1.4 How should this be book be used?

This book is not intended to be read linearly. It may profitably be read in various ways by different audiences. Researchers already knowledgeable in probabilistic expert systems may want to concentrate on the technical results contained in Chapters 6, 7, 8, 9, and 10, while newcomers to the area, or those seeking an overview of developments, could focus on the more descriptive material in Chapters 2, 3, parts of 9 and 10, and 11. Those interested in expert systems but unconcerned with learning algorithms could read Chapters 2, 3, 6, 7, 8. Chapters 4 and 5 can be used as references for important definitions and results or for individual study in their own right.

From Cowell, Dawid, Lauritzen,
Spiegelhalter : Probabilistic Networks
and Expert Systems

2

Logic, Uncertainty, and Probability

In this chapter we discuss characteristics of expert systems that relate to their ability to deal with the all-pervasive problem of uncertainty. We begin by describing one of the earliest approaches to the computer-based representation of expert knowledge, so called *rule-based* systems. The limitations of such systems when faced with uncertainty, and some of the alternatives that have been proposed in the literature, are highlighted. We then focus on the probabilistic representation of uncertainty, emphasizing both its strong theoretical basis and its possibility of a subjective interpretation. Bayes' theorem then forms the fundamental tool for belief revision, and 'Bayesian networks' can be formed by superimposing a probability model on a graph representing qualitative conditional independence assumptions. The resulting structure is capable of representing a wide range of complex domains.

Here we can only give a brief informal overview of the background to probabilistic expert systems; further reading material is indicated at the end of the chapter. The World Wide Web is a major resource in such a rapidly changing area; addresses of specific sites for information and software are given in Appendix C.

2.1 What is an expert system?

The *Concise Oxford English Dictionary* defines *expert* as "person having special skill or knowledge." Informally, an expert is someone you turn to when you are faced with a problem that is too difficult for you to solve on your own or that is outside your own particular areas of specialized knowledge, and whom you trust to reach a better solution to your problem than you could by yourself. Expert systems are attempts to crystallize and codify the knowledge and skills of one or more experts into a tool that can be used by non-specialists. Usually this will be some form of computer program, but this need not be the case.

An *expert system* consists of two parts, summed up in the equation:

$$\textbf{Expert System} = \textbf{Knowledge Base} + \textbf{Inference Engine.}$$

The *knowledge base* contains the domain-specific knowledge of a problem, encoded in some manner. The *inference engine* consists of one or more algorithms for processing the encoded knowledge of the knowledge base together with any further specific information at hand for a given application. Both parts are important for an expert system. Modern expert systems strive for the ideal of a clean separation of both components. This allows the knowledge base to be improved in the light of further information, and facilitates learning from the experience of making mistakes.

The knowledge base is the core of an expert system; no matter how sophisticated the inference procedures are for manipulating the knowledge in a knowledge base, if the content of the knowledge base is poor then the inferences will be correspondingly poor. Nevertheless it is vital to have a good inference engine to take full advantage of the knowledge base.

2.2 Diagnostic decision trees

A *diagnostic decision tree* (also known as a *classification tree*, *flowchart*, or *algorithm*) is a structured sequence of questions in which each response determines the next question to ask. The inference process involves simply walking through the algorithm, selecting the appropriate path from the answers to the questions contained in the nodes. The system encodes the expert knowledge in the order and form in which the questions are structured. At a certain stage a diagnosis or conclusion is reached.

An example of part of a diagnostic decision tree is shown in Figure 2.1. The background is as follows. The Great Ormond Street Hospital for Sick Children in London (here abbreviated to GOS) acts as a referral centre for newborn babies with congenital heart disease. Early appropriate treatment is essential, and a preliminary diagnosis must be reached using information reported over the telephone. This may concern clinical signs, blood gases, ECG, and X-ray. A decision tree, intended to help the junior doctors in GOS, was constructed from expert judgement. It contained 66 nodes, and discriminated 27 diagnostic categories in neonates, including lung disease masquerading as heart disease. It was developed and evaluated on 400 cases (Franklin et al. 1991).

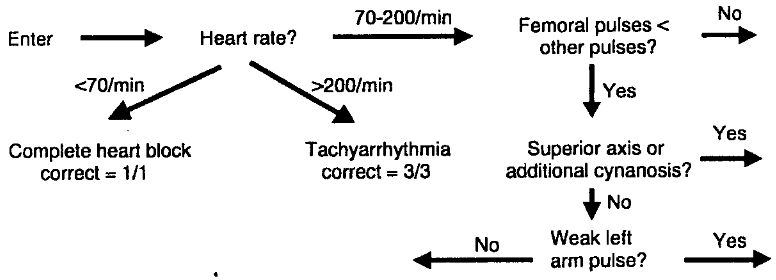

Enter
Heart rate?
70-200/min
Femoral pulses < other pulses?
No
<70/min
>200/min
Yes
Complete heart block correct = 1/1
Tachyarrhythmia correct = 3/3
Superior axis or additional cynanosis?
Yes
No
No
Weak left arm pulse?
Yes

FIGURE 2.1. Initial part of Great Ormond Street diagnosis decision tree for diagnosing problems in newborn babies. The first question is *Heart rate?*, and, depending on the answer, one of three paths is chosen. For example, if the heart rate is greater than 200 beats a minute, an immediate diagnosis of *Tachyarrhythmia* is made. The *correct* $= 3/3$ in the figure indicates that in the available database there were 3 cases that went down this path, all of which actually had a correct final diagnosis of Tachyarrhythmia.

A classification tree does not necessarily require a computer for implementation and is generally easy to explain and use. If it performs badly for a particular case, it is usually possible to pinpoint where the wrong branch was taken. However, despite their appealing nature, classification trees suffer from some drawbacks. An incorrect conclusion can be reached after a *single* unexpected response, due for example to observer error. They are inflexible with respect to missing information. Typically default responses are assumed when a question cannot be answered; for example, in the GOS algorithm the default is to assume a negative response where data are missing. Such systems usually provide little opportunity for adaptation as data become available. We might interpret their faults as stemming from the lack of separation of the knowledge base and the inference engine, leading to a rigid non-modular system.

2.3 Production systems

A more flexible type of expert system is the *production system*, also called *rule-based system*. Such a system has its origin in the attempt to perform *symbolic reasoning* using logical-rules. Generally, in a rule-based system, domain knowledge is encapsulated by a collection of implications, called *production rules*, having the form: IF $(A_1$ & A_2 & $\ldots$ & $A_k)$ THEN B; where $\{A_i\}$ are assertions and B may be an assertion or action. The following are examples of production rules (taken from Winston (1984)).

- IF the animal has hair THEN it is a mammal.
- IF the animal gives milk THEN it is a mammal.
- IF the animal has feathers THEN it is a bird.
- IF the animal flies & it lays eggs THEN it is a bird.

Since an assertion A_i may itself be a consequence of these modular rules, chains of reasoning are established. A trace through such a chain provides a degree of explanation for a particular case under consideration.

The collection of rules forms a modular knowledge base in that it is possible easily to add further rules if desired. Although there is a reliance on logical reasoning, the questions or rules do not need to be applied in a predetermined and inflexible manner, as in classification trees. Computer programs can be written, for example in languages such as LISP or Prolog, which manipulate such symbolic production rules and logic (see Lucas and van der Gaag (1991) for examples). The inference engine is embodied as a control mechanism in the program, which can select rules relevant to the particular case under consideration and suggest additional assertions that, if true, could be useful. It can also make valid deductions from a given set of assertions, a process called *forward chaining*, and perform the reverse operation to determine whether assertions exist that can validate a conjectured property (*backward chaining*).

However, there are problems with production systems. They focus on specific *assertions*, rather than *questions* with a choice of answer. They do not automatically distinguish "found to be false" and "not found to be true" (for example, the question may not be asked). The application of the laws of logic seems somewhat incomplete, particularly with respect to negation: for example, they can go from A with $A \rightarrow B$ to B, but the reverse, $\bar{B}$ with $A \rightarrow B$ may not necessarily lead to $\bar{A}$. The number of rules can grow enormously, and it is necessary to confirm consistency and eliminate redundancy. Exceptions to the rules have to be dealt with (for example penguins are birds that do not fly). Finally, the actual chains of reasoning may become too complex to comprehend.

2.4 Coping with uncertainty

Originally, production systems involved only *logical* deductions. Although this can be adequate for representing complex but determinate structures such as legislation, some problems of a general nature arise. In particular, the data available on an individual of interest may be inadequate or insufficiently reliable to enable a conclusion to be reached, or the production rules themselves may not be logically certain.

To deal with such situations, we need to quantify *uncertainty* in the conclusions. An early attempt by the artificial intelligence (AI) community concentrated on the logical certainty of the production rules, attaching a numerical value called a *certainty factor* (CF) to each production rule. For example, a system for medical diagnosis might have productions of the form:

- IF headache & fever THEN influenza (certainty 0.7)
- IF influenza THEN sneezing (certainty 0.9)
- IF influenza THEN weakness (certainty 0.6)

An early example of a backward chaining system with certainty factors is the MYCIN program (Shortliffe and Buchanan 1975), designed to assist doctors in prescribing treatment for bacteriological blood disorders. It employs about 300 productions and was the first system to separate its knowledge base from its inference engine.

It is still possible that one production can trigger others in a chain. However, with the additional numerical structure this requires that the certainty factors associated with such a chain be combined in some manner. It may also happen that two or more different productions yield the identical assertion or action, and then the various certainty factors again have to be combined in some manner. Thus, there arises the need for an *algebra* or *calculus* of certainty factors, as illustrated in Figure 2.2. To postulate or develop a plausible calculus requires some interpretation to be given to the meaning of the numbers.

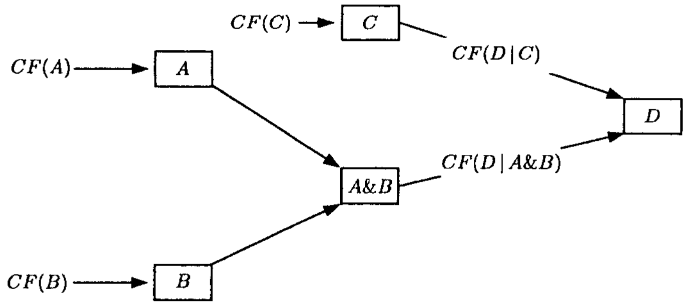


FIGURE 2.2. Combining certainty factors: How do $CF(A)$, $CF(B)$, $CF(C)$, $CF(D \mid C)$, and $CF(D \mid A \;\&\; B)$ combine to yield $CF(D \mid A \;\&\; B \;\&\; C)$?

Certainty factors can be, but have typically not been, regarded as statements of conditional probability. Although this may seem an appealing interpretation, there can be major consistency problems with this interpretation. This is because an arbitrary set of such production rules might not be compatible with any overall probability distribution, and if it is, that distribution might not be unique. Also, while 'IF A THEN B' is logically equivalent to 'IF $\bar{B}$ THEN $\bar{A}$', it is generally false that '$P(B \mid A) = q$' implies '$P(\bar{A} \mid \bar{B}) = q$'. We thus see that the desire for large modular systems made up of many smaller components or productions, together with local combination of certainty factors, appears to argue against a probabilistic interpretation of the certainty factors. This impression led to probability theory being abandoned by most of the AI community. Instead other ad hoc rules for manipulating and combining certainty factors were developed, or alternative measures of uncertainty were developed or applied, that allowed modularity to be retained, for example *fuzzy logic* (Zadeh 1983) and *belief functions* (Dempster 1967; Shafer 1976).

However, in a detailed examination of the MYCIN system, Heckerman (1986) showed that the "original definition of certainty factors is inconsistent with the functions used in MYCIN to combine the quantities." By redefining the interpretation of certainty factors he established a connection with probability theory, specifically that certainty factors can be interpreted as monotone functions of likelihood ratios. Furthermore, he showed that consistency can only be maintained or satisfied in tree-like structures.

Other reasons why probability theory has been proclaimed useless for expert systems are: first, that it is irrelevant because the uncertainty in the knowledge that is being represented does not match that of a conceptual chance mechanism underlying an observable event; secondly, that if a system is to be judged by rank order of hypotheses then a non-probabilistic calculus may be adequate; and thirdly, that a full probability distribution over many quantities would require assessment of too many numbers. We now give a short overview of how these perceived barriers were overcome.

2.5 The naïve probabilistic approach

In probabilistic terms the basic problem and solution can be stated as follows. We have a collection of unknown quantities $(A, B, \dots)$, we observe the true values for a subset of these, and we wish to derive appropriate expressions of uncertainty about the others. The solution, in principle, is quite simple. We need a *joint distribution* over all the unknown quantities, in which a probability is assigned to each possible combination of values. Then we must use the laws of probability to *condition* on the discovered facts, and hence obtain the appropriate conditional probability of those quantities that are still unknown. However, this simple solution has the following snags:

- Generally many probability assignments will be required to form the joint distribution.

- There is no modularity — just one huge computation.

- The approach is non-intuitive and lacks explanatory power.

Up to the early 1980s these problems made the application of probability theory appear infeasible in expert systems, but subsequent theoretical developments have managed to overcome or mitigate these problems and to address other conceptual concerns of the early workers in the AI community over the use of probability theory. This book deals with developments in probabilistic networks that address these and other points. The first of these is the introduction of the subjectivist Bayesian interpretation of probability into the AI context.

2.6 Interpretations of probability

The interpretation of probability continues to be a subject of intense debate, with important implications for the practice of probability modelling and statistical inference, both in general and in expert systems applications. One major division (Gillies 1994) is between *objective* and *epistemological* understandings of $P(A)$, the probability of an event A — or, more generally, of $P(A \mid B)$, the probability of A conditional on the happening of another event B.

Objective theories regard such probabilities as real-world attributes of the events they refer to, unrelated to and unaffected by the extent of our knowledge. Popper's *propensity theory* (Popper 1959), for example, in which probability measures an innate disposition of an event to occur in identified circumstances, is one such objective theory. The most influential objective interpretation has been the *frequentist* interpretation (Venn 1866; von Mises 1939), in which probabilities of events are defined as limiting proportions in an infinite ensemble or sequence of experiments. This has been the dominant interpretation of probability for most of this century and forms the basis of the influential frequentist approach to statistical inference, as developed by Neyman and Pearson (1967). However, because it only allows probabilities to be meaningfully assigned to outcomes of strictly repeatable situations and takes an uncompromising physical view of their nature, its scope is severely limited. It was the adoption of this specific interpretation by the early AI pioneers that led to the perception that there were fundamental conceptual obstacles to the incorporation of probability theory into expert systems.

Epistemological theories eschew possibly problematic 'true probabilities', instead regarding $P(A \mid B)$ as describing a state of mental uncertainty about A in the knowledge of B, where now A and B can be singular propositions, as opposed to repeatable events. From this viewpoint the probability calculus can be considered a generalization of Boolean logic (which historically came later), allowing numerical quantification of uncertainty about propositions, and describing how such numerical uncertainties should combine and change in the light of new information. Epistemological theories further divide into *logical* theories (Keynes 1921; Jeffreys 1939; Carnap 1950) and *subjectivist* theories. Logical theories posit the existence of a unique rational numerical degree of uncertainty about a proposition A in the light of information B. However, attractive though this viewpoint may be, no satisfactory theory or method for the evaluation of logical probabilities has yet been devised. In recent years the subjectivist interpretation has become popular. This does not impose any particular numerical evaluation of probabilities, but merely requires that all the probability assessments an individual makes should 'cohere' appropriately. In such a theory, $P(A)$ is a numerical measure of a particular person's subjective degree of belief in A, with probability 1 representing certain belief in the truth of A, and probability 0 expressing certainty that A is false. Thus, it is more appropriate to think of $P(A)$ as representing the probability *for* A — a characteristic of both A and the person whose probability it is. We can measure a person's subjective probabilities $P(A)$ or $P(A \mid B)$ either directly by offering bets at various odds, or indirectly by observing the subject's behaviour in situations whose uncertain consequences depend on the actions he or she takes.

From a subjectivist standpoint, it is possible to assign probabilities to individual propositions, or to treat unknown constants or parameters as random variables, even though there may be no physical stochastic mechanism at work. For example, a person could assert "My probability that the Suez canal is longer than the Panama canal is 0.2." Clearly, such a subjective probability must be relative to that person's degree of background knowledge or information. Direct frequentist interpretation of such a probability statement is not possible: there is no stochastic element or ensemble of relevant repetitions, and so no basis for assigning a frequentist probability (other than 0 or 1) to the content of the statement. However, when many subjective probability assessments are made, they do have implications for the behaviour of certain real-world frequencies, which can be investigated empirically (see Dawid (1986)).

Many authors have sought to justify a subjectivist interpretation and calculus of probability from more basic axiomatic foundations. The influential approaches of Ramsey (1926), de Finetti (1937), and Savage (1954) (see also de Finetti (1975), Savage (1971), and Lindley (1982)) are based on a decision-theoretic interpretation of subjective probability as a determinant of action, and a principle of *coherence*, which requires that an individual should not make a collection of probability assessments that could put him in the position of suffering a sure loss, no matter how the relevant uncertain events turn out. It can then be shown that coherence is attained if and only if the probability assessments satisfy the standard probability axioms (see Section 2.7).

Artificial Intelligence researchers often refer to Cox (1946), which was based on a logical interpretation of probability, but applies equally to a subjectivist one. Cox asked: if one agrees that it is possible to assign numerical values to represent degrees of rational belief in a set of propositions, how should such values combine? He assumed, for instance, the existence of some function F such that $P(C \cap B \mid A) = F\{P(C \mid A \cap B), P(B \mid A)\}$ for any three propositions A, B, and C. He then showed that there must exist a transformation of the initial numerical belief values to values in the real interval [0,1], such that the transformed values combine according to the rules of the probability calculus. Cox's paper required certain differentiability assumptions, but these have been relaxed to assumptions of continuity only (Aczél 1966); more recently Aleliunas (1990) has produced a stronger result in a discrete setting. Interestingly, the introduction of certainty factors into MYCIN was just such an attempt to use numerical values to represent uncertainty, and, as mentioned in Section 2.4, Heckerman (1986) showed that for a consistent interpretation of the way that certainty factors combine there has to be a monotonic mapping of their values to ratios of conditional probabilities.

2.7 Axioms

We assume that the reader has had some contact with probability theory, and we shall use standard results and definitions as required. However, it is useful to review the basic probability axioms. We can regard these as applying either to propositions, combining under the logical operations of the propositional calculus, or to events (subsets of a suitable sample space), combining according to the operations of set theory. Although the former interpretation is more intuitive, we shall generally use the more standard notation and terminology of the latter. For the logical conjunction $A\&B$ of events A and B we may use, interchangeably, $A \cap B$ or (A, B).

Axiom 1: $0 \leq P(A) \leq 1$, with $P(A) = 1$ if A is certain.

Axiom 2: If events (A_i) $(i = 1, 2, \ldots)$ are pairwise incompatible, then $P(\bigcup_i A_i) = \sum_i P(A_i)$.

Axiom 3: $P(A \cap B) = P(B \mid A)P(A)$.

These axioms are not quite in the form given in the standard account by Kolmogorov (1950). Our Axiom 3 relates unconditional and conditional probabilities, regarded as having independent existence on an equal footing (indeed, from our viewpoint any 'unconditional' probability is only really so by appearance, the background information behind its assessment having been implicitly assumed and omitted from the notation). Kolmogorov's approach takes unconditional probability as the primitive concept, and would therefore treat our Axiom 3 as *defining* conditional probability.

There is continuing discussion over whether the union in Axiom 2 should be restricted to finite, rather than countably infinite, collections of events (de Finetti 1975). For our purposes this makes little difference, and for convenience we shall assume full countable additivity.

Many other properties of probabilities may be deduced from the axioms, including *Bayes' Theorem* (see Section 2.8), which shows how to interchange the outcome and the conditioning events in a conditional probability.

In an epistemological approach, all quantities, be they observables or parameters, are jointly modelled as random variables with a known joint distribution. *Statistical inference* then consists simply in calculating the conditional distribution of still unknown quantities, given data. Since Bayes' theorem is the principal (though not the only) tool for performing such calculations, this approach to statistics has come to be called 'Bayesian'. This in turn has logical and subjectivist branches, although there are difficulties in constructing a fully consistent logical account (Dawid 1983). A good account of modern Bayesian statistics may be found in Bernardo and Smith (1994).

In this book we mostly adopt both the subjectivist interpretation of probability and the subjectivist Bayesian approach to statistical inference. However, in later chapters we shall also make use of non-Bayesian statistical techniques when dealing with model construction and criticism.

2.8 Bayes' theorem

Bayes' theorem is the basic tool for making inferences in probabilistic expert systems. From Axiom 3 and the fact that $P(A \cap B) = P(B \cap A)$, we immediately have

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A). \tag{2.1}$$

By rearrangement we obtain *Bayes' theorem*:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{2.2}$$

This can be interpreted as follows. Suppose we are interested in A and we begin with a *prior probability* $P(A)$, representing our belief about A before observing any relevant evidence. Suppose we then observe B. By (2.2), our revised belief for A, the *posterior probability* $P(A \mid B)$, is obtained by multiplying the prior probability $P(A)$ by the ratio $P(B \mid A)/P(B)$.

We now extend attention beyond simple events to *random variables*. Informally, a random variable is an unknown quantity that can take on one of a set of mutually exclusive and exhaustive outcomes. Such a variable, say M with values $m \in \mathcal{M}$, will have a distribution, its *prior distribution* $P(M)$, specifying the probabilities $P(m) = P(M = m)$, $m \in \mathcal{M}$. Then for any value d of another variable D, the expression $P(d \mid M)$, with values $P(d \mid m) = P(D = d \mid M = m)$, considered as a function of m, is called the *likelihood function* for M on data d. The *posterior distribution* for M given the data, $P(M \mid d)$, can then be expressed, using (2.2), by the relationship:

$$P(M \mid d) \quad \propto \quad P(d \mid M) \quad \times \quad P(M), \tag{2.3}$$

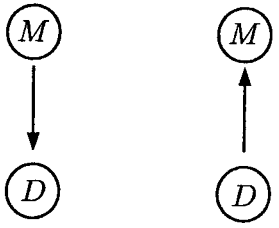
that is,

$$\textbf{Posterior} \quad \propto \quad \textbf{Likelihood} \quad \times \quad \textbf{Prior},$$

where the proportionality arises since the denominator $P(d)$ in (2.2) is the same for all values of M, and can thus be reconstructed as the normalizing constant needed to scale the right-hand side to sum to 1 over all outcomes of M.

The above discussion assumes that the random variables involved are discrete, but the identical formula (2.3) continues to hold in the case of continuous variables (or a mixture of discrete and continuous variables), so long as, when M (for example) is continuous, we interpret $P(m)$ as the *probability density* of M at m. In that case, the normalization constant $P(d)$ is given by the integral of the right-hand side.

In Figure 2.3 we display this 'prior-to-posterior inference' process pictorially. Both of the diagrams represent the structure of the joint distribution $P(M, D)$. Diagram (a) decomposes $P(M, D)$ in terms of its 'prior' components $P(M)$ and $P(D \mid M)$: often, we will think of M as a possible 'cause' of the 'effect' D, and the downward arrow represents such a causal interpretation. Diagram (b) decomposes $P(M, D)$ in terms of its 'posterior' components $P(M \mid D)$ and $P(D)$: the 'inferential' upward arrow then represents an 'argument against the causal flow', from the observed effect to the inferred cause.



(a) $P(M \mid D)P(D)$ (b) $P(D \mid M)P(M)$

FIGURE 2.3. Bayesian inference as reversing arrows.

A generalization of Figure 2.3 is illustrated in Figure 2.4. Here the variable D represents some unknown member of a set of alternative diseases, and influences the chance of the occurrence of each of a set of potential symptoms or features (F_i). We shall see later that this graph encodes an assumption that the features (F_i) are *conditionally independent* given the disease D, and hence that the joint distribution of all variables satisfies

$$P(D, F_1, \ldots, F_K) = \left(\prod_{i=1}^{K} P(F_i \mid D) \right) P(D). \tag{2.4}$$

This requires as numerical inputs only the distribution for the disease and the conditional distribution of each of the features in each of the disease categories. These can be readily estimated if we have a random sample of 'training data', in which for each case we observe the disease and some or all of the features. Calculating the posterior probability of each disease on the basis of observed findings is extremely straightforward: this simple model has been termed *naïve* — or even *idiot's* — Bayes (Titterington et al. 1981).

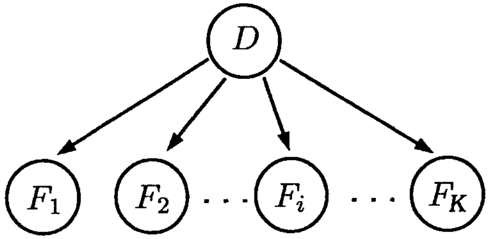


FIGURE 2.4. Directed graphical model representing conditional independence of feature variables within each disease class — the naïve Bayes model.

The naïve Bayes model was first used by Warner et al. (1961) for the diagnosis of congenital heart disease. Later applications of this model are too numerous to list, but a notable example is the acute abdominal pain system (de Dombal et al. 1972), which has been implemented in a number of hospitals and remote sites such as submarines, and has been claimed to have a significant impact on care and resources (Adams et al. 1986).

It has been argued that in most applications the assumptions underlying such a model are blatantly inappropriate. For (2.4) implies that once the disease class is known information about some feature variables is of no further relevance to predicting the values of any others. This property of the model of Figure 2.4 — the independence of features conditional on knowing the state of D — can be verified by performing the necessary calculations on the joint probability distribution. However, it can also be deduced simply from the figure, without knowing explicitly any numerical values attached to the probabilities of the model. This follows from the theory of *Markov distributions on graphs*, to be discussed in Chapter 5, which in turn relies on some aspects of graph theory described in Chapter 4.

The model of Figure 2.4 allows a simple use of Bayes' theorem, since the conditional independence assumptions mean that each item of evidence can be considered in turn, with the posterior probability distribution for the disease after observing each item becoming the prior probability distribution for the next. Thus, the sparseness of the graph leads directly to a modular form for the inference.

2.9 Bayesian reasoning in expert systems

Pearl (1982) realized that this modular approach could be generalized to more complex graphical structures and presented some elegant techniques for exploiting this vital idea of 'local computation' in graphs that are more complex than Figure 2.4 but still have a tree structure, so that removing any edge disconnects the graph. A simple example illustrates a number of points.

Suppose we wish to reason about possible personal computer failure. Let C be the variable *Computer failure?*, allowing answers "yes" and "no." The possible causes, with their assumed probabilities, are E: *Electricity failure?*, with $P(E = \text{yes}) = 0.1$, and M: *Malfunction?*, with $P(M = \text{yes}) = 0.2$. We assume that these possible precipitating events are independent, in that we have no reason to believe that the occurrence of one should influence the occurrence of the other. We also adopt the following conditional probabilities for failure:

$$\begin{aligned} P(C = \text{yes} \mid E = \text{no}, M = \text{no}) &= 0 \\ P(C = \text{yes} \mid E = \text{no}, M = \text{yes}) &= 0.5 \\ P(C = \text{yes} \mid E = \text{yes}, M = \text{no}) &= 1 \\ P(C = \text{yes} \mid E = \text{yes}, M = \text{yes}) &= 1 \end{aligned}$$

The left-hand diagram in Figure 2.5 shows a directed graphical model of this system, with each variable labelled by its current probability of taking the value "yes" (the value $P(C = \text{yes}) = 0.19$ is calculated below).

Suppose you turn your computer on and nothing happens. Then the event "$C = \text{yes}$" has occurred, and you wish to find the conditional probabilities for E and M, given this computer failure. By Bayes' theorem,

$$P(E, M \mid C = \text{yes}) = \frac{P(C = \text{yes} \mid E, M)\ P(E, M)}{P(C = \text{yes})}.$$

The necessary calculations are laid out in Table 2.1. Note that, owing to the assumed independence, $P(E, M) = P(E)P(M)$. Also $P(C = \text{yes}, E, M) = P(C = \text{yes} \mid E, M)P(E, M)$, and when summed this provides $P(C = \text{yes}) = 0.19$.

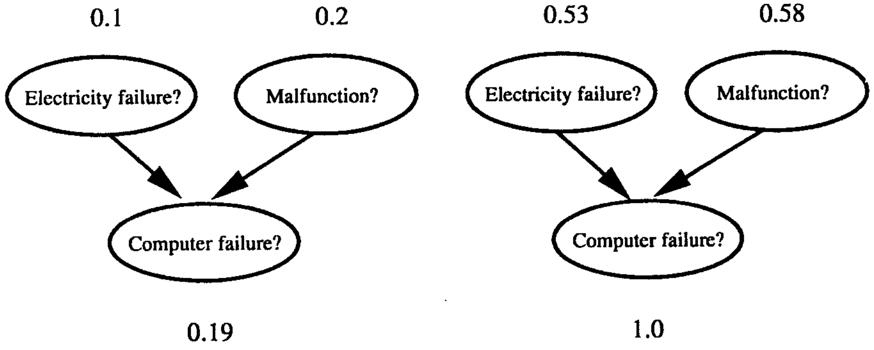


FIGURE 2.5. Directed graphical model representing two independent potential causes of computer failure, with probabilities of a 'yes' response before and after observing computer failure.

By summing over the relevant entries in the joint posterior distribution of E and M we thus obtain $P(E = \text{yes} \mid C = \text{yes}) = 0.42 + 0.11 = 0.53$ and $P(M = \text{yes} \mid C = \text{yes}) = 0.47 + 0.11 = 0.58$. These values are displayed in the right-hand diagram of Figure 2.5. Note that the observed failure has induced a strong dependency between the originally independent possible causes; for example, if one cause could be ruled, out the other *must* have occurred.

TABLE 2.1.

E $[P(E)]$	no [0.9]		yes [0.1]		
M $[P(M)]$	no [0.8]	yes [0.2]	no [0.8]	yes [0.2]	
$P(E, M)$	0.72	0.18	0.08	0.02	1
$P(C = \text{yes} \mid E, M)$	0	0.50	1	1	
$P(C = \text{yes}, E, M)$	0	0.09	0.08	0.02	0.19
$P(E, M \mid C = \text{yes})$	0	0.47	0.42	0.11	1

We now extend the system to include the possible failure, denoted by L, of the light in the room, assuming that such a failure depends only on the electricity supply, and that

$$\begin{aligned} P(L = \text{yes} \mid E = \text{yes}) &= 1 \\ P(L = \text{yes} \mid E = \text{no}) &= 0.2 \end{aligned}$$

so that $P(L = \text{yes}) = P(L = \text{yes} \mid E = \text{yes})P(E = \text{yes}) + P(L = \text{yes} \mid E = \text{no})P(E = \text{no}) = 1 \times 0.1 + 0.2 \times 0.9 = 0.28$. The extended graph is shown in Figure 2.6.

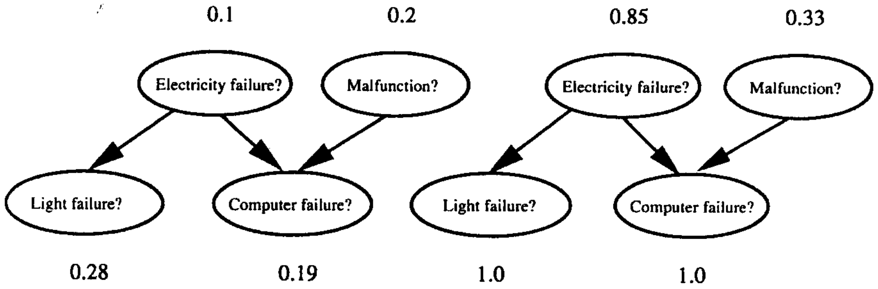


FIGURE 2.6. Introducing the roomlight into the system, before and after observing that neither the light nor the computer work.

Suppose we now find the light does not work ($L = \text{yes}$). Our previous posterior distribution $P(E, M \mid C = \text{yes})$ now becomes the prior distribution for an application of Bayes' theorem based on observing that the light has failed. Note that $P(L = \text{yes} \mid E, M, C) = P(L = \text{yes} \mid E)$, since only the electricity supply directly affects the light.

TABLE 2.2.

E	no		yes		
M	no	yes	no	yes	
$P(E, M \mid C = \text{yes})$	0	0.47	0.42	0.11	
$P(L = \text{yes} \mid E, M, C = \text{yes})$	0.2	0.2	1	1	
$P(L = \text{yes}, E, M \mid C = \text{yes})$	0	0.094	0.42	0.11	0.624
$P(E, M \mid C = \text{yes}, L = \text{yes})$	0	0.15	0.67	0.18	1

The calculations are displayed in Table 2.2. We obtain $P(E = \text{yes} \mid C = \text{yes}, L = \text{yes}) = 0.85$, $P(M = \text{yes} \mid C = \text{yes}, L = \text{yes}) = 0.33$. Thus, observing "light off" has *increased* the chance of "electricity failure," and *decreased* the chance of "malfunction": the original computer fault has been *explained away*. This ability to withdraw a tentative conclusion on the basis of further information is extremely difficult to implement within a system based on logic, even with the addition of measures of uncertainty. In contrast, it is both computationally and conceptually straightforward within a fully probabilistic system built upon a conditional independence structure.

The above example has heuristically argued for the explanatory power of probabilistic models based on Bayesian reasoning, following closely the insights of Pearl (1986b) and Pearl (1988), which largely provided the foundation for probabilistic evidence propagation in complex systems. We have not directly illustrated the specific techniques developed in these references for updating belief on any part of a tree-structured graph given evidence on any other part, techniques which can be used to organize and streamline our brute force calculations above. The approach developed in this book is more general, dealing with graphs with a more complex structure. However, many parallels with Pearl's work may be drawn.

The following fictitious example, ASIA , due to Lauritzen and Spiegelhalter (1988), illustrates the nature of the more complex graphical structures we shall be analysing in this book.

> Shortness–of–breath (dyspnoea) may be due to tuberculosis, lung cancer or bronchitis, or none of them, or more than one of them. A recent visit to Asia increases the chances of tuberculosis, while smoking is known to be a risk factor for both lung cancer and bronchitis. The results of a single chest X-ray do not discriminate between lung cancer and tuberculosis, as neither does the presence or absence of dyspnoea.

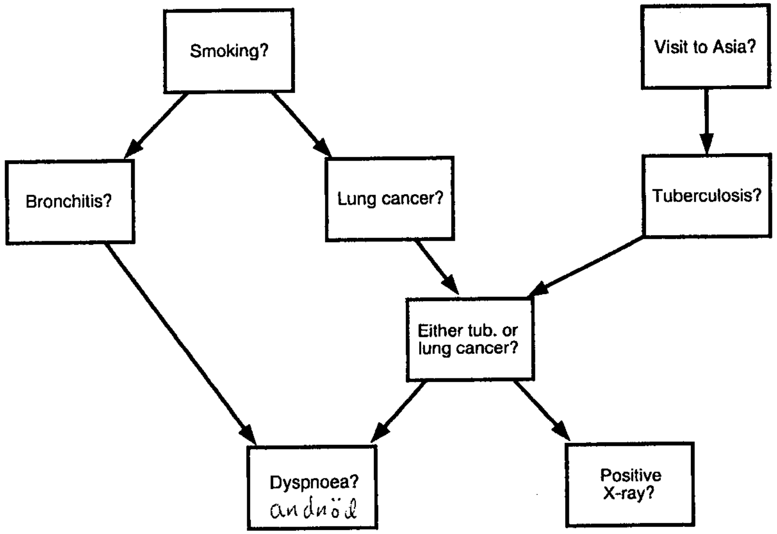


FIGURE 2.7. The ASIA network.

The qualitative structure of this example is given in Figure 2.7. Note that, as opposed to the previous example, *Smoking?* is connected to *Dyspnoea?* via two alternative routes. The quantitative specification is given in Table 2.3. Here we use B (for example) to denote the variable *Bronchitis?*, and $b, \bar{b}$ respectively for "*Bronchitis?* = yes," "*Bronchitis?* = no."

The model might be applied to the following hypothetical situation. A patient presents at a chest clinic with dyspnoea, and has recently visited Asia. Smoking history and chest X-ray are not yet available. The doctor would like to know the chance that each of these diseases is present, and if tuberculosis were ruled out by another test, how would that change the belief in lung cancer? Also, would knowing smoking history or getting an X-ray contribute more information about cancer, given that smoking may 'explain away' the dyspnoea since bronchitis is considered a possibility? Finally, when all information is in, can we identify which was the most influential in forming our judgement?

TABLE 2.3. Conditional probability specifications for the ASIA example.

A:	$p(a)$	$=$	0.01	L:	$p(l \mid s)$	$=$	0.1
					$p(l \mid \bar{s})$	$=$	0.01
B:	$p(b \mid s)$	$=$	0.6	S:	$p(s)$	$=$	0.5
	$p(b \mid \bar{s})$	$=$	0.3				
D:	$p(d \mid b, e)$	$=$	0.9	T:	$p(t \mid a)$	$=$	0.05
	$p(d \mid \bar{b}, e)$	$=$	0.7		$p(t \mid \bar{a})$	$=$	0.01
	$p(d \mid b, \bar{e})$	$=$	0.8				
	$p(d \mid \bar{b}, \bar{e})$	$=$	0.1				
E:	$p(e \mid l, t)$	$=$	1	X:	$p(x \mid e)$	$=$	0.98
	$p(e \mid \bar{l}, t)$	$=$	1		$p(x \mid \bar{e})$	$=$	0.05
	$p(e \mid l, \bar{t})$	$=$	1				
	$p(e \mid \bar{l}, \bar{t})$	$=$	0				

2.10 A broader context for probabilistic expert systems

We have informally introduced the idea of representing qualitative relationships between variables by graphs and superimposing a joint probability model on the unknown quantities. When the graph is directed and does not contain any (directed) cycles, the resulting system is often called a *Bayesian network*, although later we shall see how broader classes of graphs may be used. Using the terms introduced earlier, we may think of this network and its numerical inputs as forming the knowledge base, while efficient methods of implementing Bayes' theorem form the inference engine used to draw conclusions on the basis of possibly fragmentary evidence.

While Bayesian networks have now become a standard tool in artificial intelligence, it is important to place them in a wider context of what might be called *highly structured stochastic systems* (HSSS). This broad term attempts to bring together areas in which complex interrelationships can be expressed by local dependencies, and hence a graphical representation can be exploited both to help communication and as a basis for computational algorithms. We are led naturally to a unifying system of Bayesian reasoning on graphical structures: by embedding apparently unrelated topics in this common framework, strong similarities are revealed which can lead to valuable cross-fertilization of ideas. Further information can be found on the HSSS Web page (see Appendix C).

A natural example area is genetics, in which familial relationships form the basis for the graph and Mendelian laws of inheritance and relationships between genotype and phenotype provide the elements of the probability distribution. The 'peeling' algorithm for pedigree analysis derived by Cannings et al. (1978) was shown by Spiegelhalter (1990) to be very similar to the local computation algorithm of Lauritzen and Spiegelhalter (1988). Similarly, much of image analysis is dominated by Markov field models which are defined in terms of local dependencies and can be described graphically (Besag and Green 1993), although simulation methods are generally required for inference. Such spatial models are also used in geographical epidemiology (Bernardinelli et al. 1997) and agricultural field trials (Besag et al. 1995). Within the artificial intelligence community, neural networks are natural candidates for interpretation as probabilistic graphical models, and are increasingly being analysed within a Bayesian statistical framework (see, for example, Neal (1996)). Hidden Markov models, which form the basis for work in such diverse areas as speech recognition (Rabiner and Juang 1993) and gene sequencing (Durbin et al. 1998), can likewise be considered as special cases of Bayesian networks (Smyth et al. 1997).

Further reading

Recent years have seen an explosion of interest in graphical models as a basis for probabilistic expert systems, with a number of dedicated books and a wide range of theoretical and practical publications. Textbooks on probabilistic expert systems include the classic Pearl (1988). Neapolitan (1990) explains the basic propagation algorithms, and these are studied in detail by Shafer (1996). Jensen (1996) is a very good tutorial introduction, while Castillo et al. (1997) provides another sound introduction with many worked examples.

Perhaps the best guide to current research is to be found in the *Proceedings* of the annual meeting of the Association for Uncertainty in Artificial Intelligence, which hosts an excellent Web page providing many relevant links, and provides a forum for discussion of a wide range of issues concerning uncertainty in expert systems (although the arguments between the advocates of probabilistic and non-probabilistic approaches appear to have died down as each group tries to identify the most appropriate domains for its work). Other electronic sources of information include the Bayesian network Web page of the US Air Force Institute of Technology Artificial Intelligence Laboratory, which in particular features informal comments of people who work in industry, and the Web page of the Microsoft Decision Theory and Adaptive Systems group. See Appendix C for addresses and more details. Appendix C also details some free and commercial software available over the World Wide Web.

Other good sources of tutorial material are the special issues of *AI Magazine* (Charniak 1991; Henrion et al. 1991) and of the *Communications of the ACM* (Heckerman and Wellman 1995).

See also Section 3.5 for some pointers to various implementations and applications of probabilistic expert systems.