

Exercise 3.8. ^{E} Consider the insemination example from Section 3.1.3. Let the probabilities be as in Table 3.21 ($Ho = y$ means that hormonal changes have taken place) $P(Pr) = (0.87, 0.13)$.

	$Pr = y$	$Pr = n$
$Ho = y$	0.9	0.01
$Ho = n$	0.1	0.99

	$Ho = y$	$Ho = n$
$BT = y$	0.7	0.1
$BT = n$	0.3	0.9

	$Ho = y$	$Ho = n$
$UT = y$	0.8	0.1
$UT = n$	0.2	0.9

Table 3.21. Tables for Exercise 3.8.

(i) What is $P(Pr \mid BT = n, UT = n)$?
(ii) Construct a naive Bayes model. Determine the conditional probabilities for the model using the model above. What is $P(Pr \mid BT = n, UT = n)$ in this model?

Exercise 3.9. ^{E} Use the model from Exercise 3.8 to calculate $P(UT = y, BT = y)$. Enter the two pieces of evidence into the model and prompt your system to update probabilities. As a side effect, the system computes $P(e)$, the probability of the evidence entered. Find out how your system provides it.

Exercise 3.10. ^{E}

(i) Implement the seven-day model in Figure 3.13. Are the initial probabilities stable over time?

(ii) Consider the conditional probability tables $P(\mathit{Inf}_2 \,|\, \mathit{Inf}_1)$ and $P(\mathit{Inf}_1) = (0.0007, 0.9993)$ and assume that the risk of becoming infected is 0.0002. We require that the initial probabilities be stable: $P(\mathit{Inf}_2) = P(\mathit{Inf}_1) = (0.0007, 0.9993)$. Show that the chance of being cured must be 2/7.

(iii) Consider the conditional probabilities $P(\mathit{Inf}_{i+2} \,|\, \mathit{Inf}_i,\ \mathit{Inf}_{i+1})$, and assume that the risk of being infected is the same as above. We require stable initial probabilities. Show that the chance of being cured for a more than one day infection must be 0.4.

Exercise 3.11. Show that the procedure described in Section 3.1.5 is equivalent to updating in the model in Figure 3.12.

Exercise 3.12. ^{E} Consider the stud farm example in Section 3.2.2 and let the prior probability for aA be 0.005.

(i) Enter the model into your Bayesian network system.

(ii) Add to the model the frequency 0.001 for mutation of the gene from A to a.

(iii) Construct a model for the situation in part (ii), but for a recessive gene borne by the female sex chromosome. (Note that horses with the disease are taken out of production.)

Exercise 3.13. ^{E} Consider the transmission example from Section 3.2.4.

(i) From Table 3.10, calculate the remaining conditional probabilities for the model in Figure 3.18.

(ii) Implement the model.

(iii) The sequence *baaca* is received. What is the most-probable symbol transmitted according to the model in Figure 3.18? What is the most-probable word?

(iv) What is the most-probable word according to the model in Figure 3.19?

Exercise 3.14. ^{E} Consider the simplified poker game in Sections 3.1.4 and 3.2.3.

(i) Implement the system.

(ii) Extend the system with a facility giving the chances that your hand is better than your opponent's hand.

Exercise 3.15. ^{E} Construct a naive Bayes model of the simplified poker game example in Sections 3.1.4 and 3.2.3 with *OH2* being the class variable. Use your implemented model from Exercise 3.14 to calculate the needed probabilities for the naive Bayes model. What is $P(OH2 \,|\, FC1 = 1, FC2 = 2)$ using the model from Exercise 3.14? What is $P(OH2 \,|\, FC1 = 1, FC2 = 2)$ using the naive Bayes model?

Exercise 3.16. You are confronted with three doors, A, B, and C. Behind exactly one of the doors there is $10,000. When you have pointed at a door, an official will open another door with nothing behind it. After he has done so, you are allowed to alter your choice. Should you do that?

Exercise 3.17. Extend the model in Figure 3.23 to incorporate constraints on color and pattern for the same sock.

Exercise 3.18. The *drive* in golf is the first shot in playing a hole. If you drive with a *3-wood* (a particular type of golf club), there is a 2% risk of a miss (a bad drive), and $\frac{1}{4}$ of the good drives have a length of 180 m, $\frac{1}{2}$ are 200 m, and $\frac{1}{4}$ have a length of 220 m. You may also use a *driver* (another type of golf club). This will on average increase the length by 10%, but you will also have 3 times as high a risk of a miss. Both wind and the slope of the hole may affect the result of the drive. Wind doubles the risk of a miss, and the length is affected by 10% (longer if the wind is from behind and shorter otherwise). A downhill slope yields 10% longer drives, and an uphill slope decreases the length of the drive by 10%.

Estimate the probabilities for miss and length given the various factors.

Exercise 3.19. The *putt* is (usually) the last shot on a golf hole. My ball is lying 1 m away from the hole, and under normal circumstances I will miss 1 putt out of 10. However, when it rains, I miss 1 out of 7; if it is windy, I miss 1 out of 4; if the green is curved, I miss 1 out of 3; and if I am putting for a birdie (one under par), I miss 1 out of 2.

Estimate the probabilities for success and failure given the various factors.

Exercise 3.20. Show that the model in Figure 3.26 corresponds to the one in Figure 3.25.

Exercise 3.21. ^{E} Show that noisy or may be modeled as described in Figures 3.30 and 3.31. Apply this model to the putting problem of Exercise 3.19, and compare the number of quantities to specify.

Exercise 3.22.

(i) Complete the model in Section 3.3.4.

$$\begin{aligned}
P(\mathit{Ha}) = P(\mathit{Ha} \,|\, \mathit{Ot} = y) &= (0.93, 0.04, 0.02, 0.01),\\
P(\mathit{Ha} \,|\, \mathit{Fe} = y) = P(\mathit{Ha} \,|\, \mathit{Ho} = y) = P(\mathit{Ha} \,|\, \mathit{Fb} = y) &= (0.1, 0.8, 0.1, 0),\\
P(\mathit{Ha} \,|\, \mathit{Bt} = y) &= (0.3, 0.2, 0.2, 0.3).
\end{aligned}$$

$As \setminus Ha_1$	*no*	*mild*	*moderate*	*severe*
y	(1, 0, 0, 0)	(0.7, 0.3, 0, 0)	(0.1, 0.7, 0.2, 0)	(0, 0.1, 0.7, 0.2)
n	(1, 0, 0, 0)	(0, 1, 0, 0)	(0, 0, 1, 0)	(0, 0, 0, 1)

Table 3.22. $P(Ha \mid Ha_1, As)$ for Exercise 3.22.

(ii) Include aspirin in the basis of Table 3.22.

Exercise 3.23. Specify the model in Figure 3.4 as an OOBN.

Exercise 3.24. Construct an OOBN model for the stud farm in Section 3.2.2. Use default potentials for horses with parents outside the model.

Exercise 3.25. ^{E} Consider the model in Figure 3.52. All variables have ten states.

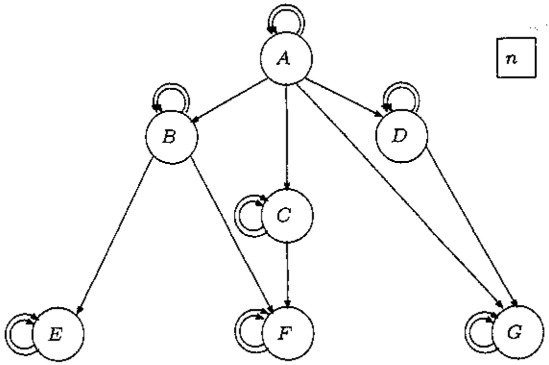


Fig. 3.52. A compact representation of a dynamic Bayesian network for Exercise 3.25.

(i) Implement one time slice (with any set of potentials).
(ii) Implement three time slices.
(iii) How many time slices can you implement before your system reports that it requires extra memory?

Exercise 3.26. ^{E} Consider a soccer tournament with 8 teams. Teams 1 to 4 are poor ones, and Teams 5 to 8 are good ones. Each match is between two teams drawn at random from those that have played the same number of matches previously in the tournament. The loser of each match is eliminated from the tournament. The probability of a good team winning a match against another team is 0.8 if the other team is a poor one, and 0.5 if the other team is a good one. The probability of a poor team winning a match against another poor team is 0.5. What is the probability of a poor team making it to the final? (Hint: For each match, generate a variable that represents the winner (with states *poor team* and *good team*), and variables that represent each contestant in the opening matches (with states *poor team* and *good team*). Finally, use constraint nodes to ensure compliance with the restrictions in the exercise.)

Exercise 3.27. ^{E} The following relations hold for the Boolean variables $A, B, C, D, E,$ and F:

$$(A \vee \neg B \vee C) \wedge (B \vee C \vee \neg D) \wedge (\neg C \vee E \vee \neg F) \wedge (\neg A \vee D \vee F) \wedge$$
$$(A \vee B \vee \neg C) \wedge (\neg B \vee \neg C \vee D) \wedge (C \vee \neg E \vee \neg F) \wedge (A \vee \neg D \vee F).$$

(i) Is there a truth value assignment to the variables making the expression true? (Hint: Represent the expression as a Bayesian network.)
(ii) We receive the evidence that A is false and B is true. Is there a truth value assignment to the other variables making the expression true?
(iii) The *satisfiability problem* for propositional calculus is, given a Boolean expression $\mathbb{E}$ (over n Boolean variables), is there a truth-value assignment to the variables that makes $\mathbb{E}$ true?
Show that a method for calculation of probabilities in Bayesian networks yields a method for solving the satisfiability problem for propositional calculus. (Hint: Assume that $\mathbb{E}$ is in conjunctive normal form.)
(iv) Show that probability calculation in Bayesian networks is NP-hard.

Exercise 3.28. You have the model $A \rightarrow B$ and $P(A) = (0.7, 0.3)$. Two experts give the tables in Table 3.23, and you have no reason to believe more in one expert than in the other.

You receive the evidence $A = y$. What are the posterior probabilities for B and the experts?

$B \setminus A$	y	n
y	0.9	0.4
n	0.1	0.6

$P_1(B \mid A)$

$B \setminus A$	y	n
y	0.6	0.4
n	0.4	0.6

$P_2(B \mid A)$

Table 3.23. Table for Exercise 3.28.

Exercise 3.29. ^{E}

(i) Take your model from Exercise 3.7 and enter the evidence $e = \{Fever? = no, Sore\ Throat? = no, See\ Spots? = yes\}$. How does your system react? Change the potentials such that $P(Sore\ Throat? = no \mid Angina? = severe, Cold? = no) = 0.001$, and $P(See\ Spots? \mid Angina? = mild) = 0.01$.
(ii) Calculate $P(Cold?, Angina? \mid e)$.
(iii) Calculate MPE(e).
(iv) Calculate conf(e).
(v) Determine $P(Angina? = mild \mid e)(s)$, where $s = P(See\ Spots? = yes \mid Angina? = mild)$.

4

Belief Updating in Bayesian Networks

In this chapter, we present algorithms for probability updating. An efficient updating algorithm is fundamental to the applicability of Bayesian networks. As shown in Chapter 2, access to $P(\mathcal{U}, e)$ is sufficient for the calculations. However, because the joint probability table increases exponentially with the number of variables, we look for more-efficient methods. Unfortunately, no method guarantees a tractable calculational task. However, the method presented here represents a substantial improvement, and it is among the most-efficient methods known.

We shall use the framework of potentials. A conditional probability table $P(A \mid \mathrm{pa}(A))$ is a function $\phi : \mathrm{pa}(A) \cup \{A\} \to [0:1]$, and we call it a potential. For the algebra of probability tables we shall for notational convenience use functional notation. That is, the product $P(A \mid \mathrm{pa}(A)) \cdot P(B \mid \mathrm{pa}(B))$ is considered as a product of two functions $\phi_1(A, \mathrm{pa}(A))\phi_2(B, \mathrm{pa}(B))$. The reader is expected to be familiar with Section 1.4.

Sections 4.1–4.6 present the junction tree algorithm, a version of the variable elimination method. Section 4.7 presents an alternative method with any-space properties, recursive conditioning, and in Sections 4.8 and 4.9 we outline different approximation methods.

4.1 Introductory Examples

To repeat the fundamentals from Chapter 2 and for pinpointing the issues in belief updating for Bayesian networks, we consider in this section a simple example. Consider the Bayesian network in Figure 4.1 over the universe $\mathcal{U}$. The potentials specified for BN are $\phi_1 = P(A_1), \phi_2 = P(A_2 \mid A_1), \phi_3 = P(A_3 \mid A_1), \phi_4 = P(A_4 \mid A_2), \phi_5 = P(A_5 \mid A_2, A_3)$, and $\phi_6 = P(A_6 \mid A_3)$.