AN INTRODUCTION TO EXPERT SYSTEMS

by

Bryan S. Todd

Technical Monograph PRG-95
ISBN 0-902928-73-2

February 1992

Oxford University Computing Laboratory
Programming Research Group
11 Keble Road
Oxford OX1 3QD
England

Oxford University Computing Laboratory
Programming Research Group
11 Keble Road
Oxford OX1 3QD
England

Electronic mail: todd@comlab.ox.ac.uk

An Introduction to Expert Systems

Bryan S. Todd

Abstract

This monograph provides an introduction to the theory of expert systems. The task of medical diagnosis is used as a unifying theme throughout. A broad perspective is taken, ranging from the role of diagnostic programs to methods of evaluation. While much emphasis is placed on probability theory, other calculi of uncertainty are given due consideration.

Contents

1 Synopsis 1
- 1.1 Scope of Monograph . . . 1
- 1.2 Outline of Monograph . . . 2

2 Decision Support Systems 3
- 2.1 Purpose and Role . . . 3
  - 2.1.1 Checklists . . . 3
  - 2.1.2 Decision Aids . . . 3
- 2.2 Early Attempts . . . 4
  - 2.2.1 Flowcharts . . . 5
- 2.3 Observer Variation . . . 5
- 2.4 Statistical Methods . . . 6
  - 2.4.1 The Value of Raw Data . . . 6
  - 2.4.2 Probability Theory . . . 6
  - 2.4.3 Bayes' Theorem . . . 10

3 Data-Based Approaches 13
- 3.1 Validity of the Independence Assumption . . . 13
- 3.2 Avoiding the Independence Assumption . . . 14
  - 3.2.1 Lancaster Model . . . 14
  - 3.2.2 Clustering Methods . . . 14
  - 3.2.3 Kernel Method . . . 14
- 3.3 Nearest-Neighbours Method . . . 15
- 3.4 Logistic Model . . . 16
  - 3.4.1 The Spiegelhalter-Knill-Jones Method . . . 17
- 3.5 Recursive Partitioning . . . 19
- 3.6 Neural Networks . . . 22

4 Rule-Based Methods 24
- 4.1 Types of Knowledge . . . 24
- 4.2 Categorical Knowledge . . . 25

4.2.1 Knowledge Base 25
4.2.2 Inference Engine 26
4.3 MYCIN 30
4.3.1 Certainty Factors 31
4.3.2 Belief 32
4.3.3 Inference Strategy 33
4.3.4 EMYCIN 34
4.4 PROSPECTOR 34
4.4.1 Inference 34

5 Descriptive Methods **37**
5.1 INTERNIST 37
5.1.1 Knowledge Representation 37
5.1.2 Inference Algorithm 38
5.1.3 Performance 39
5.1.4 CADUCEUS 39
5.2 Discussion 39
5.2.1 Patient Specific Models 40

6 Causal Networks **41**
6.1 Combining Statistical and Knowledge-Based Methods 41
6.1.1 A Generalization 41
6.2 Causal Networks as a Representation 42
6.2.1 Simplification 43
6.2.2 An Example 43
6.2.3 Separation 46
6.2.4 Assumed Models 46
6.3 Inference 47
6.3.1 Inference in Causal Trees 47
6.3.2 Inference in Sparse Causal Graphs 50
6.3.3 Monte Carlo Inference Methods 54

7 A Probabilistic Rule-Based System **57**
7.1 A Causal Graph Representation 57
7.1.1 Car Faults Revisited 58
7.2 Assuming a Logistic Model 61
7.2.1 Allowing Expressions 62
7.2.2 Transforming the Weights 62
7.2.3 Decomposition into Rules 64
7.3 Inference 65
7.3.1 Monte Carlo Propagation 65

7.4 Inferential versus Causal Representations 67
7.4.1 Insufficiency of Causation 68
7.4.2 Scarcity of Training Data 68
7.4.3 Explanations 69

8 Alternative Calculi of Uncertainty **70**
8.1 Fuzzy Sets 70
8.1.1 Paradoxes of Gradual Change 70
8.1.2 A Representation for Fuzzy Sets 71
8.1.3 Operations on Fuzzy Sets 72
8.1.4 Linguistic Hedges 74
8.1.5 Fuzzy Inference 75
8.1.6 Production Rules 76
8.1.7 Fuzzy Inference and Medical Diagnosis 77
8.2 Dempster-Shafer Theory of Evidence 77
8.2.1 Some Difficulties with Probability Theory 77
8.2.2 Mass Functions 78
8.2.3 Dempster's Rule of Combination 79

9 Testing and Evaluation of Decision Aids **81**
9.1 Evaluation 81
9.1.1 Test Data 81
9.1.2 Trial Design 82
9.2 Performance Parameters 83
9.2.1 Diagnostic Accuracy 83
9.2.2 ROC Curves 84
9.2.3 Discriminant Matrices 86

Chapter 1

Synopsis

1.1 Scope of Monograph

What is an *expert system*? Opinions differ, and definitions vary from functional requirements, which may be undemanding

> *a program intended to make reasoned judgements or give assistance in a complex area in which human skills are fallible or scarce [Lau88]*

or exacting

> *a program designed to solve problems at a level comparable to that of a human expert in a given domain [Coo89],*

to more operational descriptions, usually in terms of 'knowledge' and 'inference':

> *a computer system that operates by applying an inference mechanism to a body of specialist expertise represented in the form of 'knowledge' [Goo85].*

The scope of this monograph is not restricted to any specific kind of implementation method, such as that embodied by the last of the three definitions above. Instead, a broader view is taken. Other kinds of system meeting the first definition are included for comparison.

Application to medical diagnosis is used as a recurring theme throughout. This is one of the most intensive fields of expert system research, and it provides a unifying context for discussing the merits of different approaches. The arguments are, however, transferable to other domains, and other applications are also described and used as examples where relevant.

1.2 Outline of Monograph

Chapter 2 discusses the possible roles of medical expert systems, and briefly reviews some early methods for providing decision support. These include one of the most successful: the use of Bayes' theorem with the assumption of conditional independence.

Chapter 3 reviews a variety of alternative statistical methods which in one way or another avoid some of the disadvantages associated with the simpler use of Bayes' theorem.

Chapter 4 introduces rule-based methods by illustrating some of the components of a categorical expert system, by means of a simple example in Prolog. Two well-known systems, MYCIN and PROSPECTOR, which reason under uncertainty, are then described.

Chapter 5 explains an alternative knowledge representation: the descriptive paradigm. This is exemplified by two large medical expert systems, INTERNIST and its successor CADUCEUS.

Chapter 6 introduces causal networks as a descriptive knowledge representation based soundly on probability theory. Considerable emphasis is given to the theory of causal networks. This is because they appear to be emerging as one of the most important methods for constructing expert systems which reason under uncertainty.

Chapter 7 counters the claim that inference rules are unsuitable as a knowledge representation when uncertainty is involved. A rule-based representation is derived, employing a model first introduced in Chapter 3: the logistic form.

Chapter 8 describes two alternative formalisms for handling uncertainty. The motivation for seeking new techniques is explained, and the methods are contrasted with probability theory.

Chapter 9 discusses both how to evaluate a diagnostic expert system, and how to present the results in a clear and comprehensive way.

Chapter 2

Decision Support Systems

2.1 Purpose and Role

Consider the problem of medical diagnosis. How might a computer program assist a doctor to interpret his clinical findings and make a correct diagnosis? There are two, quite different ways, and it is possible for a computer program to help to some extent in both.

2.1.1 Checklists

Firstly, from time to time a particular kind of diagnostic challenge is encountered, with the following characteristics.

1. All the information necessary to reach the correct diagnosis has been gathered.
2. It is hard, however, to think of the correct diagnosis.
3. Once suggested, though, the correct diagnosis is easily verified.

A loose analogy can be drawn with solving a crossword clue. For this kind of problem, a computer program would be useful if it could suggest a sensible list of possible interpretations. The role of such a program ought to be uncontroversial because judgement and decision are left entirely to the clinician. The program can be regarded simply as an 'intelligent checklist' which prevents a possible oversight. However, while such problems are often thought to be quite common, they are actually extremely rare [Dom78].

2.1.2 Decision Aids

A more controversial role for a computer program is as a direct aid to deciding between a few possible alternatives, others having been ruled out.

It has been suggested that the results of a computer analysis can be regarded just like those of any other test which assist the doctor in making a decision [Dom84]. Indeed, computer analysis is an entirely non-invasive test carrying no direct risk to the patient, only the indirect risk that it may mislead the doctor. Moreover, if the program is carefully designed and implemented, it is inexpensive too!

However, there is a special distinction between analysing clinical findings by computer and carrying out a blood test or an X-ray; no *new* diagnostic evidence is obtained. The computer simply analyses the clinician's own findings. Furthermore, the facts entered into the computer are an abstraction of those findings, so some of the information available to the clinician is inevitably lost in the process. (Can you think of a practical way of estimating how much is lost?)

Despite these constraints, programs can be developed which, in trials, appear useful. One approach entails trying to formalize a specialist's own knowledge and to simulate his reasoning processes; the program may then assist non-experts ('dissemination of expertise'). A recent example of such a program in a medical domain is the PLEXUS system for advice on the diagnosis and management of nerve injuries [Jas87]. We will examine others in more detail later.

If, though, the intention is to assist the specialist himself, then the program must incorporate 'knowledge' he does not possess, and (if possible) use it in a more effective way. Surprisingly, quite simple techniques go some considerable way to attaining this objective, although no systems yet exist which have been shown to be of unequivocal use to a medical specialist.

2.2 Early Attempts

Before computers became widely available, efforts were made to provide diagnostic assistance using mechanical devices. Nash designed a wooden frame down the side of which were marked some 300 diseases [Nas54]. Wooden strips, one for each symptom the patient had, could be hung on the frame. Each strip was marked across with lines corresponding in position to the diseases which could explain the symptom. Diseases which could explain all the patient's symptoms were then easily read off the frame; they were against continuous lines running across all the strips. Lipkin and Hardy describe a similar method for the identification of 26 blood disorders, using punched cards [Lip58]. They tested their system using the case records of 80 patients who had been previously diagnosed. In 73 of these cases, only one disease explained all the findings, and this was invariably the correct

diagnosis. In the remaining 7 cases, the system failed because each patient had multiple disorders, and no single disease could explain everything.

The strength of these systems is their simplicity; it is transparently obvious to the user how the results are obtained and what they mean. Furthermore, the inherent limitations of mechanical devices are readily overcome by implementing the methods as computer programs instead. For example, it would then be easy to look for all *pairs* of diseases which explain the findings. A system in current use based on these principles assists in the diagnosis of rare malformation syndromes [Win84].

Exercise 2.1 *Choose some diagnostic task with which you are familiar (for example, working out why a car won't start). Design and implement in your preferred programming language, a system based on the principle of Nash's apparatus to help localize the cause.*

2.2.1 Flowcharts

Once computers became readily accessible, a favoured method of encoding medical reasoning was by means of flowcharts using branch chain logic (so-called 'clinical algorithms'). Flowcharts can be useful because they make lines of reasoning explicit, so errors and omissions can be more readily identified than with some more complicated techniques. Quite complex diagnostic procedures can he formalized in this way, and explanations can be assembled during program execution from fragments of prose attached to arcs in the diagram; see for example a program to interpret biochemical abnormalities [Ble72]. Other medical applications include the diagnosis of dysphagia [Edw70], and screening for neurological disease [Vas73].

Exercise 2.2 *Repeat Exercise 2.1 using a flowchart representation instead. Which method is easier, and why?*

2.3 Observer Variation

The diagnostic value of any computer analysis is ultimately limited by the reliability of the clinical information entered about a given patient, and this principle applies equally to non-medical applications. How reliable then are clinical findings? In 1973 Gill and co-workers reported the results of a study of observer variation amongst clinicians [Gil73]. Three clinicians attended patient interviews conducted by a fourth. They recorded which questions were asked, and whether the symptoms were present or not. Surprisingly, the three observers disagreed in 20% of instances as to whether a particular

question was actually asked, and in 16% of instances as to whether the patient's response was positive or negative!

This high degree of variation was attributed to a lack of standard definitions of symptoms. When agreed formal definitions were introduced, and the experiment repeated, disagreement occurred in only about 4% of instances [Gil73]. Further evidence of this wide divergence of opinion regarding the definition of common symptoms is provided by a study of 40 experienced gastroenterologists and surgeons [Kni85]. Clearly, any proposed development of an expert system to assist diagnosis should be preceded wherever possible by agreeing standard definitions of findings. This may prove to have a greater effect on the final performance than any particular choice of implementation method.

2.4 Statistical Methods

In general, what sources of medical 'knowledge' are available for constructing an expert system? There are of course journal articles, textbooks and medical specialists themselves. There is, however, another important source of information: databases of previously diagnosed cases, particularly when compiled using agreed formal definitions of symptoms and signs.

2.4.1 The Value of Raw Data

In an interesting study [Kni85], four gastroenterologists were asked independently to specify which symptoms might discriminate between duodenal and gastric ulcers. When compared with a database of several hundred actual cases, only four of the twelve most trusted symptoms were subsequently found to be significantly discriminating, one of which discriminated in the reverse direction to that expected. This demonstrates the potential diagnostic value of databases, and to some extent casts doubt on 'expert opinion' as a primary source of knowledge for diagnostic programs.

2.4.2 Probability Theory

In order to draw from previous cases, possibly uncertain inferences regarding a new case, we require a calculus of uncertainty. Although there exist several such calculi pertinent to expert systems (two modern alternatives are described in Chapter 8), probability theory is the most firmly established. The following is a brief summary of the basics of discrete probability theory. A more complete account can be found in almost any standard text (for example, [Nea89]).

Definitions and Axioms

Consider an experiment whose set Ω of possible outcomes is known in advance. The set Ω is known as the *sample space* of the experiment, and each element of Ω is known as a *sample point.* (For simplicity we will assume that Ω is fiuite.) Thus if the experiment consists of rolling a die, then $\Omega = \{1,2,3,4,5,6\}$.

Any subset of Ω is referred to as an *event.* (We will denote events by upper-case letters.) An event E is said to occur precisely when the outcome of the experiment lies in E. For example regarding dice, $\{2,4,6\}$ is tbe event 'an even number is thrown', and $\{1,2,3\}$ is the event 'a number less than four is thrown'. The entire sample space Ω denotes the *certain event,* and the empty set $\{\}$ denotes the *impossible event.*

The probability of an event E is a real number denoted $p(\mathrm{E})$, and every probability function p satisfies three axioms.

Axiom 1 *Probabilities are non-negative.*

$$0 \leq p(\mathrm{E})$$

Axiom 2 *The probability of the certain event is one.*

$$p(\Omega) = 1$$

Axiom 3 *If two events (*E *and* F*) are* mutually exclusive *(disjoint) then the probability that at least one of them occurs is the sum of their respective probabilities.*

$$\mathrm{E} \cap \mathrm{F} = \{\} \Rightarrow p(\mathrm{E} \cup \mathrm{F}) = p(\mathrm{E}) + p(\mathrm{F})$$

The Complement of an Event

The *complement* (or negation) of an event E is written $\overline{\mathrm{E}}$. By definition, $\overline{\mathrm{E}}$ occurs if and only if E does not occur.

$$\overline{\mathrm{E}} \triangleq \Omega - \mathrm{E} \tag{2.1}$$

Consequently,

$$p(\overline{\mathrm{E}}) = 1 - p(\mathrm{E}) \tag{2.2}$$

Joint Probabilities and Conditional Probabilities

The probability $p(\mathrm{E} \cap \mathrm{F})$ that both event E and event F occur is termed the *joint probability* of E and F. By convention, a comma is used to denote intersection of events; given any two events E and F,

$$p(\mathrm{E}, \mathrm{F}) \;\widehat{=}\; p(\mathrm{E} \cap \mathrm{F}) \tag{2.3}$$

The *conditional probability* of E given F is denoted $p(\mathrm{E} \mid \mathrm{F})$. When $p(\mathrm{F})$ is non-zero, $p(\mathrm{E} \mid \mathrm{F})$ is defined to be the ratio of the joint probability to the probability of F.

$$p(\mathrm{E} \mid \mathrm{F}) \;\widehat{=}\; \frac{p(\mathrm{E}, \mathrm{F})}{p(\mathrm{F})} \tag{2.4}$$

When $p(\mathrm{F})$ is zero, $p(\mathrm{E} \mid \mathrm{F})$ is undefined.

Continuing with the example of a die, let E be the event 'an even number is thrown' and let F be the event 'a number less than four is thrown'. The probability of any event is given by the sum of the probabilities associated with its constituent sample points (from Axiom 3). We assume that the die is unbiased, so the probability associated with each sample point is the same (1/6). Thus

$$\begin{array}{lcllcl} \mathrm{E} & = & \{2,4,6\} & \text{and} & p(\mathrm{E}) & = & 1/2 \\ \mathrm{F} & = & \{1,2,3\} & \text{and} & p(\mathrm{F}) & = & 1/2 \\ \mathrm{E} \cap \mathrm{F} & = & \{2\} & \text{and} & p(\mathrm{E}, \mathrm{F}) & = & 1/6 \end{array}$$

Therefore, the conditional probability that an even number has been thrown, given that the number is less than four, is 1/3 (i.e. 1/6 divided by 1/2).

Random Variables

A *random variable* is a function from Ω to the reals $\mathbf{R}$. We will use lower-case Greek letters to denote random variables. In this course we will consider only the boolean variety ($\Omega \rightarrow \{0,1\}$) which we will call *propositional variables.*

By convention, the event that a random variable α takes value a, is denoted by '$\alpha = a$'. Thus, given any propositional variable $\alpha : \Omega \rightarrow \{0,1\}$ and value $a : \{0,1\}$,

$$\alpha = a \quad \widehat{=} \quad \{s : \Omega \mid \alpha(s) = a\} \tag{2.5}$$

We will denote sets of propositional variables by the letters $\mathcal{A}, \mathcal{B}, \ldots, \mathcal{Z}$. Given any set $\mathcal{A}$ of propositional variables ($\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$) and corresponding sequence a of values ($a = [a_1, a_2, \ldots, a_n]$), by convention,

$$\mathcal{A} = a \quad \widehat{=} \quad \bigcap_{1 \leq i \leq n} (\alpha_i = a_i) \tag{2.6}$$

In order to reduce the notational burden, a propositional variable (or set of propositional variables) will often appear in a formula without reference to a particular value. In such cases, there is an implicit universal quantification over all possible values. For example,

$$p(\alpha, \beta) = p(\alpha)p(\beta)$$

is short for

$$\forall a, b : \{0, 1\} \bullet p(\alpha = a, \beta = b) = p(\alpha = a)p(\beta = b)$$

Furthermore, the event that a propositional variable takes value 1 will often be abbreviated to the corresponding upper-case letter. Thus

$$\begin{array}{lll} \alpha = 1 & \text{becomes} & \mathrm{A} \\ \beta = 1 & \text{becomes} & \mathrm{B} \end{array}$$

and so forth. Similarly,

$$\begin{array}{lll} \alpha = 0 & \text{becomes} & \overline{\mathrm{A}} \\ \beta = 0 & \text{becomes} & \overline{\mathrm{B}} \end{array}$$

etc.

Independence

Two events E and F are said to be *independent* exactly when the probability $p(\mathrm{E}, \mathrm{F})$ of the joint event is equal to the product of the individual probabilities, $p(\mathrm{E})$ and $p(\mathrm{F})$. Clearly, independence is a symmetric relationship. Furthermore, it follows that if E and F are independent then, whenever $p(\mathrm{E} \mid \mathrm{F})$ is defined, $p(\mathrm{E})$ is equal to $p(\mathrm{E} \mid \mathrm{F})$. Thus knowledge that event F has occurred does not influence the likelibood of E occurring.

Similarly, two propositional variables α and β are said to be (unconditionally) independent precisely when

$$p(\alpha, \beta) = p(\alpha)p(\beta) \tag{2.7}$$

and *conditionally independent* given a set of variables $\mathcal{C}$ precisely when

$$p(\alpha, \beta \mid \mathcal{C}) = p(\alpha \mid \mathcal{C})p(\beta \mid \mathcal{C}) \tag{2.8}$$

Application to Medical Diagnosis

In the context of medical diagnosis, Ω is some real or imagined population (for example, the set of all patients who have been, or ever will be, referred to the John Radcliffe Hospital). Now suppose δ represents some arbitrary disease: formally, $\delta = 1$ (abbreviated to D) is the set of all patients who have disease δ. Furthermore, let $\mathcal{S}$ $(= \{\sigma_1, \sigma_2, \ldots, \sigma_n\})$ be a set of propositional variables corresponding to possible symptoms, signs or other items of diagnostic value. Thus, if say σ_3 is 'raised temperature' then $\sigma_3 = 0$ is the event 'the patient does *not* have a raised temperature', and $\sigma_3 = 1$ is the event 'the patient *does* have a raised temperature'.

Suppose a patient is drawn randomly from the same population. The actual symptom values we record are s $(= [s_1, s_2, \ldots, s_n])$, and we wish to predict whether he or she has disease δ. We are therefore interested in $p(\mathrm{D} \mid \mathcal{S} = s)$, the conditional probability that our patient has disease δ.

Unfortunately, in practice any attempt to estimate $p(\mathrm{D} \mid \mathcal{S} = s)$ directly from a random sample of previously diagnosed patients will almost certainly fail because it is unlikely that the sample will include any cases with exactly the findings s. One solution, however, is to make some modelling assumptions; Bayes' theorem allows this.

2.4.3 Bayes' Theorem

Two applications of the definition of conditional probability (Equation 2.4) leads to

$$p(\mathrm{D} \mid \mathcal{S}) = \frac{p(\mathcal{S} \mid \mathrm{D})p(\mathrm{D})}{p(\mathcal{S})} \tag{2.9}$$

Unless disease δ is very rare, it is generally feasible to estimate $p(\mathrm{D})$ directly as the relative frequency with which $\delta = 1$ in a random sample (e.g. a database of several hundred cases). One solution to the problem of estimating $p(\mathcal{S} \mid \mathrm{D})$ is to assume that the individual symptoms are conditionally independent given the presence of disease δ. Thus,

$$p(\mathcal{S} \mid \mathrm{D}) = \prod_{1 \leq i \leq n} p(\sigma_i \mid \mathrm{D}) \tag{2.10}$$

Direct estimation of the conditional probability $p(\sigma_i \mid \mathrm{D})$ is usually feasible.

The denominator $p(\mathcal{S})$ of Equation 2.9 is also problematic. The usual procedure is to assume that all diseases $(\delta_1, \delta_2, \ldots, \delta_m)$ are mutually exclusive (each patient has exactly one such disease δ_j). It then follows from Axiom 3 (Page 7) and the definition of conditional probability (Equation 2.4)

that

$$p(\mathcal{S}) = \sum_{1 \le j \le m} p(\mathcal{S} \mid \mathrm{D}_j) p(\mathrm{D}_j) \tag{2.11}$$

(The numerator in Equation 2.9 is one of the terms in the sum; the others are evaluated similarly.)

Exercise 2.3 *As an alternative to Equation 2.11 with its implicit assumption that every patient has exactly one disease, we could assume instead that findings are* unconditionally *independent as well. Thus we could write*

$$p(\mathcal{S}) = \prod_{1 \le i \le n} p(\sigma_i)$$

Suppose two symptoms (σ_1 and σ_2) are recorded from 1000 patients each of whom has one of two possible diseases (δ_1 and δ_2).

δ_1	δ_2	σ_1	σ_2	*Cases*
0	*1*	*0*	*0*	*730*
0	*1*	*0*	*1*	*20*
0	*1*	*1*	*0*	*20*
0	*1*	*1*	*1*	*30*
1	*0*	*0*	*0*	*20*
1	*0*	*0*	*1*	*80*
1	*0*	*1*	*0*	*80*
1	*0*	*1*	*1*	*20*
				1000

Calculate $p(\mathrm{D}_1 \mid \mathrm{S}_1, \mathrm{S}_2)$ using Equation 2.9. Obtain the numerator by assuming conditional independence and applying Equation 2.10. Obtain the denominator by assuming unconditional independence and applying the formula suggested above. What is the meaning of the result?

An Application of Bayes' Theorem: The Leeds Program

One of the most successful medical applications of Bayes' theorem has been to the diagnosis of acute abdominal pain. De Dombal and co-workers in Leeds noted that 95% of patients presenting to hospital with abdominal pain of recent onset fall into exactly one of seven diagnostic categories [Dom72].

1. Appendicitis
2. Diverticular disease
3. Perforated duodenal ulcer

4. Non-specific abdominal pain
5. Cholecystitis
6. Small bowel obstruction
7. Pancreatitis

Using data from 400 patients, conditional probabilities for each possible clinical finding, given each of the seven diagnostic categories, were estimated. Bayes' theorem was used to classify 304 new cases; the computer diagnosis was taken to be the disease δ_j with highest $p(\mathrm{D}_j \mid \mathcal{S})$, where $\mathcal{S}$ stands for the registrar's findings at his first contact with the patient. The computer achieved a correct diagnosis rate of 91.8% compared to 79.6% for the most senior clinician who saw the case.

This very high computer accuracy has not been sustained in subsequent trials, however, and doubts are now being expressed about the true value of this method [Sut89].

Exercise 2.4 *In 43% of cases referred to hospital with acute abdominal pain, the pain resolves spontaneously and no specific cause is found ('non-specific abdominal pain'). Another 24% of cases turn out to have appendicitis. In 74% of cases of appendicitis, the pain is in the right lower quadrant, whereas in only 29% of cases of non-specific pain is this the site. What is the relative likelihood of appendicitis as opposed to non-specific abdominal pain if the site is the right lower quadrant? (Published data [Dom80])*

Exercise 2.5 *Continuing Exercise 2.4, in 57% of cases of appendicitis, the pain is aggravated by movement, but this is true in only 9% of cases of non-specific pain. Assuming that the site of the pain is conditionally independent of aggravation of the pain by movement, both in the presence of acute appendicitis and when the pain has no specific cause, what is the relative likelihood of appendicitis if we also learn that the pain is* not *aggravated by movement?*

Chapter 3

Data-Based Approaches

3.1 Validity of the Independence Assumption

The most common criticism of the use of Bayes' theorem as described in Chapter 2 is the assumption of conditional independence. In practice, many symptoms and signs are correlated (for example, pulse rate and temperature). Several studies (for example [Fry78, Cha89]) have assessed the importance of the independence assumption with respect to medical data; a small but significant reduction of diagnostic accuracy was generally found.

To see the effect of ignoring interactions, consider the following hypothetical example (taken from [Nor75a]) of the joint distributions of two symptoms (σ_1 and σ_2) given the presence of each of two diseases (δ_1 and δ_2).

$$\begin{array}{lcllcl} p(S_1, S_2 \mid D_1) & = & 0.5 & p(S_1, S_2 \mid D_2) & = & 0 \\ p(S_1, \overline{S_2} \mid D_1) & = & 0 & p(S_1, \overline{S_2} \mid D_2) & = & 0.5 \\ p(\overline{S_1}, S_2 \mid D_1) & = & 0 & p(\overline{S_1}, S_2 \mid D_2) & = & 0.5 \\ p(\overline{S_1}, \overline{S_2} \mid D_1) & = & 0.5 & p(\overline{S_1}, \overline{S_2} \mid D_2) & = & 0 \end{array}$$

The conditional probabilities of each symptom are the same given each disease, since

$$p(S_1 \mid D_1) = p(S_1 \mid D_2) = 0.5$$

and

$$p(S_2 \mid D_1) = p(S_2 \mid D_2) = 0.5$$

So taken alone, each symptom provides no discriminatory power between the diseases. Yet, considered in combination, the two symptoms enable perfect discrimination.

This chapter describes a variety of approaches which make weaker assumptions than does the simpler application of Bayes' theorem.

3.2 Avoiding the Independence Assumption

3.2.1 Lancaster Model

Lancaster has generalized the definition of independence between variables to one of independence between sets of variables [Zen75]. This enables the following alternative to Equation 2.10 (Page 10); Equation 3.1 takes into account pairwise interactions between symptoms, but assumes that no higher order interactions occur.

$$p(\mathcal{S} \mid \mathrm{D}) = \left(\sum_{1 \le i < j \le n} p(\sigma_i, \sigma_j \mid \mathrm{D}) \prod_{k \neq i,j} p(\sigma_k \mid \mathrm{D}) \right) - (C_2^n - 1) \prod_{1 \le k \le n} p(\sigma_k \mid \mathrm{D}) \tag{3.1}$$

Notice, however, that the number of parameters to estimate is now quadratic rather than linear with respect to the number of symptoms. In most applications, this requires a large amount of training data.

The effect of weakening the independence assumption in this way was assessed with respect to the diagnosis of acute abdominal pain using 5916 training cases [Ser86]. A small improvement in diagnostic accuracy was found.

3.2.2 Clustering Methods

The principal interactions that do occur are probably between small clusters of symptoms which share a common cause. Norussis and Jacquez have suggested identifying these clusters by analyzing correlation coefficients, and then regarding each such group of variables as a single, multi-valued variable [Nor75b].

3.2.3 Kernel Method

If sufficient training data were available, the conditional probability $p(\mathcal{S} \mid \mathrm{D})$ could be estimated directly, and no independence assumption would be necessary. One way to compensate for a shortage of training data is to 'blur' the cases that are available; each case is replaced by a collection of similar cases. This is the basis of the 'kernel' method of smoothing [Ait76]. It offers another alternative to Equation 2.10.

$$p(\mathcal{S} \mid \mathrm{D}) = \frac{1}{T} \sum_{1 \le t \le T} \lambda_\delta^{n - s_t} (1 - \lambda_\delta)^{s_t} \tag{3.2}$$

where

T = Total number of training cases.

λ_δ = Smoothing parameter for disease δ. $(0.5 \leq \lambda_\delta \leq 1)$

s_t = Hamming distance (number of differing values) between the instantiation of $\mathcal{S}$ and the corresponding findings of the t^{th} training case.

The success of this method depends on the choice of the smoothing parameter λ_δ. Several optimization methods have been described [Ait76, Tit80, Tut86].

3.3 Nearest-Neighbours Method

Actually, if sufficient training data really were available, then Equation 2.9 (Page 10) would be irrelevant; $p(\mathrm{D} \mid \mathcal{S} = s)$ itself could be estimated directly as the relative frequency with which $\delta = 1$ amongst cases which have exactly the clinical findings s. This is defeated in practice, however, because it is very unusual to find in the training set even a single exact match (identical symptom values) to a new patient.

A simple relaxation of this is to define a metric on vectors of findings, and identify (for some pre-set value k, such as $k = 10$) the k cases in the training set which are closest to the new patient. The conditional probability $p(\mathrm{D} \mid \mathcal{S})$ is then estimated as the relative frequency of disease δ amongst this set of partial matches. The simplest metric to use is the Hamming distance. However, greater diagnostic accuracy may be achieved if each of the symptoms is assigned a positive weight, and the distance defined as the sum of the weights of the symptoms whose values differ. Notice that application of this method entails no assumption of mutual exclusion between diseases; multiple disorders can be detected.

It has been proposed to implement this method on a connectionist architecture in which the task of storing a very large training set and retrieving close matches to new cases is distributed over a large number of processors [Sta86]. However, when the nearest-neighbours method was applied to the diagnosis of acute abdominal pain (5916 training cases and 1000 test cases), results were markedly inferior to those obtained simply from applying Bayes' theorem with the assumption of conditional iudependence [Ser85]. More encouraging results were obtained in a similar comparative study of the methods for the diagnosis of liver disorders (1991 training cases and

437 test cases), but Bayes' theorem was still marginally better [Cro72]. In conclusion, it seems that the nearest-neighbours method is not effective unless a very large amount of training data is available, and this is generally impracticable.

3.4 Logistic Model

For any events E and F, the *odds* are defined by

$$\text{odds (E)} \;\widehat{=}\; \frac{p(\text{E})}{p(\overline{\text{E}})} \tag{3.3}$$

and the *conditional odds* are defined by

$$\text{odds (E} \mid \text{F)} \;\widehat{=}\; \frac{p(\text{E} \mid \text{F})}{p(\overline{\text{E}} \mid \text{F})} \tag{3.4}$$

Notice that the corresponding probabilities are easily recovered.

$$p(\text{E}) \;=\; \frac{\text{odds (E)}}{1 + \text{odds (E)}} \tag{3.5}$$

$$p(\text{E} \mid \text{F}) \;=\; \frac{\text{odds (E} \mid \text{F)}}{1 + \text{odds (E} \mid \text{F)}} \tag{3.6}$$

The logistic approach to discrimination assumes a linear form for the log-odds [And82]. Thus if a is a sequence of real-valued coefficients ($a = [a_0, a_1, \ldots, a_n]$),

$$\ln \text{odds (D} \mid \mathcal{S} = s) = a_0 + \sum_{1 \le i \le n} a_i s_i \tag{3.7}$$

The coefficients $a_0, \ldots, a_n$ are chosen to maximize the probability of correct classification of the training cases. This entails iterative optimization.

Equation 3.7 is consistent with several families of distribution, including that in which symptoms are either conditionally independent or mutually exclusive given D and, conversely, given $\overline{\text{D}}$. It is also consistent with log-linear distributions in which the interaction terms are equal. Therefore, the logistic model is more general than independence Bayes, and this is usually reflected by higher diagnostic accuracy.

3.4.1 The Spiegelhalter-Knill-Jones Method

Indeed, whatever the underlying distribution, the conditional log-odds for a disease can be expressed as the sum of the 'weights of evidence' provided by the findings.

$$\ln\text{odds}\ (\mathrm{D} \mid \mathcal{S}) = \sum_{0 \le i \le n} w_i \tag{3.8}$$

The term w_0 stands for the prior weight of evidence before any of the findings are considered. It is simply the prior log-odds.

$$w_0 \ \hat{=}\ \ln\text{odds}\ (\mathrm{D}) \tag{3.9}$$

Each of the other terms represents the weight of evidence provided by the corresponding finding.

$$(i \neq 0), \qquad w_i \ \hat{=}\ \ln\left(\frac{p(\sigma_i \mid \sigma_1, \sigma_2, \ldots, \sigma_{i-1}, \mathrm{D})}{p(\sigma_i \mid \sigma_1, \sigma_2, \ldots, \sigma_{i-1}, \overline{\mathrm{D}})}\right) \tag{3.10}$$

Notice that the value of weight w_i depends on the values of all symptoms $\sigma_1 \ldots \sigma_i$. So w_i is really a family of 2^i terms, one for each possible assignment of symptom values. Therefore the number of parameters to estimate from training data is infeasibly large, in general.

One solution is to assume that symptoms are conditionally independent given D and, conversely, given $\overline{\mathrm{D}}$. Equation 3.10 then simplifies to Equation 3.11. Now only two parameters are required for each symptom σ_i: the weight of evidence provided by $\sigma_i = 0$ and the weight of evidence provided by $\sigma_i = 1$.

$$(i \neq 0), \qquad w_i \ \hat{=}\ \ln\left(\frac{p(\sigma_i \mid \mathrm{D})}{p(\sigma_i \mid \overline{\mathrm{D}})}\right) \tag{3.11}$$

We refer to these weights (Equation 3.11) as 'simple weights of evidence', because they rely upon a naive assumption of independence. If symptoms are in fact associated statistically, then the procedure implied by Equation 3.8 tends to count their evidence twice. To compensate for this, Spiegelhalter and Knill-Jones [Spi84] introduce 'shrinkage coefficients'.

$$\ln\text{odds}\ (\mathrm{D} \mid \mathcal{S}) = \sum_{0 \le i \le n} a_i w_i \tag{3.12}$$

Thus, a logistic relationship is assumed between $p(\mathrm{D} \mid \mathcal{S})$ and the weights of evidence w. The coefficients $a_0, \ldots, a_n$ are optimized iteratively over the same training data used to determine w.

Exercise 3.1 *Derive Equations 3.8, 3.9, 3.10 and 3.11 from first principles. Hence justify the assertion that the logistic form (Equation 3.7) is consistent with distributions in which symptoms are conditionally independent in the presence of the disease and in the absence of the disease.*

The Glasgow Dyspepsia System

This method was first applied to the diagnosis of dyspepsia (abdominal discomfort) [Spi84]. About 150 symptoms were recorded in 1200 patients referred to a specialist gastrointestinal clinic with dyspepsia. From this data simple weights of evidence for each of 7 diagnostic categories were obtained, and then shrinkage coefficients were derived. Multiplication of a simple weight of evidence by its shrinkage coefficient gives the actual weight.

For example, tabulated below are some weights of evidence for the diagnosis of gallstones.

Finding		Simple Weight	Actual Weight
Starting score (w_0)		−2.97	−3.00
History ≤ 12 months	No	−0.52	−0.44
	Yes	+0.56	+0.52
Attacks of pain	No	−1.75	−1.41
	Yes	+2.18	+1.77
Pain in RUQ	No	−0.88	−0.53
	Yes	+1.28	+0.77
Pain radiates to shoulder	No	−0.37	−0.19
	Yes	+2.53	+1.29

So for example, if a patient presents (−3.00) with a two-year history (−0.44) of attacks (+1.77) of pain in the right upper quadrant (+0.77) radiating to the shoulder (+1.29), then the total score is +0.39. So, the conditional log-odds are 0.39. Taking antilogs and applying Equation 3.6 (Page 16), we find that the probability that the patient has gallstones is 0.60.

$$\frac{e^{0.39}}{1+e^{0.39}} = 0.60$$

The strength of this method is that the user can clearly see which findings count for and which count against the final diagnosis, and to what extent. Furthermore, the method has an attractive simplicity. The entire table of weights, and a graph or reference table for performing the final transformation from score to probability, can be printed on a piece of card. The user can then calculate $p(\mathrm{D} \mid \mathcal{S})$ even without the aid of a computer.

More recently this method has been applied to the problem of predicting postoperative respiratory complications in elderly surgical patients [Sey90].

3.5 Recursive Partitioning

Rather than make the independence assumption that is implicit in Equation 3.11, it may be preferred to retain the generality of Equation 3.10. This is actually possible because, although the number of parameters to estimate is exponentially large, in practice the most reasonable estimate of nearly all of these is zero by default.

This is because in order to estimate w_i for some particular symptom values $(s_1, \ldots, s_i)$, sufficient training cases are required with precisely the findings $s_1, \ldots, s_{i-1}$, and these become rarer as i increases. If no such training cases are available, or if their number is too small to permit reliable estimation, then there is no alternative but to take w_i to be zero both for $\sigma_i = 0$ and for $\sigma_i = 1$. It follows that the number of weights that can actually be estimated cannot exceed the total number of training cases available.

The effect of each finding on the running total of evidence in favour of diagnosis δ can be expressed as a kind of flowchart (see Figure 3.1).

The accuracy of the probabilities depends critically on the order in which symptoms are considered. The worst decision would be to choose as σ_1 a symptom which is present in about half the training cases, but which provides hardly any evidence for or against the diagnosis of disease δ. Whatever the value of σ_1, only about half the training data would then be available to guide interpretation of subsequent findings.

When choosing the next symptom to consider, the objective should be to select one which partitions the training data into two sets of roughly similar size, but in which the relative frequency of disease δ is as different as possible. A measure advocated by Michie [Mic89] for this purpose is the expected magnitude of the weight of evidence that the finding will provide. In general, for symptom σ_i this is given by

$$E(w_i) = p_i \times |w_i^1| + (1 - p_i) \times |w_i^0| \tag{3.13}$$

where p_i is the probability that $\sigma_i = 1$ given the values of the preceding symptoms.

$$p_i \;\hat{=}\; p(\mathrm{S}_i \mid \sigma_1, \sigma_2, \ldots, \sigma_{i-1}) \tag{3.14}$$

and w_i^0 and w_i^1 are the weights of evidence provided by $\sigma_i = 0$ and $\sigma_i = 1$, respectively.

$$w_i^0 \;\hat{=}\; \ln\left(\frac{p(\overline{\mathrm{S}_i} \mid \sigma_1, \sigma_2, \ldots, \sigma_{i-1}, \mathrm{D})}{p(\overline{\mathrm{S}_i} \mid \sigma_1, \sigma_2, \ldots, \sigma_{i-1}, \overline{\mathrm{D}})}\right) \tag{3.15}$$

$$w_i^1 \;\hat{=}\; \ln\left(\frac{p(\mathrm{S}_i \mid \sigma_1, \sigma_2, \ldots, \sigma_{i-1}, \mathrm{D})}{p(\mathrm{S}_i \mid \sigma_1, \sigma_2, \ldots, \sigma_{i-1}, \overline{\mathrm{D}})}\right) \tag{3.16}$$

Overfitting is avoided by considering only symptoms which are significant according to Fisher's exact test, or χ^2 if numbers are large. If none of the remaining symptoms are significant then partitioning stops. The entire recursive procedure is as shown below, where T is the training set.

if T is 'partitionworthy' **then**

- Choose as the next σ_i the symptom with highest expected weight of evidence among the significant candidates.
- Partition T into T_0 (those for which $\sigma_i = 0$) and T_1 (those for which $\sigma_i = 1$).
- Apply this same procedure recursively to T_0 and T_1.

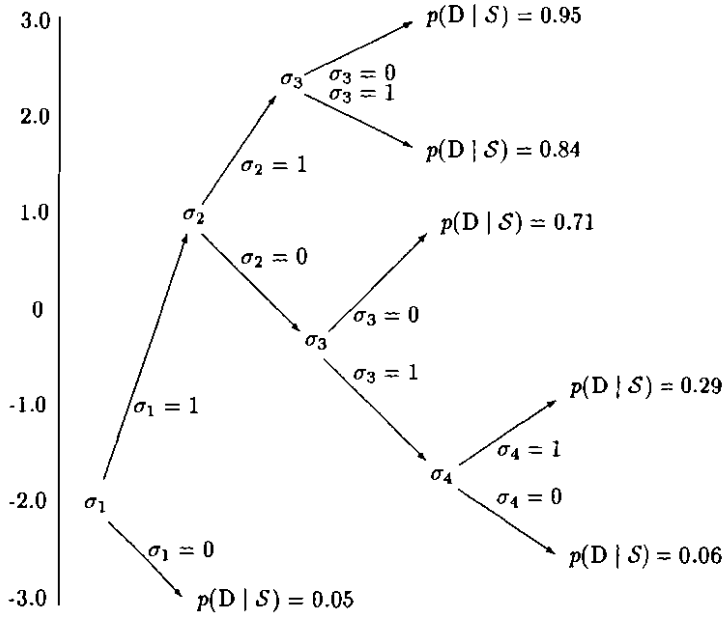
else label T with the log-likelihood ratio of disease δ estimated as

$$\ln\left(\frac{\text{Number of cases in } T \text{ for which } \delta = 1}{\text{Number of cases in } T \text{ for which } \delta = 0}\right)$$

A training set T is said to be *partitionworthy* exactly when

1. There is some symptom which does not appear anywhere on the path to T from the root, and
2. this symptom is present in at least one member of T and absent in at least one member of T, and
3. this symptom is a statistically significant discriminant for disease δ.

Figure 3.1: Example of a flowchart showing influence of each finding on the total evidence in favour of diagnosis δ, expressed on a vertical scale.



Notice that in Figure 3.1 each symptom is numbered according to its distance from the root, but different occurrences of σ_i need not stand for the same symptom.

Exercise 3.2 *Since recursive partitioning avoids making independence assumptions, is it necessarily a more accurate method than application of Bayes' theorem with the assumption of conditional independence? (Consider the case that findings* really *are conditionally independent. Consider also the example of extreme conditional* dependence *given in Section 3.1 (Page 13).)*

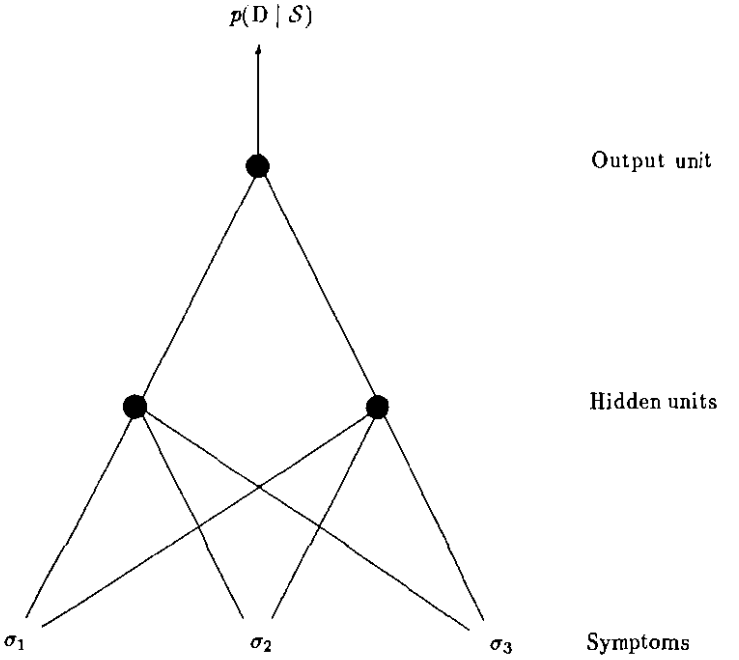
3.6 Neural Networks

Although motivated by the desire to model biological neural systems, the study of computational neural networks has led to more flexible discriminant functions, capable of computing more accurate conditional probabilities in the presence of interactions [Lip87]. One kind, a perceptron, is constructed from an ordered set of logistic functions (called either 'neurons' or simply 'units'). Each variable ('input') now consists of either a symptom σ_i or the output of another logistic function lower in the order. The final value computed by the top element ('output unit') can be regarded as an estimate of $p(\mathrm{D} \mid \mathcal{S})$. The lower elements are referred to as 'hidden units'; their purpose is to detect important, but unspecified features of the finding vector.

Figure 3.2 shows a three-unit perceptron. For simplicity, it has been assumed that there are only three symptoms: $\mathcal{S} = \{\sigma_1, \sigma_2, \sigma_3\}$.

A simple, iterative algorithm has become available in recent years for optimizing the coefficients over training data [Rum86]. It is a gradient descent method entailing the propagation of errors back from the output unit to those lower in the order.

Figure 3.2: Three-unit perceptron computing $p(D \mid \mathcal{S})$.



Chapter 4

Rule-Based Methods

4.1 Types of Knowledge

Although databases of previous cases provide a great deal of useful diagnostic information, other kinds of medical knowledge exist too. They include

- *Heuristic knowledge* - recognized associations between diseases and symptoms. For example,
  - Appendicitis nsually causes right lower quadrant abdominal pain.
  - Right lower quadrant pain suggests appendicitis.
- *Deep knowledge* - knowledge about underlying causal and anatomical mechanisms. For example,
  - The appendix usually lies in the right lower quadrant of the abdomen.
  - Inflammation of an ahdominal organ usually causes local pain.
- *Meta-knowledge* - knowledge about knowledge. This includes explicit awareness both of the reliability of particular knowledge, and of the strategy for using knowledge. For example,
  - Probabilities derived from statistical databases are more reliable than subjective estimates.
  - If a diagnosis cannot be reached directly by application of heuristic knowledge, then reason from first principles hy applying deep knowledge.

Knowledge of this kind can be gathered from textbooks and journals, and elicited from experts through interview. This knowledge is invaluable for the construction of expert systems when training data are scarce. Also, expert systems which use explicit knowledge of their domain to reason, have the potential to explain and justify their conclusions.

The ability of an expert system to explain its conclusions is often said to be an important prerequisite if it is to gain acceptance into routine use [Tea81, Fox83, San85]. (However, in a national trial of the Leeds system few doctors (under 10%) complained about the program's numerical output or its lack of reasoning [Dom84].)

4.2 Categorical Knowledge

Knowledge which consists only of logical relationships between facts, and which contains no element of doubt, is called *categorical.* Categorical knowledge can be expressed as 'IF-THEN' rules. In their simplest form, they have the structure

IF *Antecedent* THEN *Conclusion*

The antecedent is a conjunction of facts, and the conclusion is some new fact which may be inferred. By analogy with the term 'database', a collection of these rules is said to constitute a *knowledge base.*

4.2.1 Knowledge Base

For example, listed below are some rules to identify animals, written in the logic programming language Prolog [Clo81]. The first argument of each term is the rule's antecedent, and the second argument is the conclusion. The antecedent is a list of facts which must all be established before the conclusion can be drawn. Notice that the conclusions (e.g. 'is_bird') of some rules appear within the antecedents of others.

```
rule( [has_feathers,lays_eggs                ], is_bird    ).
rule( [has_scales,lives_on_land,lays_eggs ], is_reptile ).
rule( [has_scales,lives_in_water,lays_eggs], is_fish    ).
rule( [has_fur,drinks_milk                   ], is_mammal  ).
rule( [is_viviparous,drinks_milk             ], is_mammal  ).
rule( [is_bird,is_flightless,swims           ], is_penguin ).
rule( [is_bird,is_flightless,is_big          ], is_ostrich ).
rule( [is_mammal,lives_in_water,is_big       ], is_whale   ).
rule( [is_fish,is_big                        ], is_shark   ).
rule( [is_reptile,has_no_legs                ], is_snake   ).
```

4.2.2 Inference Engine

In order to apply a set of rules to solve a particular problem, we require an *inference engine*. Several different inference strategies are possible, and explicit separation of the declarative knowledge expressed in the rules from details of the inference algorithm is one of the distinct merits of the rule-based approach (as opposed to the procedural approach of say the flowchart). This makes it much easier to modify the knowledge as the expert system is being developed.

Suppose the following observations have been made about an animal. (They too are expressed as Prolog assertions.)

```
is_big.
is_flightless.
has_feathers.
lays_eggs.
```

Backward Chaining

Suppose now we wish to prove that the animal is an ostrich. After first checking that this fact is not already established, we choose any rule which concludes 'is_ostrich', and try to prove recursively all the facts in its antecedent. If this is unsuccessful, we choose an alternative rule (there are none in this example) and try again.

This inference algorithm is a depth-first backward-chaining method, and is often referred to as *goal driven*. It may he expressed in Prolog as follows.

```
bak(G):- G.
bak(G):- rule(A,G), map1(bak,A), assert(G).
```

Here, 'map1(P,L)' means that the single-argument predicate P holds for each member of the list L. It is defined

```
map1(_,[]).
map1(F,[H|T]):- P =.. [F,H], P, map1(F,T).
```

The goal 'bak(G)' succeeds precisely when the fact G can he established through hackward chaining. Notice, however, that the goal may not terminate if the knowledge base is cyclic. (Consider the effect of including '`rule([is_ostrich],is_ostrich).`' at the top of the list of rules.)

Identification of the animal entails trying to prove each of the possible hypotheses (in a medical context: 'diagnoses') in turn until one is successfully established, or none remain (in which case the rules are insufficient to permit identification). Expressing this in Prolog,

```
identity(G):- animal(G), bak(G).

animal(is_ostrich).
animal(is_penguin).
animal(is_ostrich).
animal(is_whale  ).
animal(is_shark  ).
animal(is_snake  ).
```

Forward Chaining

Alternatively, we can choose any rule whose antecedent is already established, but whose conclusion is not, and add this conclusion to the growing database of established facts. We repeat this procedure until the fact of interest is finally proven, or no further rules can be found.

This strategy is forward chaining, and is often referred to as *data driven*. It can be expressed in Prolog as follows. (Whether inference proceeds in a depth-first or breadth-first manner depends on the relative positions of the rules in the Prolog database.)

```
fwd(G):- G.
fwd(G):- rule(A,H), map1(call,A), not H, assert(H), fwd(G).
```

The goal 'fwd(G)' succeeds precisely when fact G can be established through forward chaining. Given the present form of rule, this inference strategy is impervious to cycles in the knowledge base. However, if the knowledge base is acyclic[1] then 'fwd(G)' is logically equivalent to 'bak(G)'.

The principal difference between forward and backward chaining concerns efficiency. Forward chaining tends to be more efficient when the number of available diagnoses is large, whereas backward chaining tends to be more efficient when the number is small.

The inference strategy is pertinent also to interactive programs. We have assumed that all possible observations about the animal are included in the Prolog database before inference begins. However, diagnostic expert systems are usually required to seek whatever further information is necessary in order to reach a diagnosis, by questioning the user. A simple extension to the definitions of 'bak(G)' and 'fwd(G)' will cause the program to ask the user the truth value of G when all else fails. The order in which questions are asked depends on the inference strategy employed.

[1]The facts can be ordered totally so that facts in the antecedent of any given rule are strictly lower than the conclusion.

Explanations

One of the advantages of the rule-based representation is the ease with which diagnostic conclusions can be supported by reasoned explanations. This is because each rule amounts to a justification for its own conclusion. A complete trace of the rules used to establish a final diagnosis thus provides a coherent and complete argument.

This is how our Prolog program can be extended to generate explanations. Firstly, if a fact G is found in the Prolog database, then the explanation for G is simply that it is 'given' (by the user).

```
exp(G,given(G)):- G.
```

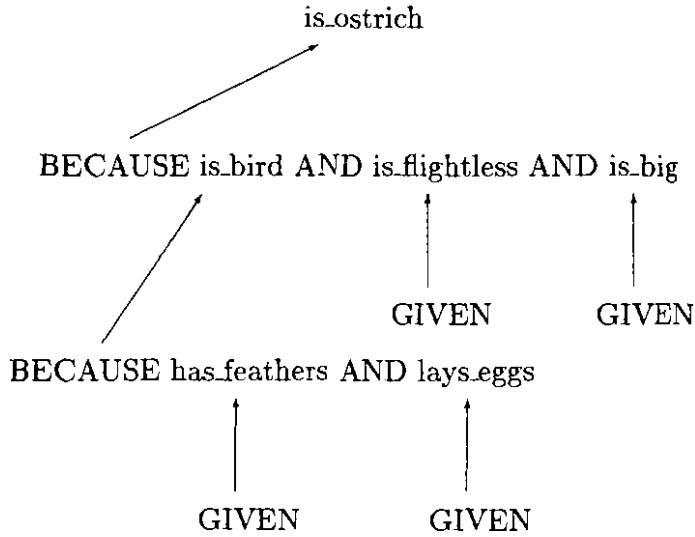
Backward chaining need proceed no further if an explanation for G is known. If not, then G is explained by the list A of proven facts, each supported by their own explanation, provided that a suitable rule with antecedent A and conclusion G can he found.

```
bak(G,E):- exp(G,E).
bak(G,since(G,E)):- rule(A,G), map2(bak,A,E),
                    assert(exp(G,since(G,E))).
```

Here, 'map2(P,$L0$,$L1$)' means that the binary predicate P holds for each pair of corresponding members of the lists $L0$ and $L1$. It is defined

```
map2(_,[],[]).
map2(F,[H0|T0],[H1|T1]):- P =.. [F,H0,H1], P, map2(F,T0,T1).
```

The explanation we obtain, when presented in a suitably readable format, is as follows.



Conflict Resolution

Rules can be a good deal more complicated than this. The antecedent can be an arbitrary boolean expression, the expression can include predicates as well as propositions, and the conclusion can be generalized to an 'action' to be taken if the antecedent is satisfied. An action may entail, for example, assertion or retraction of a fact, assignment of a value to a global variable, or printing of a message. When an expert system is used in a real-time control application, an action might be, say, to open or close a particular valve.

When forward chaining, if more than one rule's antecedent is satisfied, there is said to be *conflict* in the choice of rule to apply next. The inferences made, the actions taken, and the advice given by the expert system depend on the way these choices are made. The protocol for selecting among alternative rules is called a 'conflict resolution strategy'. Possible strategies are

1. *Priority Ordering* - Give rules a fixed priority, and choose the rule with the highest priority. This is the strategy employed in the animal

classification example; the order of the rules is simply the order in which they appear in the Prolog database.

2. *Specificity Ordering* - Choose a rule whose antecedent is maximally strong (logically). The rationale for this is that, if an antecedent holds, the stronger it is, the more pertinent is the corresponding rule to the present situation.

3. *Utility Ordering* - Choose the rule which employs cheapest materials, or which entails the least hazardous remedy.

4. *Recency Ordering* - Choose the rule which was most (or least) recently used.

5. *Context Limiting* - Partition rules into disjoint sets. Only one set of rules is active at any one time.

XCON/R1 is a rule-based expert system of this kind [McD82]. It is used by DEC to configure Vax systems. The program checks that the customer's order is complete and consistent, and then configures a layout of the computer system. XCON is said to have reduced the error rate on orders from 35% to 2%, and to have saved \$18-20m per year [Goo85, Jac86].

Medical reasoning, however, is almost inevitably associated with some uncertainty. One area, though, in which categorical decisions can be made is the planning of therapy. ONCOCIN is a rule-based system sharing similar principles with XCON, which assists a clinician to plan cancer treatment [Sho81].

Exercise 4.1 *Modify the definitions of 'bak(G)' and 'fwd(G)' so that instead of the observations having to be asserted into the Prolog database at the outset, the expert system questions the user. Define 'fwd(G,E)' so that it provides explanations.*

Exercise 4.2 *Repeat Exercise 2.1 (Page 5) using a rule-based representation instead.*

Exercise 4.3 *How might recursive partitioning (Section 3.5, Page 19) be used to induce rules automatically?*

4.3 MYCIN

If knowledge is uncertain, the degree of certainty in a rule can be expressed by some suitable parameter attached to the rule. The first system to incorporate an explicit mechanism for handling uncertainty in this way was MYCIN [Sho76].

MYCIN assists in the diagnosis and treatment of bacterial infections, and it has several hundred rules. The antecedent of a rule is a conjunction of clauses, each of which is a boolean expression. The conclusion of a rule is a list of new facts which may be inferred, and it is associated with a numerical *certainty factor* ranging from -1 to $+1$.

Here is an example of a MYCIN rule. The number appearing in brackets in the conclusion is the certainty factor.

RULE 85

IF
1) The site of the culture is blood, and
2) The Gram stain of the organism is negative, and
3) The morphology of the organism is rod, and
4) The patient is a compromised host

THEN
There is suggestive evidence (0.6) that the identity of the organism is pseudomonas aeruginosa.

This means that if all the conditions in the antecedent are satisfied, then our belief that the infecting organism is pseudomonas aeruginosa is significantly increased (by an amount '0.6').

4.3.1 Certainty Factors

A rule's certainty factor is elicited directly from the same expert who formulates the rule. The number is understood as the degree to which belief in the rule's conclusion would change if it were learned that the rule's antecedent were true. A certainty factor of $+1$ indicates that the conclusion would follow logically, and a certainty factor of -1 indicates that the conclusion would be completely refuted.

In the original MYCIN experiment, certainty factors were given a formal interpretation by relating them to subjective probabilities (Equation 4.1). Nevertheless, the certainty factors were still elicited directly; they were not calculated using these equations.

With reference to an arbitrary rule, we adopt the following notation for events (corresponding to *evidence* and *hypothesis*).

$$\begin{aligned} E &= \text{'The antecedent is true.'} \\ H &= \text{'The conclusion is true.'} \end{aligned}$$

The rule's certainty factor (CF) is then defined

$$\mathrm{CF} \;\widehat{=}\; \begin{cases} \dfrac{p(\mathrm{H} \mid \mathrm{E}) - p(\mathrm{H})}{1 - p(\mathrm{H})} & p(\mathrm{H}) \leq p(\mathrm{H} \mid \mathrm{E}) \\ \dfrac{p(\mathrm{H} \mid \mathrm{E}) - p(\mathrm{H})}{p(\mathrm{H})} & p(\mathrm{H}) > p(\mathrm{H} \mid \mathrm{E}) \end{cases} \tag{4.1}$$

4.3.2 Belief

The current belief 'bel $(\alpha = 1)$' in a fact α is also represented on a scale -1 to $+1$. (The user can therefore express doubt in his findings.) For example, the user might assert the following.

bel ('The site of the culture is blood.')	=	1.0
bel ('The Gram stain of the organism is negative.')	=	1.0
bel ('The morphology of the organism is rod.')	=	0.9
bel ('The patient is a compromised host.')	=	0.4

Propagation of Belief

The belief in a rule's antecedent is calculated from those of its component facts by taking the minimum over conjunctions and the maximum over disjunctions. The belief in the antecedent of Rule 85, for example, would thus be 0.4. The intuitive justification for this procedure is that a chain of necessary conditions is only as strong as its weakest link.

If the belief 'bel (E)' in a rule's antecedent is negative, then that rule causes no change in belief in its conclusion. This is because a rule is meant to influence belief in its conclusion only if we have some reason to believe that the antecedent holds.

However, if the belief in a rule's antecedent is positive then the change CF′ in belief in the rule's conclusion is taken to be the product of 'bel (E)' and the rule's certainty factor CF. This is because CF is defined to be the change in belief in the conclusion when the antecedent is known to hold for certain. If there is doubt, then the change must be attenuated. This is summarized by

$$\mathrm{CF}' \;\widehat{=}\; (0 \sqcup \mathrm{bel}\,(\mathrm{E})) \times \mathrm{CF} \tag{4.2}$$

where $\sqcup$ denotes the binary infix 'maximum' operator.

Combination of Belief

If more than one rule share the same conclusion, then separate changes in belief CF'_1 and CF'_2 are combined to form a resulting total change in belief $\mathrm{CF}'_1 \oplus \mathrm{CF}'_2$ using the following commutative and associative rule of combination.

$$\mathrm{CF}'_1 \oplus \mathrm{CF}'_2 \;\hat{=}\; \begin{cases} \dfrac{\mathrm{CF}'_1 + \mathrm{CF}'_2}{1 - (|\mathrm{CF}'_1| \sqcap |\mathrm{CF}'_2|)} & \mathrm{CF}'_1\mathrm{CF}'_2 < 0 \\ \mathrm{CF}'_1 + \mathrm{CF}'_2(1 - |\mathrm{CF}'_1|) & \mathrm{CF}'_1\mathrm{CF}'_2 \geq 0 \end{cases} \tag{4.3}$$

where $\sqcap$ denotes the binary infix 'minimum' operator.

Since the initial belief in the conclusion is zero, the resulting belief is given simply by the *total change* in belief. So if the conclusion appears within the antecedent of another rule, belief can be propagated by repeating the same procedure described above.

The certainty factor formalism has, however, been criticized for its *ad hoc* nature [Spi84]. Adams [Ada76] has shown that the definition of the MYCIN combinator (Equation 4.3) involves implicit assumption of both conditional and unconditional independence. Furthermore, Heckerman [Hec86] has also poiuted out that the original interpretation of certainty factors (Equation 4.1) is inconsistent with the combinator.

Exercise 4.4 *Show that the MYCIN combinator* $(\_ \oplus \_)$ *has identity element* 0, *and two zero elements* 1 *and* -1. *(Remember that* '$-1 \oplus 1$' *is undefined.)*

Exercise 4.5 *Show that the MYCIN combinator* $(\_ \oplus \_)$ *is commutative and associative.*

Hint — first show that funetion f *defined by*

$$f(x) \;\hat{=}\; \begin{cases} -\ln(1-x) & x > 0 \\ \ln(1+x) & x \leq 0 \end{cases}$$

has the property that

$$f(x \oplus y) = f(x) + f(y)$$

4.3.3 Inference Strategy

The inference strategy of MYCIN is backward chaining. Rules that bear on the current goal (determination of the value of a particular variable) are retrieved and evaluated. Any antecedent fact encountered whose current

belief is unknown causes a subgoal to be generated, and the process recurses. Subgoals, however, are generalizations of the unknown fact: so, for example, if the fact 'the identity of the organism is pseudomonas aeruginosa' is encountered, but its belief value is as yet undetermined, then a subgoal 'determine the identity of the organism' is created. If, after application of this recursive procedure, the total weight of evidence about the current goal remains small, the user is asked the value of the variable.

The search space is limited by means of 'meta-rules'. These have the same form as ordinary rules, but prescribe which rules to evaluate. In other words, the inference strategy is itself encoded by rules to some extent.

4.3.4 EMYCIN

MYCIN was found to perform at expert level [Yu79], but has never found a role in clinical practice. However, a derivative of MYCIN (with different rules), PUFF [Aik83a], has been applied successfully to the routine interpretation of lung function tests. An expert system without a knowledge base is referred to as an expert system *shell*. EMYCIN (standing for *Essential MYCIN*) is MYCIN's shell.

4.4 PROSPECTOR

The expert system PROSPECTOR assists geologists to evaluate exploration sites for mineral ores [Dud79]. It contains several hundred inference rules. The antecedent of a rule is a boolean combination of facts, and the conclusion of a rule is a single fact. Like MYCIN, the belief in an antecedent is calculated from the beliefs in the component facts by minimizing over conjunctions, and maximizing over disjunctions. Unlike MYCIN, however, beliefs are expressed on a scale 0 to 1, and they are interpreted as probabilities conditioned on the evidence available to the user.

4.4.1 Inference

Regarding a particular rule, if E is the event 'the antecedent holds', and H is the event 'the conclusion holds', then the *likelihood ratios*[2] λ and $\overline{\lambda}$ are defined by

$$\lambda \;\widehat{=}\; \frac{p(\mathrm{E} \mid \mathrm{H})}{p(\mathrm{E} \mid \overline{\mathrm{H}})} \tag{4.4}$$

[2]Note these are real numbers, and not random variables or events.

$$\overline{\lambda} \;\triangleq\; \frac{p(\overline{E} \mid H)}{p(\overline{E} \mid \overline{H})} \tag{4.5}$$

It follows that

$$\text{odds}\,(H \mid E) = \lambda \times \text{odds}\,(H) \tag{4.6}$$

and

$$\text{odds}\,(H \mid \overline{E}) = \overline{\lambda} \times \text{odds}\,(H) \tag{4.7}$$

Every rule is associated with a pair of values $(\lambda, \overline{\lambda})$, and every fact which is the conclusion of any rule is associated with prior odds 'odds (H)'. All these quantities were estimated subjectively by expert geologists.

Propagation of Probabilities

Equations 4.6 and 4.7 allow us to compute the conditional probability of H when it is known for certain either that E has occurred or that E has not occurred. However, in general, it is not known for certain whether E has occurred, either because E is directly observable but the user is doubtful about it, or because E is not observable and must be inferred by means of other uncertain rules. Either way, the probability that E has occurred is conditioned on the event U representing all the evidence the user has regarding E. The probability $p(E \mid U)$ is known, and we would like to compute $p(H \mid U)$. It follows that

$$\begin{aligned} p(H \mid U) &= p(H, E \mid U) + p(H, \overline{E} \mid U) \\ &= p(H \mid E, U)p(E \mid U) + p(H \mid \overline{E}, U)p(\overline{E} \mid U) \end{aligned} \tag{4.8}$$

If now we assume that E subsumes all evidence provided by U about H, as does $\overline{E}$, then

$$p(H \mid E, U) = p(H \mid E) \tag{4.9}$$

and

$$p(H \mid \overline{E}, U) = p(H \mid \overline{E}) \tag{4.10}$$

So, substituting and rearranging, Equation 4.8 becomes

$$p(H \mid U) = p(H \mid \overline{E}) + p(E \mid U)\left(p(H \mid E) - p(H \mid \overline{E})\right) \tag{4.11}$$

This means that $p(H \mid U)$ can be calculated by linear interpolation between the value it would have if E did not occur and the value it would have if

E did occur. One way to view this is to imagine the two likelihood ratios λ and $\overline{\lambda}$ being combined to form a single *effective likelihood ratio* λ' depending on the amount of evidence for E. This is defined

$$\lambda' \triangleq \frac{\text{odds}\,(\mathrm{H} \mid \mathrm{U})}{\text{odds}\,(\mathrm{H})} \tag{4.12}$$

Notice that

$$\begin{aligned} p(\mathrm{H} \mid \mathrm{U}) = p(\mathrm{H} \mid \mathrm{E}) &\Rightarrow \lambda' = \lambda \\ p(\mathrm{H} \mid \mathrm{U}) = p(\mathrm{H} \mid \overline{\mathrm{E}}) &\Rightarrow \lambda' = \overline{\lambda} \end{aligned}$$

Combination of Probabilities

If several (k) rules share the same conclusion, their separate evidence must be combined. Let E_i ($1 \leq i \leq k$) be the event 'the antecedent of the i^{th} rule holds', and let H be the event 'the (common) conclusion holds'. If all the E_i are independent both given H and given $\overline{\mathrm{H}}$, then

$$\text{odds}\,(\mathrm{H} \mid \mathrm{E}_1, \ldots, \mathrm{E}_k) = \lambda_1 \times \ldots \times \lambda_k \times \text{odds}\,(\mathrm{H}) \tag{4.13}$$

and

$$\text{odds}\,(\mathrm{H} \mid \overline{\mathrm{E}}_1, \ldots, \overline{\mathrm{E}}_k) = \overline{\lambda}_1 \times \ldots \times \overline{\lambda}_k \times \text{odds}\,(\mathrm{H}) \tag{4.14}$$

where λ_i and $\overline{\lambda}_i$ are the respective likelihood ratios for the i^{th} rule. (If the occurrence of E_i is not known for certain, then the corresponding effective likelihood ratio is used instead.)

Performance

Although the propagation formulae used by PROSPECTOR make strong independence assumptions, a close correspondence was found between the computed probabilities and an expert's subjective estimates with respect to three test cases. Furthermore, when put to practical use, PROSPECTOR was instrumental in the discovery of a deposit of molybdenum near Mount Tolman (Washington State), and later in the discovery of another in Alberta Canada (worth $100m).

Exercise 4.6 *What is the sample space Ω in relation to PROSPECTOR? To which (imagined) population do the probabilities relate?*

Exercise 4.7 *Derive Equations 4.6 and 4.7.*

Chapter 5

Descriptive Methods

A difficulty with the rule-based approach is that in many applications the validity of inferences is highly context-sensitive. So antecedents tend to be long, containing many preconditions, and the number of rules required tends to be large. Also, formulation of inference rules is a largely subjective and *ad hoc* procedure, and furthermore, experts tend to experience difficulty in articulating their expertise.

An alternative, and often more satisfactory method of representing medical knowledge, is to describe the *consequences* of diseases, rather than to say explicitly how to interpret symptoms. This descriptive knowledge of diseases must then be coupled to some suitable inference engine which performs the inverse task of finding the disease which most closely matches the actual findings.

5.1 INTERNIST

One of the largest medical expert systems that employs a descriptive representation, is INTERNIST [Mil82, Pop85]. It covers about 80% of general medicine, and it has descriptions of about 750 disorders. These were compiled from the medical literature and from interviews with specialists.

5.1.1 Knowledge Representation

Each disease description consists of a list of the manifestations (symptoms, signs *etc.*) that the presence of the disease can explain. Each manifestation in the list is associated with two numbers: a *frequency* and an *evoking strength*. Respectively, these are estimates of the frequency with which the disease produces the manifestation, and the frequency with which the dis-

ease explains the manifestation. They are expressed on a discrete, subjective scale from 0 to 5.

Every manifestation is also assigned an *importance*, irrespective of any disease, which indicates the necessity with which the manifestation must be explained by the final diagnosis. This is also expressed on a discrete scale, from 1 to 5. An importance of 1 means that the manifestation occurs commonly in normal persons and is easily disregarded. Whereas, an importance of 5 means that it is absolutely essential to explain the manifestation. The importance of a manifestation can thus be thought of as the frequency with which it can be explained by some identifiable disease.

Some of the 750 'diseases' are actually generalizations of others. For example, 'inflammatory hepatocellular disease' is a generalization of 'infectious mononucleosis', both being represented in the model. Conversely, others are more properly termed 'pathophysiological states'; for example, 'anaemia'. Links of various types (e.g. 'is_caused_by', 'is_subtype_of') exist between the diseases.

5.1.2 Inference Algorithm

When the clinical findings of a patient are entered into the computer, a list is compiled of the diseases which can explain any of the manifestations present. A score is calculated for each disease on the list using a heuristic scoring system. The score is based on the evoking strengths and importance values of the manifestations that are present, and the frequency values of the manifestations that are absent. Bonus points are awarded if there are links in the database to previously concluded diseases. Precise details of the scoring system can be found in [Mil82]. More general diseases are retained on the list in place of the more specific diseases they subsume, provided the latter are indistinguishable in their ability to explain the observed data.

Next, a set of competitors for the highest scoring disease is identified. Two diseases are said to be *competitors* precisely when the positive findings explained by one disease are a subset of those explained by the other. If there are no competitors, or if the nearest ones are 90 or more points below, then INTERNIST concludes that the highest scoring disease is present. Otherwise one of three diagnostic strategies is adopted according to the relative scores:

1. Closest competitor 46 to 89 points below $\Longrightarrow$ *Pursuing Mode*: questions are asked about manifestations with high evoking strength for the leading disease.

2. More than 4 competitors within 45 points of leading disease $\Longrightarrow$ *Rul-*

ing Out Mode: questions are asked about manifestations with high frequency numbers amongst the competitors.

3. From 1 to 4 competitors within 45 points of leading disease $\Longrightarrow$ *Discriminating Mode*: questions are asked which 'attempt to maximize the spread in scores'.

When the presence of a disease is concluded, the manifestations explained by that disease are removed from further consideration, and the procedure is repeated. This enables the diagnosis of multiple co-existent disorders.

5.1.3 Performance

When tested on 19 cases that had been published in the medical literature because of their abstruse nature, INTERNIST was found to have roughly the same diagnostic accuracy as hospital physicians. The principal weaknesses identified in INTERNIST's reasoning were the inabilities to synthesize a broad overview of the case, to reason temporally, and to reason anatomically.

5.1.4 CADUCEUS

A second version of INTERNIST, called CADUCEUS [Pop85], has an embellished knowledge base, and employs an improved diagnostic algorithm. Whereas INTERNIST has only one strategy (identify the common cause of a set of manifestations), CADUCEUS has several more operations to assist in constructing an explanation. These include identifying one evoked disease as a subtype or cause of another, and identifying shared subtypes of two evoked diseases. A search for a coherent explanation is then performed by repeated execution of these procedures, with the facility for back-tracking (unlike INTERNIST) when unfavourable evidence is obtained.

5.2 Discussion

INTERNIST and CADUCEUS utilize a *semantic network* representation in which links between entities (diseases, manifestations) are of more than one kind, denoting various types of relationship. Several other systems employ a similar representation: for example, CASNET [Wei78] and ABEL [Pat82]. Others have found *frames* useful which allow procedural information to be combined with declarative: for example, the systems PIP [Pau76] and CENTAUR [Aik83b].

5.2.1 Patient Specific Models

Common to all these approaches is the construction of a *patient specific model.* This means that the inference engine constructs an explanation for the specific set of findings under consideration, which is in some sense the *most likely.*

However, a maximum likelihood classification rule can be misleading. Consider the restricted case that there are precisely three alternative complete explanations that are consistent with the patient's clinical findings s:

$$\begin{aligned} \mathrm{E}_1 &\triangleq \text{'The patient has precisely diseases } \delta_1 \text{ and } \delta_2\text{.'} \\ \mathrm{E}_2 &\triangleq \text{'The patient has precisely diseases } \delta_1 \text{ and } \delta_3\text{.'} \\ \mathrm{E}_3 &\triangleq \text{'The patient has precisely diseases } \delta_4 \text{ and } \delta_5\text{.'} \end{aligned}$$

Suppose that the probabilities of these explanations conditioned on the findings are

$$\begin{aligned} p(\mathrm{E}_1 \mid \mathcal{S} = s) &= 0.3 \\ p(\mathrm{E}_2 \mid \mathcal{S} = s) &= 0.3 \\ p(\mathrm{E}_3 \mid \mathcal{S} = s) &= 0.4 \end{aligned}$$

Then the most likely explanation is E_3 that the patient has precisely the diseases δ_4 and δ_5. Yet, disease δ_1 is more likely to be present ($p(\mathrm{D}_1 \mid \mathcal{S} = s) = 0.6$) than either δ_4 or δ_5 ($p(\mathrm{D}_4 \mid \mathcal{S} = s) = p(\mathrm{D}_5 \mid \mathcal{S} = s) = 0.4$). This potential anomaly becomes much more pronounced as the number of possible alternative explanations increases.

Exercise 5.1 *In the design of an expert system employing a descriptive representation, how might one guard against the anomaly described above?*

Chapter 6

Causal Networks

6.1 Combining Statistical and Knowledge-Based Methods

When reasoning under uncertainty, inappropriate assumption of independence can lead to loss of diagnostic accuracy. Yet, general statistical methods which avoid independence assumptions (e.g. nearest neighbours, recursive partitioning) use training data inefficiently. However, it is often possible to predict statistical interactions and dependencies, from a knowledge of the underlying causal mechanisms of the given domain. A possible solution is therefore to assume a (possibly complicated) statistical model based on a knowledge of causal mechanisms, and then to estimate the numerical parameters of the model objectively, by reference to training data. This combined statistical and knowledge-based approach is now increasingly advocated for the design of diagnostic expert systems. Recent discoveries of efficient algorithms for propagating probabilities through graphical structures have made this approach much more feasible.

6.1.1 A Generalization

When symptoms tend to occur together in the presence of a particular disease, they are generally produced by some shared mechanism (often termed a 'pathophysiological state'). For example, gallstones (disease) can sometimes block the common bile duct causing obstructive jaundice (pathophysiological state), which causes the skin to turn yellow and the urine to become dark (symptoms). Analogously, with regard to fault diagnosis in cars for example, failure of the alternator ('disease') causes the battery to run down ('pathophysiological state') which both dims the headlights and makes it difficult

to start ('symptoms'). The key to making statistical dependencies explicit, therefore, is to introduce pathophysiological states into the computer model.

However, some symptoms can cause others: for example, a raised temperature causes sweating. Furthermore, sweating contributes to dehydration, which is not a symptom but a pathophysiological state! Also, the clinician using the expert system may *know* (or wish to hypothesize) that his patient has a particular pathophysiological state or co-existent disease. Ideally, therefore, it should be possible for the user to make assertions about pathophysiological states and diseases as well as symptoms. So the distinction between these three kinds of entity seems unhelpful.

Let us generalize and consider symptoms, pathophysiological states and diseases all simply as *propositional variables*. The diagnostic task given the values of any subset of these variables is to determine the likely values of the rest.

6.2 Causal Networks as a Representation

Let $\mathcal{A}$ be an indexed set of variables $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$. (Although, for simplicity, we assume these are binary variables, most of the techniques and results described here generalize easily to the multi-valued case). The knowledge representation task amounts to finding some suitable way of describing the joint probability distribution $p(\mathcal{A})$.

$$p(\mathcal{A}) = p(\alpha_1, \alpha_2, \ldots, \alpha_n) \tag{6.1}$$

Explicit definition of $p(\mathcal{A} = a)$ for every sequence (a) of values would, however, require tabulating 2^n separate probabilities. However, it follows from the definition of conditional probability (Equation 2.4, Page 8) that any joint probability can be defined by a chain of conditional probabilities (Equation 6.2). Furthermore, any such chain defines a valid joint probability; the two representations are equivalent.

$$p(\mathcal{A}) = p(\alpha_1) \times p(\alpha_2 \mid \alpha_1) \times p(\alpha_3 \mid \alpha_1, \alpha_2) \times \ldots \times p(\alpha_n \mid \alpha_1, \alpha_2, \ldots, \alpha_{n-1}) \tag{6.2}$$

The first few terms in Equation 6.2 are easily specified. Only one value ($p(\alpha_1 = 1)$) is required for the first term since $p(\alpha_1 = 0)$ is equal to $1 - p(\alpha_1 = 1)$. Similarly, two values are needed for the second term, four values for the third, and eight values for the fourth. However, the number of values increases exponentially, the last term requiring 2^{n-1}, so no saving is yet achieved.

6.2.1 Simplification

In practice, however, evidence about the state of any given variable α_i is exhausted by only a few anterior variables. They are called the *parents* of α_i, and their set is denoted by 'par(α_i)'.

Suppose, for example, the parents of α_{21} are α_3, α_7 and α_{16}. That is to say, $\text{par}(\alpha_{21}) = \{\alpha_3, \alpha_7, \alpha_{16}\}$. Therefore knowledge of the values of just those three variables exhausts the evidence provided by all the variables $\alpha_1, \ldots, \alpha_{20}$ regarding the value of the variable α_{21}. This means

$$p(\alpha_{21} \mid \alpha_1, \alpha_2, \ldots, \alpha_{20}) = p(\alpha_{21} \mid \alpha_3, \alpha_7, \alpha_{16})$$

So, the number of probabilities to specify is only eight instead of more than a million.

The greatest savings are likely to be achieved if the variables are indexed so that those which represent direct physical causes of any other lie anterior to it in the chain. Knowledge of the state of the direct causes of a given variable thus exhausts the anterior evidence regarding that variable's own state. For example, if it is known whether or not a car's battery is flat, then the state of the alternator does not affect the probability that the car will start.

The dependencies between the variables $\alpha_1, \ldots, \alpha_n$ can be expressed as a directed acyclic graph (DAG). The nodes are the variables, and the arcs indicate direct dependence; an arc from α_i to α_j indicates that α_i is one of the parents of α_j. Irrespective of whether the variables really are indexed so as to respect true physical causation (not possible if causation happens to be cyclic), we will refer to any such graph as a *causal graph*. Associated with each node α_i in the causal graph is a table specifying the conditional probability that $\alpha_i = 1$ given all possible states of its parents.

6.2.2 An Example

Let us develop the previous example. Suppose we wish to construct an expert system to help garage mechanics determine the likely faults with cars. The number of variables in any useful system would run into hundreds, but for simplicity let us select just five.

α = 'Alternator is ok'
β = 'Battery is charged'
γ = 'Carburettor is ok'
ε = 'Engine starts'
λ = 'Lights work'

With respect to any of these five variables, a value of 1 corresponds to 'true', and a value of 0 corresponds to 'false'.

Next we must decide how to order the variables. Let us do it as follows, because this accords with our knowledge of causation.

$$\alpha < \gamma < \beta < \lambda < \varepsilon$$

Now we must consider each variable in turn and decide which anterior variables are its parents. It may be helpful at this stage to refer to any available training data. If we do, we will probably be surprised to find that γ ('carburettor is ok') depends on α ('alternator is ok') even though there is no apparent causal connection. This is because the joint distribution we are describing is implicitly conditioned on the event 'the car is taken to the garage to be mended'. This makes rare, independent faults become almost mutually exclusive (but much more prevalent than in the unselected population of vehicles). We must therefore retain α as a parent of γ. Figure 6.1 shows this and other dependencies.

The conditional probability tables (Table 6.1) are derived from training data. The numbers are fictitious in this case, and serve only as an example.

The joint probabilities are easily recovered from these tables. For example,

$$\begin{aligned} p(\mathrm{A},\mathrm{B},\overline{\mathrm{C}},\overline{\mathrm{E}},\mathrm{L}) &= p(\mathrm{A})p(\mathrm{B} \mid \mathrm{A})p(\overline{\mathrm{C}} \mid \mathrm{A})p(\overline{\mathrm{E}} \mid \mathrm{B},\overline{\mathrm{C}})p(\mathrm{L} \mid \mathrm{B}) \\ &= 0.81 \times 0.89 \times (1-0.95) \times (1-0.23) \times 0.94 \\ &\approx 0.026 \end{aligned}$$

Exercise 6.1 *Calculate (or write a program to compute) the rest of the joint probability distribution from Table 6.1. Use the explicit joint probability distribution to calculate* $p(\mathrm{A} \mid \overline{\mathrm{L}},\overline{\mathrm{E}})$.

Figure 6.1: Causal graph of car faults.

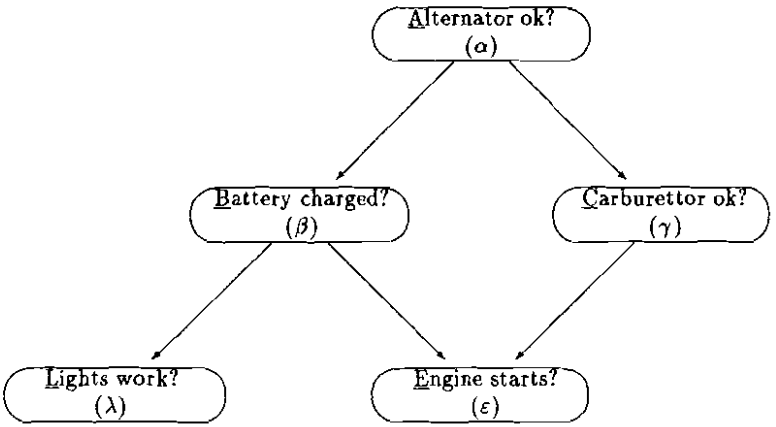


Table 6.1: Conditional probability tables for network shown in Figure 6.1.

α	$p(\mathrm{A})$	=	0.81	... *Alternators sometimes fail.*
β	$p(\mathrm{B} \mid \overline{\mathrm{A}})$	=	0.00	... *Battery discharges if alternator fails.*
	$p(\mathrm{B} \mid \mathrm{A})$	=	0.89	... *Batteries can fail for other reasons.*
γ	$p(\mathrm{C} \mid \overline{\mathrm{A}})$	=	0.99	... *Two separate faults are very unlikely.*
	$p(\mathrm{C} \mid \mathrm{A})$	=	0.95	... *Carburettors sometimes fail.*
λ	$p(\mathrm{L} \mid \overline{\mathrm{B}})$	=	0.10	... *Lights usually fail if battery is low.*
	$p(\mathrm{L} \mid \mathrm{B})$	=	0.94	... *Lights can fail for other reasons.*
ε	$p(\mathrm{E} \mid \overline{\mathrm{B}}, \overline{\mathrm{C}})$	=	0.05	... *Starting less likely if both faults.*
	$p(\mathrm{E} \mid \overline{\mathrm{B}}, \mathrm{C})$	=	0.12	... *A low battery hinders starting.*
	$p(\mathrm{E} \mid \mathrm{B}, \overline{\mathrm{C}})$	=	0.23	... *A faulty carburettor hinders starting.*
	$p(\mathrm{E} \mid \mathrm{B}, \mathrm{C})$	=	0.85	... *Engines can fail for other reasons.*

6.2.3 Separation

A variety of conditional independencies can be read from a causal graph. We start with some definitions.

An undirected path between variables α and β in a causal graph is said to be *blocked* by a set of variables $\mathcal{C}$ if one of the following holds.

1. Two arcs on the path meet tail to tail at a variable γ in $\mathcal{C}$.
2. Two arcs on the path meet head[1] to tail at a variable γ in $\mathcal{C}$.
3. Two arcs on the path meet head to head at a variable γ such that neither γ nor any of γ's descendants[2] are in $\mathcal{C}$.

Two variables α and β in a causal graph are said to be *separated* by a set of variables $\mathcal{C}$ if every undirected path between α and β is blocked by $\mathcal{C}$. By extension, two disjoint sets of variables $\mathcal{A}$ and $\mathcal{B}$ are said to be separated by $\mathcal{C}$ if every member of $\mathcal{A}$ is separated by $\mathcal{C}$ from every member of $\mathcal{B}$.

The notion of separation is the graphical equivalent of conditional independence [Pea86]. If $\mathcal{C}$ separates $\mathcal{A}$ from $\mathcal{B}$ then the variables $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent given $\mathcal{C}$. See [Nea89] Chapter 6 for a formal treatment of separation.

Exercise 6.2 *Regarding the causal graph shown in Figure 6.1, prove that*

$$p(\lambda, \varepsilon, \alpha \mid \beta, \gamma) = p(\lambda, \varepsilon \mid \beta, \gamma) \times p(\alpha \mid \beta, \gamma)$$

first without appeal to 'separation', and then again by arguing that $\{\beta, \gamma\}$ *separates* $\{\lambda, \varepsilon\}$ *from* $\{\alpha\}$.

6.2.4 Assumed Models

If a variable has more than a few direct causes, it may be infeasible to estimate all the entries in its conditional probability table from training data. If so, then it may be reasonable to assume a statistical model for the dependence of the variable upon the state of its direct causes.

One such model is the so-called 'noisy OR gate'. Here it is assumed that a variable α can be true only if at least one of its parents is also true. Suppose that $\text{par}(\alpha) = \mathcal{B}$, and $\mathcal{B} = \{\beta_1, \ldots, \beta_m\}$. According to the model, each such β_i has some specified probability p_i of causing α to be true, and these causation events are statistically independent. This can be expressed

$$p(A \mid \mathcal{B} = b) = 1 - \prod_{1 \le i \le m} (1 - b_i p_i) \tag{6.3}$$

[1] Arrow head

[2] Variables reachable via a *directed* path from γ.

where b stands for some arbitrary sequence of values, $[b_1, \ldots, b_m]$.

A network in which all the tables are defined in this way is sometimes referred to as a *probabilistic causal graph*. A medical expert system using this representation has been proposed by Peng and Reggia [Pen87], although earlier experiments with a similar representation were unrewarding [Lud83].

An alternative to the 'noisy OR gate' is the logistic model (Equation 3.7, Page 16). Reference to training data will help to decide which model is the most appropriate for any given application.

6.3 Inference

Although causal graphs are an efficient way of representing a joint distribution over a set of variables, the inference task is unfortunately NP-Hard [Coo89]. Nevertheless, efficient algorithms are known for restricted kinds of causal graphs.

6.3.1 Inference in Causal Trees

If the causal graph is restricted to a tree, then design of an inference algorithm is particularly straightforward. Consider the tree shown in Figure 6.2. In a medical context, the upper variables would correspond to diseases, and the lower variables to symptoms. Suppose we observe that symptoms π and σ are present and symptoms δ, ν, and τ are absent. If, furthermore, we choose to assume that disease γ is present, how likely is disease β to be present too? The conditional probability we wish to compute can be expressed as the ratio of two marginal probabilities.

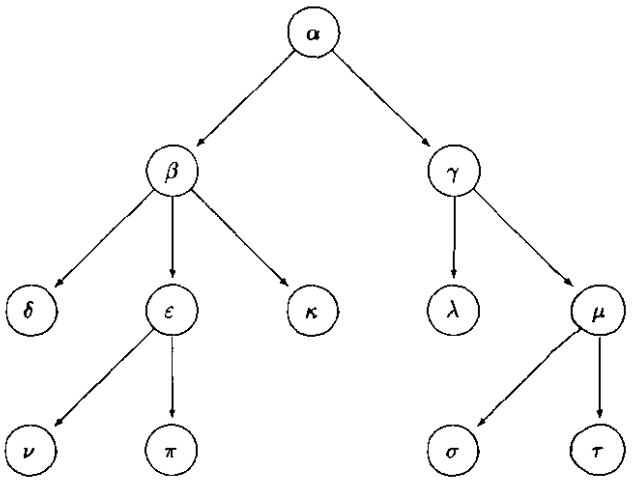
$$p(\mathrm{B} \mid \mathrm{C}, \overline{D}, \overline{N}, P, S, \overline{T}) = \frac{p(\mathrm{B}, \mathrm{C}, \overline{D}, \overline{N}, P, S, \overline{T})}{p(\mathrm{C}, \overline{D}, \overline{N}, P, S, \overline{T})} \tag{6.4}$$

The task of computing conditional probabilities therefore reduces to one of computing marginal probabilities. A general algorithm for this is easily derived.

Let $\mathcal{A}$ be the set of variables in a causal tree ($\mathcal{A} = \{\alpha_1, \ldots, \alpha_n\}$), where α_1 is the root. Suppose $\mathcal{I}$ is the set of variables in the tree which are instantiated to particular values ($\mathcal{I} \subseteq \mathcal{A}$), and we wish to determine the probability $p(\mathcal{I})$. Regardless of whether α_1 is one of the instantiated variables in $\mathcal{I}$

$$\begin{aligned} p(\mathcal{I}) &= p(\mathcal{I}, \mathrm{A}_1) + p(\mathcal{I}, \overline{\mathrm{A}_1}) \\ &= p(\mathcal{I} \mid \mathrm{A}_1)p(\mathrm{A}_1) + p(\mathcal{I} \mid \overline{\mathrm{A}_1})p(\overline{\mathrm{A}_1}) \end{aligned} \tag{6.5}$$

Figure 6.2: An example of a causal tree. (Conditional probabilities are given in Table 6.2.)



Let $\mathcal{I}_i$ denote the subset of $\mathcal{I}$ in the tree rooted at α_i, for any i $(1 \leq i \leq n)$. Also, for any value u $(u \in \{0,1\})$ let us define $\omega_i(u)$ to be the conditional probability associated with the variables $\mathcal{I}_i$ given that α_i takes value u.

$$\omega_i \;\widehat{=}\; \lambda u : \{0,1\} \bullet p(\mathcal{I}_i \mid \alpha_i = u) \tag{6.6}$$

Thus Equation 6.5 may be rewritten

$$p(\mathcal{I}) = \omega_1(1)p(\mathrm{A}_1) + \omega_1(0)p(\overline{\mathrm{A}_1}) \tag{6.7}$$

Notice that $p(\mathrm{A}_1)$, and hence $p(\overline{\mathrm{A}_1})$, can be found in the conditional probability table associated with the root node of the causal tree. Furthermore, ω_1 is determined by the following recursive equations.

For all i and u, if α_i is in $\mathcal{I}$, and u is not its instantiated value then

$$\omega_i(u) = 0 \tag{6.8}$$

Table 6.2: Conditional probability tables for causal tree shown in Figure 6.2.

α	$p(\mathrm{A})$	$=$	0.3	λ	$p(\mathrm{L} \mid \overline{\mathrm{C}})$	$=$	0.1
					$p(\mathrm{L} \mid \mathrm{C})$	$=$	1.0
β	$p(\mathrm{B} \mid \overline{\mathrm{A}})$	$=$	0.9	μ	$p(\mathrm{M} \mid \overline{\mathrm{C}})$	$=$	0.0
	$p(\mathrm{B} \mid \mathrm{A})$	$=$	0.1		$p(\mathrm{M} \mid \mathrm{C})$	$=$	0.7
γ	$p(\mathrm{C} \mid \overline{\mathrm{A}})$	$=$	0.2	ν	$p(\mathrm{N} \mid \overline{\mathrm{E}})$	$=$	0.5
	$p(\mathrm{C} \mid \mathrm{A})$	$=$	0.8		$p(\mathrm{N} \mid \mathrm{E})$	$=$	0.1
δ	$p(\mathrm{D} \mid \overline{\mathrm{B}})$	$=$	0.5	π	$p(\mathrm{P} \mid \overline{\mathrm{E}})$	$=$	0.1
	$p(\mathrm{D} \mid \mathrm{B})$	$=$	0.4		$p(\mathrm{P} \mid \mathrm{E})$	$=$	0.6
ε	$p(\mathrm{E} \mid \overline{\mathrm{B}})$	$=$	0.7	σ	$p(\mathrm{S} \mid \overline{\mathrm{M}})$	$=$	0.8
	$p(\mathrm{E} \mid \mathrm{B})$	$=$	0.2		$p(\mathrm{S} \mid \mathrm{M})$	$=$	0.5
κ	$p(\mathrm{K} \mid \overline{\mathrm{B}})$	$=$	0.3	τ	$p(\mathrm{T} \mid \overline{\mathrm{M}})$	$=$	0.3
	$p(\mathrm{K} \mid \mathrm{B})$	$=$	0.9		$p(\mathrm{T} \mid \mathrm{M})$	$=$	0.7

whereas if u is its instantiated value, or if α_i is not in $\mathcal{I}$, then

$$\omega_i(u) = \prod_{\alpha_j \in \mathrm{chn}(\alpha_i)} \Big(\omega_j(1)p(\mathrm{A}_j \mid \alpha_i = u) + \omega_j(0)p(\overline{\mathrm{A}_j} \mid \alpha_i = u)\Big) \qquad (6.9)$$

where '$\mathrm{chn}(\alpha_i)$' denotes the children of α_i in the tree. Again, notice that $p(\alpha_j \mid \alpha_i)$ is given by the conditional probability table associated with node α_j. Provided that these equations are applied starting at the leaves and working back to the root of the tree, they enable efficient calculation of $p(\mathcal{I})$.

Proof

Equation 6.8 follows directly from the definition of ω (Equation 6.6). Equation 6.9 is derived as follows. If α_i is in $\mathcal{I}$ and u is its instantiated value

then

$$\begin{aligned}
\omega_i(u) = \; & p(\mathcal{I}_i \mid \alpha_i = u) \\
& \qquad \ldots \text{Definition of } \omega. \\
= \; & p(\mathcal{I}_i - \{\alpha_i\} \mid \alpha_i = u) \\
& \qquad \ldots \text{Since } u \text{ is the instantiated value of } \alpha_i. \\
= \; & p(\bigcup_{\alpha_j \in \mathrm{chn}(\alpha_i)} \mathcal{I}_j \mid \alpha_i = u) \\
& \qquad \ldots \text{From definition of } \mathcal{I}_i. \\
= \; & \prod_{\alpha_j \in \mathrm{chn}(\alpha_i)} p(\mathcal{I}_j \mid \alpha_i = u) \\
& \qquad \ldots \text{Since } \alpha_i \text{ separates each of the } \mathcal{I}_j. \\
= \; & \prod_{\alpha_j \in \mathrm{chn}(\alpha_i)} (p(\mathcal{I}_j, \mathrm{A}_j \mid \alpha_i = u) + p(\mathcal{I}_j, \overline{\mathrm{A}_j} \mid \alpha_i = u)) \\
& \qquad \ldots \text{Partitions event.} \\
= \; & \prod_{\alpha_j \in \mathrm{chn}(\alpha_i)} (p(\mathcal{I}_j \mid \mathrm{A}_j, \alpha_i = u) p(\mathrm{A}_j \mid \alpha_i = u) \\
& \qquad + p(\mathcal{I}_j \mid \overline{\mathrm{A}_j}, \alpha_i = u) p(\overline{\mathrm{A}_j} \mid \alpha_i = u)) \\
= \; & \prod_{\alpha_j \in \mathrm{chn}(\alpha_i)} (p(\mathcal{I}_j \mid \mathrm{A}_j) p(\mathrm{A}_j \mid \alpha_i = u) + p(\mathcal{I}_j \mid \overline{\mathrm{A}_j}) p(\overline{\mathrm{A}_j} \mid \alpha_i = u)) \\
& \qquad \ldots \text{Since } \alpha_j \text{ separates } \mathcal{I}_j \text{ from } \alpha_i. \\
= \; & \prod_{\alpha_j \in \mathrm{chn}(\alpha_i)} (\omega_j(1) p(\mathrm{A}_j \mid \alpha_i = u) + \omega_j(0) p(\overline{\mathrm{A}_j} \mid \alpha_i = u)) \\
& \qquad \ldots \text{Definition of } \omega.
\end{aligned}$$

The case where α_i is not in $\mathcal{I}$ follows similarly.

Exercise 6.3 *Calculate the conditional probability specified by Equation 6.4 (Page 47) by applying Equations 6.8 and 6.9 to the data given in Table 6.2. (Calculate first the denominator of the right-hand side of Equation 6.4, and then the numerator.) What is the computational complexity of this procedure?*

6.3.2 Inference in Sparse Causal Graphs

Overview

Recently, an inference algorithm has been described for causal graphs which is efficient provided that the graph is sparse [Lau88]. The method entails clustering together interacting variables in such a way that the dependence between the sets of variables has a tree structure. This is carried out as a

single pre-processing step at the time of building the expert system. Only the tree is then retained for calculating conditional probabilities as and when required. The algorithm for computing these conditional probabilities makes special use of the fact that the conditional probability of an event is proportional to the joint probability when the conditioning event F is held constant.

$$p(\mathrm{E} \mid \mathrm{F}) \propto p(\mathrm{E}, \mathrm{F})$$

A more detailed description of the entire method is now given in reverse order so as to motivate each preceding step. We start with some definitions.

Definitions

Let $\mathcal{A}$ be a set of propositional variables ($\mathcal{A} = \{\alpha_1, \ldots, \alpha_n\}$), and let Γ denote a collection of sets of these variables ($\Gamma = \{\mathcal{C}_1, \ldots, \mathcal{C}_p\}$); for example, $\mathcal{C}_3 = \{\alpha_2, \alpha_7, \alpha_9\}$. If ψ is a function which maps instantiations of the variables in $\mathcal{C}_i$ to the reals, for each i ($1 \leq i \leq p$), such that for some constant k

$$p(\mathcal{A}) = k \prod_{1 \leq i \leq p} \psi(\mathcal{C}_i) \tag{6.10}$$

then (Γ, ψ) is said to be a *potential representation* of the joint probability distribution over $\mathcal{A}$.

For each set $\mathcal{C}_i$ we define the *separator* $\mathcal{S}_i$ and the *residual* $\mathcal{R}_i$ as follows ($\mathcal{S}_1$ is simply the empty set).

$$\mathcal{S}_i \;\hat{=}\; \mathcal{C}_i \cap (\mathcal{C}_1 \cup \mathcal{C}_2 \cup \ldots \cup \mathcal{C}_{i-1}) \tag{6.11}$$

$$\mathcal{R}_i \;\hat{=}\; \mathcal{C}_i - \mathcal{S}_i \tag{6.12}$$

The set Γ is said to have the *running intersection* property if for all $i > 1$ there exists a $j < i$ such that $\mathcal{S}_i \subseteq \mathcal{C}_j$. The set $\mathcal{C}_j$ is then called the *parent* of $\mathcal{C}_i$. If more than one such $\mathcal{C}_j$ exists then the choice as to which one is the parent is arbitrary. Thus the relationship between the sets in Γ has a tree structure, the root being $\mathcal{C}_1$.

It follows that if (Γ, ψ) is a potential representation of the joint distribution such that Γ has the running intersection property then marginal probabilities (for example, $p(\mathrm{A}_8)$) can be computed using the method given below. The theorems are stated here without proof. For a fuller and more formal treatment, see [Lau88, Nea89].

Conditioning on Evidence

Firstly, if any of the variables in $\mathcal{A}$ are instantiated to particular values then a new potential function ψ' is obtained from ψ by substituting the instantiated

values. Thus, continuing the earlier example, if α_2 is instantiated to the value 1, and α_9 to the value 0, then $\mathcal{C}_3$ shrinks to the singleton set $\mathcal{C}'_3 = \{\alpha_7\}$ and ψ' is defined on $\mathcal{C}'_3$ as follows.

$$\begin{aligned} \psi'(\overline{\mathrm{A}_7}) &= \psi(\mathrm{A}_2, \overline{\mathrm{A}_7}, \overline{\mathrm{A}_9}) \\ \psi'(\mathrm{A}_7) &= \psi(\mathrm{A}_2, \mathrm{A}_7, \overline{\mathrm{A}_9}) \end{aligned}$$

The other $\mathcal{C}_i$ are treated similarly.

It follows, because of the constant of proportionality k in Equation 6.10, that (Γ', ψ') is itself a potential representation of the joint distribution over the uninstantiated variables in $\mathcal{A}$, conditioned on the instantiated variables. Furthermore, Γ' inherits the running intersection property from Γ. Thus, by first conditioning the potential representation on available evidence before recovering marginal probabilities, we obtain conditional probabilities (for example, $p(\mathrm{A}_8 \mid \mathrm{A}_2, \overline{\mathrm{A}_9})$) instead.

Computing Marginal Probabilities

Marginal probabilities are recovered in three stages.

1. The conditional probabilities $p(\mathcal{R}_i \mid \mathcal{S}_i)$ are computed for each i, starting with $i = p$ and working down to $i = 1$. This entails repeated application of the following two steps. Firstly, it follows that

$$p(\mathcal{R}_p \mid \mathcal{S}_p) = \frac{\psi(\mathcal{C}_p)}{\sum_{\mathcal{R}_p} \psi(\mathcal{C}_p)} \tag{6.13}$$

where the sum is over all possible instantiations of the variables in $\mathcal{R}_p$. Secondly, let $\mathcal{C}_j$ be the parent of $\mathcal{C}_p$. We now define a new potential function ψ' according to

$$\psi'(\mathcal{C}_i) \triangleq \begin{cases} \psi(\mathcal{C}_i) & i \neq j \\ \psi(\mathcal{C}_i) \sum_{\mathcal{R}_p} \psi(\mathcal{C}_p) & i = j \end{cases} \tag{6.14}$$

It follows that $(\{\mathcal{C}_1, \ldots, \mathcal{C}_{p-1}\}, \psi')$ is a potential representation of the joint distribution over $\mathcal{C}_1 \cup \ldots \cup \mathcal{C}_{p-1}$. So by repeating these steps we can recover all the $p(\mathcal{R}_i \mid \mathcal{S}_i)$.

2. From these, the probabilities $p(\mathcal{C}_i)$ are computed for each i, starting with $i = 1$ and working up to $i = p$. Since $\mathcal{S}_1 = \{\}$, it follows that $p(\mathcal{C}_1) = p(\mathcal{R}_1 \mid \mathcal{S}_1)$. The rest are calculated using $p(\mathcal{C}_i) = p(\mathcal{R}_i \mid \mathcal{S}_i) p(\mathcal{S}_i)$. Probability $p(\mathcal{S}_i)$ is determined by summing $p(\mathcal{C}_j)$ over all possible instantiations of the variables in $\mathcal{C}_j - \mathcal{S}_i$, where $\mathcal{C}_j$ is $\mathcal{C}_i$'s parent.

3. The marginal probability of any given variable is then determined from $p(C_i)$, for any C_i containing that variable, by summing out over all possible instantiations of the remaining variables in C_i.

Exercise 6.4 *Prove Equation 6.13.*

Obtaining a Potential Representation ψ

We have now seen how to calcnlate conditional probabilities from a potential representation (Γ, ψ) with the running intersection property. How can we obtain a suitable potential function ψ from a causal graph representation of the joint distribution over the variables $\mathcal{A}$?

If Γ is chosen such that for every variable α_i there exists a C_j which contains both α_i and all α_i's parents (i.e. $\alpha_i \in C_j \wedge \mathrm{par}(\alpha_i) \subseteq C_j$) then ψ can be obtained by multiplication of the conditional probability tables associated with the causal graph. This is accomplisbed by assigning every variable α_i to exactly one C_j which contains both that variable and its parents. If there is more than one such C_j then the choice is arbitrary.

For each C_j let I_j be the (possibly empty) set of variables assigned to C_j. The potential function ψ is then defined for each C_j by

$$\psi(C_j) \;\widehat{=}\; \prod_{\alpha_i : I_j} p(\alpha_i \mid \mathrm{par}(\alpha_i)) \tag{6.15}$$

This forms a valid potential representation of the joint distribution (with constant $k = 1$) because

$$\prod_{1 \le j \le p} \psi(C_j) = \prod_{1 \le i \le n} p(\alpha_i \mid \mathrm{par}(\alpha_i)) \tag{6.16}$$

Obtaining a Cover Γ

It remains, therefore, only to find a suitable collection Γ of sets of variables. It must have the running intersection property, and each variable must appear together with its parents in at least one of these sets.

Lauritzen and Spiegelhalter's method consists of first forming the *moral graph* from the original causal graph by marrying all common parents: that is to say, inserting an undirected edge between any two parents of a variable that are not already joined, and then dropping directions of all edges.

Next, the nodes of the moral graph are ordered (assigned rank 1 to n) by *maximum cardinality search* [Tar84], which proceeds as follows. First, rank 1 is assigned to an arbitrary variable. Tben, repeatedly, the variable

adjacent to the greatest number of previously numbered variables (breaking ties arbitrarily) is chosen as the next to number.

It follows that if the moral graph is triangulated[3] and if Γ is taken to be the set of cliques[4], ordered by their highest ranked variable, then Γ has the running intersection property. Furthermore, since each variable together with its direct causes forms a complete set[5] in the moral graph, they must all appear together, as required, in at least one clique. If the moral graph is not already triangulated, then a simple algorithm [Tar84] fills in with extra edges until the graph is triangulated.

Computational Complexity

The Lauritzen-Spiegelhalter algorithm is applicable to any causal graph, yet computation of conditional probabilities is known to be NP-Hard [Coo89]. For which kinds of graph is the algorithm efficient, and which component of the algorithm becomes infeasible when the method is applied to an unsuitable kind of graph?

Pre-processing of the causal graph can always be completed in polynomial time. Algorithms are available for performing maximum cardinality search, and for triangulating graphs by computing the fill-in, which are $O(n+e)$ where n is the number of nodes (variables) and e is the number of edges in the causal graph [Tar84]. Furthermore, an $O(n+e)$ algorithm is known for enumerating the cliques of a triangulated graph [Gol80]. This is possible because in the case of triangulated graphs, the number of cliques is no greater than the number of nodes.

However, initialization of the potential function ψ and computation of marginal probabilities are $\Omega(2^m)$ where m is the number of variables in the largest clique. This is the critical factor which determines the feasibility of the algorithm for any particular causal graph, and the size of the largest clique is discovered during the pre-processing step.

In one medical application of this method, MUNIN [And87] a system to assist the interpretation of electromyographic findings, no clique was found to contain more than four variables.

6.3.3 Monte Carlo Inference Methods

One technique worth considering when others are found to be infeasible, is Monte Carlo simulation. Pearl [Pea87] has proposed a stochastic simulation

[3] (i.e. contains no cycle of more than three nodes without a bridging edge)

[4] A *clique* is a maximal complete set of nodes.

[5] A *complete set* of nodes is that which induces a complete subgraph.

method in which the known variables are clamped to their respective and the unknown variables are assigned random values. Then, each of the unknown variables in turn is assigned a (possibly new) random value with probability determined by the conditional probability tables. This procedure is repeated many times, nntil the system reaches a stationary distribution. For any i, the relative frequency with which variable α_i takes value 1 during the simulation thus provides an estimate of the probability $p(\mathrm{A}_i)$ conditioned on the known variables.

Underlying this method is the principal that a variable α_i depends on all others only through its parents, its children and their parents. These are said to constitute the variable's *Markov blanket.*

The nature of a variable's dependence on its Markov blanket is quite simple. Let $\mathcal{R}_i$ here denote the set of all variables except α_i.

$$\mathcal{R}_i \triangleq \mathcal{A} - \{\alpha_i\} \tag{6.17}$$

The conditional probability associated with α_i given the values of all other variables is

$$\begin{aligned} p(\alpha_i \mid \mathcal{R}_i) &= \frac{p(\alpha_i, \mathcal{R}_i)}{p(\mathcal{R}_i)} \\ &= \frac{p(\mathcal{A})}{\sum_{\alpha_i} p(\mathcal{A})} \end{aligned} \tag{6.18}$$

where the sum is over all possible values (0 and 1) of variable α_i. Now, the joint distribution specified by the causal graph is

$$p(\mathcal{A}) = \prod_{1 \le j \le n} p(\alpha_j \mid \mathrm{par}(\alpha_j)) \tag{6.19}$$

Thus, from Equations 6.18 and 6.19,

$$p(\alpha_i \mid \mathcal{R}_i) = \frac{\prod_{1 \le j \le n} p(\alpha_j \mid \mathrm{par}(\alpha_j))}{\sum_{\alpha_i} \prod_{1 \le j \le n} p(\alpha_j \mid \mathrm{par}(\alpha_j))} \tag{6.20}$$

However, α_i appears only in the term $p(\alpha_i \mid \mathrm{par}(\alpha_i))$ and in each term $p(\alpha_j \mid \mathrm{par}(\alpha_j))$ where α_j is a child of α_i (and equivalently $\alpha_i \in \mathrm{par}(\alpha_j)$). The other terms therefore factor and cancel.

$$p(\alpha_i \mid \mathcal{R}_i) = \frac{p(\alpha_i \mid \mathrm{par}(\alpha_i)) \prod_{\alpha_j \in \mathrm{chn}(\alpha_i)} p(\alpha_j \mid \mathrm{par}(\alpha_j))}{\sum_{\alpha_i} \left(p(\alpha_i \mid \mathrm{par}(\alpha_i)) \prod_{\alpha_j \in \mathrm{chn}(\alpha_i)} p(\alpha_j \mid \mathrm{par}(\alpha_j)) \right)} \tag{6.21}$$

However, the denominator is independent of the actual value of α_i since the sum is over all possible values. So

$$p(\alpha_i \mid \mathcal{R}_i) \propto \left(p(\alpha_i \mid \text{par}(\alpha_i)) \prod_{\alpha_j \in \text{chn}(\alpha_i)} p(\alpha_j \mid \text{par}(\alpha_j)) \right) \tag{6.22}$$

where the constant of proportionality does not depend on α_i, it depends only on the values of α_i's Markov blanket. Equation 6.22 thus provides a more efficient way of calculating the conditional probability $p(\alpha_i \mid \mathcal{R}_i)$ than does Equation 6.21. This is because, provided no variable's Markov blanket is too large, the term on the right-hand side of the proportionality can be pre-computed as a reference table.

Although this method seems a powerful technique for small but highly connected causal graphs that would otherwise not yield to exact methods, convergence tends to become unacceptably slow as the number of nodes in the network increases [Coo89].

Exercise 6.5 *Without recourse to Equation 6.22, use the definition of* separation *(Section 6.2.3, Page 46) to argue that a variable depends on all others in a causal graph only through its Markov blanket.*

Chapter 7

A Probabilistic Rule-Based System

It has often been argued that the rule-based approach is inappropriate for reasoning under uncertainty [Hec86, Nea89]. In this chapter, we first look at how a rule-based system could be constructed, and then discuss the relative merits of adopting a rule-based rather than a descriptive approach.

7.1 A Causal Graph Representation

The direction of inference in a rule-based system is usually in the reverse direction to that of causation. In general, we observe symptoms, and we wish to determine which disease has caused them, rather than the other way around. A causal graph representation would therefore appear to be the converse to that required for direct inferential knowledge. However, Equation 6.2 (Page 42) remains valid for any indexing of the variables, not just those which respect causation. The reason the latter are preferred when describing a joint probability distribution is that knowledge of the state of the direct causes of a variable tends to exhaust all other anterior evidence, so the resulting graph is sparser. If we are prepared to suspend temporarily any consideration of efficiency, we can use a causal graph to represent inferential knowledge.

Recall the earlier example (Section 6.2.2, Page 43) of a diagnostic program for faults in cars. Let us develop the causal graph again, but this time representing inferential rather than descriptive knowledge.

7.1.1 Car Faults Revisited

Given the task of developing an expert system to assist garage mechanics localize faults in cars, we have decided upon the following (trivial) list of binary variables.

$$\begin{array}{rcl} \alpha & = & \text{'Alternator is ok'} \\ \beta & = & \text{'Battery is charged'} \\ \gamma & = & \text{'Carburettor is ok'} \\ \varepsilon & = & \text{'Engine starts'} \\ \lambda & = & \text{'Lights work'} \end{array}$$

A value of 1 corresponds to 'true', and a value of 0 corresponds to 'false', as before. This time, however, let us order the variables according to the sequence in which we infer their values. Lowest are those variables whose values are directly observable (ε and λ). Higher variables are inferred from lower ones.

$$\lambda < \varepsilon < \beta < \gamma < \alpha$$

Next we construct the causal graph. We consider each variable in turn, and decide on which anterior variables in the new order it depends. We can discover this by inspection of the original causal graph (Figure 6.1, Page 45). If $\mathcal{A}$ is the set of variables anterior to a particular variable ϕ in the new order, then we wish to restrict $\mathcal{A}$ to the smallest subset $\mathcal{A}'$ which separates $\{\phi\}$ and $\mathcal{A} - \mathcal{A}'$ in the original causal graph.

1. Variable λ has no anterior variables. It is the root of the new graph.

2. Variable ε depends on λ because failure of the lights is evidence that the battery is flat, and therefore makes it less likely that the engine will start.

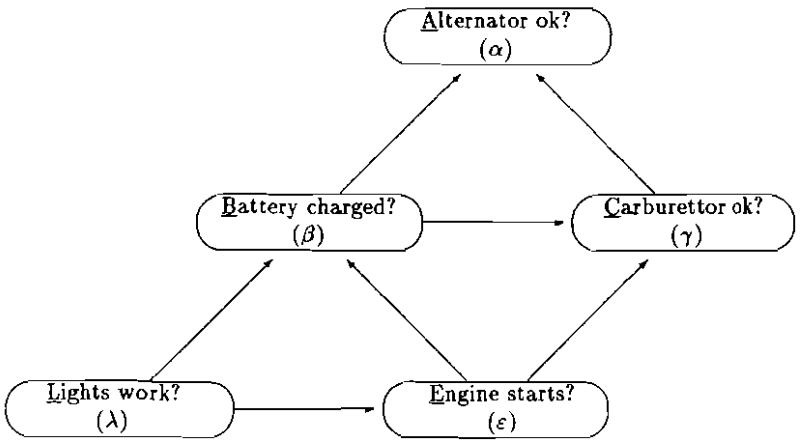
3. Variable β depends on both λ and ε since if it found that the lights work normally, or that the engine starts, then this is evidence that the battery is charged. Both λ and ε must be retained as parents of β in the new graph because neither separates the other from β in the old graph.

4. Variable γ, however, depends on λ only through β, since $\{\beta\}$ separates γ from λ in the old graph. Only ε and β need to be retained as pareuts of γ in the new graph.

5. Variable α depends on the other variables only through β and γ, since $\{\beta, \gamma\}$ separates $\{\alpha\}$ from $\{\lambda, \varepsilon\}$ in the descriptive graph. Therefore, only β and γ need to be retained as parents of α in the new graph.

So some economy of representation can still be achieved. Figure 7.1 shows the new 'causal' graph.

Figure 7.1: Causal graph of car faults.



For this example, let us derive the new conditional probability tables (Table 7.1, Page 60) from the previous ones (Table 6.1, Page 45) using, say, the Lauritzen-Spiegelhalter method, so that both the old and the new causal graphs specify the same joint probability distribution. In practice, though, we would derive the probabilities directly from training data.

Table 7.1: Conditional probability tables for network shown in Figure 7.1.

λ	$p(\mathrm{L})$	=	0.706	... *Lights usually work normally.*
ε	$p(\mathrm{E} \mid \overline{\mathrm{L}})$	=	0.221	... *If lights fail then usually no starting.*
	$p(\mathrm{E} \mid \mathrm{L})$	=	0.791	... *If lights work then probably starts ok.*
β	$p(\mathrm{B} \mid \overline{\mathrm{L}}, \overline{\mathrm{E}})$	=	0.034	... *No start/lights ⇒ battery probably low.*
	$p(\mathrm{B} \mid \overline{\mathrm{L}}, \mathrm{E})$	=	0.544	... *Engine starts ok suggests battery ok.*
	$p(\mathrm{B} \mid \mathrm{L}, \overline{\mathrm{E}})$	=	0.833	... *If lights work then battery probably ok.*
	$p(\mathrm{B} \mid \mathrm{L}, \mathrm{E})$	=	0.994	... *Start/lights ok ⇒ battery probably ok.*
γ	$p(\mathrm{C} \mid \overline{\mathrm{B}}, \overline{\mathrm{E}})$	=	0.975	... *Two separate faults are unlikely.*
	$p(\mathrm{C} \mid \overline{\mathrm{B}}, \mathrm{E})$	=	0.990	... *Starts ok ⇒ carburettor probably ok.*
	$p(\mathrm{C} \mid \mathrm{B}, \overline{\mathrm{E}})$	=	0.787	... *Carburettor fault can explain no start.*
	$p(\mathrm{C} \mid \mathrm{B}, \mathrm{E})$	=	0.986	... *Starts ok ⇒ carburettor probably ok.*
α	$p(\mathrm{A} \mid \overline{\mathrm{B}}, \overline{\mathrm{C}})$	=	0.701	... *Two separate faults are unlikely.*
	$p(\mathrm{A} \mid \overline{\mathrm{B}}, \mathrm{C})$	=	0.310	... *Flat battery suggests alternator fault.*
	$p(\mathrm{A} \mid \mathrm{B}, \overline{\mathrm{C}})$	=	1.000	... *If battery charged then alternator ok.*
	$p(\mathrm{A} \mid \mathrm{B}, \mathrm{C})$	=	1.000	... *If battery charged then alternator ok.*

The same joint probabilities as before are easily recovered from these tables, allowing for the fact that we are working to three decimal places only. For example,

$$\begin{aligned} p(\mathrm{A},\mathrm{B},\overline{\mathrm{C}},\overline{\mathrm{E}},\mathrm{L}) &= p(\mathrm{A} \mid \mathrm{B},\overline{\mathrm{C}})p(\mathrm{B} \mid \mathrm{L},\overline{\mathrm{E}})p(\overline{\mathrm{C}} \mid \mathrm{B},\overline{\mathrm{E}})p(\overline{\mathrm{E}} \mid \mathrm{L})p(\mathrm{L}) \\ &= 1.000 \times 0.833 \times (1-0.787) \times (1-0.791) \times 0.706 \\ &\approx 0.026 \end{aligned}$$

7.2 Assuming a Logistic Model

Actually, it was possible to invert the original causal graph as shown above only because of its small size. In general, we would find that a variable has so many new parents that it is infeasible to specify by explicit enumeration all the conditional probabilities in its table. That, after all, is the reason for trying to order the variables in a manner consistent with causation. We are deliberately doing the opposite here.

A simple solution, however, is to specify each variable's conditional probability table implicitly, by assuming a statistical model. A reasonable model to choose is the logistic one, since the parents of a variable now represent evidence rather than causative factors. (The 'noisy OR gate' model might have been a better choice had the latter been the case instead.)

Let us see how we can specify the conditional probability tables in the above example (Table 7.1). The first two variables present no difficulty because neither has more than one parent.

$$\ln \mathrm{odds}\ (\mathrm{L}) = 0.874 \tag{7.1}$$

$$\ln \mathrm{odds}\ (\mathrm{E} \mid \lambda) = -1.26 + 2.59\,\lambda \tag{7.2}$$

Notice that in Equation 7.2 we allow random variables to appear on the right-hand side. This is a shorthand for the more cumbersome

$$\forall u : \{0,1\} \bullet \ln \mathrm{odds}\ (\mathrm{E} \mid \lambda = u) = -1.26 + 2.59u$$

The next variable β has two parents (λ and ε). However, we are rather fortunate: β does indeed depend logistically on λ and ε. This is because λ and ε are conditionally independent given β; inspection of the original graph (Figure 6.1) confirms that $\{\beta\}$ separates λ and ε. Calculating the appropriate weights we obtain

$$\ln \mathrm{odds}\ (\mathrm{B} \mid \lambda,\varepsilon) = -3.342 + 4.949\,\lambda + 3.517\varepsilon \tag{7.3}$$

Thus only three parameters instead of four are required to specify the joiut probability table. In general, only $n + 1$ parameters rather than 2^n are required, where n is the number of parents.

7.2.1 Allowing Expressions

We are less fortunate with the next variable γ. It has two parents, ε and β, and $\{\gamma\}$ fails to separate them in the original graph, so they are not conditionally independent. A logistic relationship might still have held ueverthe-less, but not in this particular case.

However, a logistic form of dependence can always be obtained if we replace the variables in the logistic equation with arbitrary boolean expressions. Thus,

$$\begin{aligned}\ln\text{odds}\,(\mathrm{C} \mid \varepsilon, \beta) = \\ 3.683(\neg\varepsilon \wedge \neg\beta) + 4.635(\varepsilon \wedge \neg\beta) + 1.309(\neg\varepsilon \wedge \beta) + 4.252(\varepsilon \wedge \beta)\end{aligned} \tag{7.4}$$

where, for any values u, v ($u, v \in \{0, 1\}$),

$$\begin{aligned}\neg v &\;\widehat{=}\; 1 - v \\ u \wedge v &\;\widehat{=}\; u \times v\end{aligned}$$

A logistic relationship now holds because all of the terms are mutually exclusive. However, we have as many terms as there are entries in γ's conditional probability table. Nevertheless, we can approximate the required function by combining terms with similar weights. For example, combining the second and fourth terms (and averaging the weights) we obtain

$$\text{lnodds}\,(\mathrm{C} \mid \varepsilon, \beta) = 3.683(\neg\varepsilon \wedge \neg\beta) + 1.309(\neg\varepsilon \wedge \beta) + 4.443\,\varepsilon \tag{7.5}$$

In practice, we would use our knowledge of the domain to decide upon a set of terms that were mutually independent or exclusive, and then derive the relevant weights directly from training data. Although this involves an element of approximation and assumption, so too does the construction of a descriptive causal graph.

7.2.2 Transforming the Weights

The last variable α poses a special problem. How are we to handle logical constraints? Although α depends logistically ou its new parents, β and γ, because it separates them in the original graph, $p(\mathrm{A} \mid \mathrm{B}) = 1$. This means variable β would require an infinitely large weight in the logistic equation!

A simple solution is to transform the weights to the interval $[-1, +1]$. This is easily achieved [Haj85] by applying a suitable transformation such as $\mathcal{F}$.

$$\mathcal{F} \;\widehat{=}\; \lambda r : \mathbf{R} \bullet \frac{e^r - 1}{e^r + 1} \tag{7.6}$$

We will refer to these transformed weights as *certainty factors*. They combine not by simple addition, but by a new operator '$\oplus$'. (Let a and b denote arbitrary certainty factors.)

$$a \oplus b \;\widehat{=}\; \mathcal{F}(\mathcal{F}^{-1}(a) + \mathcal{F}^{-1}(b)) \tag{7.7}$$

Since $\mathcal{F}$ is bijective, the operator $\oplus$ inherits the properties of commutativity and associativity from simple addition, in terms of which it is defined. The operator also has 0 as its identity element. Furthermore, substituting for $\mathcal{F}$ in the definition above we obtain the following more familiar rule of combination [Haj85].

$$a \oplus b = (a + b)/(ab + 1) \tag{7.8}$$

Logical constraints are now represented by the certainty factors $+1$ and -1. These are both zero elements of the operator $\oplus$: in the presence of complete certainty, further evidence makes no difference. (Note that, as one would expect, $+1$ cannot be combined with -1 because that denotes contradiction.)

$$a \neq -1 \;\Rightarrow\; +1 \oplus a = +1 \tag{7.9}$$

$$a \neq +1 \;\Rightarrow\; -1 \oplus a = -1 \tag{7.10}$$

In general, let us denote the conditional *certainty* in an event E given F by 'cert (E | F)'. This is defined as the transformed log-odds. Thus, for any pair of events E and F,

$$\text{cert (E | F)} \;\widehat{=}\; \mathcal{F}(\ln\text{odds (E | F)}) \tag{7.11}$$

It follows that

$$\begin{aligned} \text{cert (E | F)} &= p(\text{E} \mid \text{F}) - p(\overline{\text{E}} \mid \text{F}) \\ &= 2p(\text{E} \mid \text{F}) - 1 \end{aligned} \tag{7.12}$$

So conditional probabilities are easily recovered from conditional certainties. Let us now transform Equations 7.1, 7.2, 7.3 and 7.4, before going on to deal with α. From Equation 7.1,

$$\mathcal{F}(\ln\text{odds (L)}) = \mathcal{F}(0.874) \tag{7.13}$$

Thus,

$$\text{cert}\ (\mathrm{L}) = 0.411 \tag{7.14}$$

From Equation 7.2

$$\mathcal{F}(\ln \text{odds}\ (\mathrm{E} \mid \lambda)) = \mathcal{F}(-1.26 + 2.59\,\lambda) \tag{7.15}$$

Thus, from the definition of $\oplus$, and since 0 is a fixed point of transformation $\mathcal{F}$,

$$\begin{aligned} \text{cert}\ (\mathrm{E} \mid \lambda) &= \mathcal{F}(-1.26 + 2.59\,\lambda) \\ &= \mathcal{F}(-1.26) \oplus \mathcal{F}(2.59\,\lambda) \\ &= \mathcal{F}(-1.26) \oplus \lambda\,\mathcal{F}(2.59) \\ &= -0.557 \oplus 0.860\,\lambda \end{aligned} \tag{7.16}$$

Similarly, from Equations 7.3 and 7.4

$$\text{cert}\ (\mathrm{B} \mid \lambda, \varepsilon) = -0.932 \oplus 0.986\,\lambda \oplus 0.942\,\varepsilon \tag{7.17}$$

$$\text{cert}\ (\mathrm{C} \mid \varepsilon, \beta) = 0.951(\neg\varepsilon \wedge \neg\beta) \oplus 0.981(\varepsilon \wedge \neg\beta) \oplus 0.575(\neg\varepsilon \wedge \beta) \oplus 0.972(\varepsilon \wedge \beta) \tag{7.18}$$

Lastly, the equation for variable α is

$$\text{cert}\ (\mathrm{A} \mid \beta, \gamma) = 0.402 \oplus \beta \oplus -0.678\,\gamma \tag{7.19}$$

7.2.3 Decomposition into Rules

If we so desire, we can rewrite Equations 7.14, 7.16, 7.17, 7.18 and 7.19 as a collection of inference rules. Each rule corresponds to one term in an equation. All that is required is to introduce 'true' as an expression that has constant value 1 in order to represent the constant term in each equation. The certainty factor of each term we write as a superscript to the implication symbol. Thus,

$$\text{true} \Rightarrow^{+0.411} \lambda \tag{7.20}$$

$$\text{true} \Rightarrow^{-0.557} \varepsilon \tag{7.21}$$

$$\lambda \Rightarrow^{+0.860} \varepsilon \tag{7.22}$$

$$\text{true} \Rightarrow^{-0.932} \beta \tag{7.23}$$

$$\lambda \Rightarrow^{+0.986} \beta \tag{7.24}$$

$$\varepsilon \Rightarrow^{+0.942} \beta \tag{7.25}$$

$$\neg\beta \wedge \neg\varepsilon \Rightarrow^{+0.951} \gamma \tag{7.26}$$

$$\neg\beta \wedge \varepsilon \Rightarrow^{+0.981} \gamma \tag{7.27}$$

$$\beta \wedge \neg\varepsilon \Rightarrow^{+0.575} \gamma \tag{7.28}$$

$$\beta \wedge \varepsilon \Rightarrow^{+0.972} \gamma \tag{7.29}$$

$$\text{true} \Rightarrow^{+0.402} \alpha \tag{7.30}$$

$$\beta \Rightarrow^{+1.000} \alpha \tag{7.31}$$

$$\gamma \Rightarrow^{-0.678} \alpha \tag{7.32}$$

Notice that acyclicity of rules obtained this way follows from acyclicity of causal graphs.

7.3 Inference

Since a set of rules, such as the one above, is equivalent to a causal graph, the Lauritzen-Spiegelhalter algorithm provides a method for drawing inferences ('applying the rules'). If this is found to be infeasible because the size of the resulting cliques is too great, then there is another suitable alternative: Monte Carlo propagation [Cor86].

7.3.1 Monte Carlo Propagation

Since the causal graph on which the rules are based has been orientated in the direction of inference, the variables whose values are known for any given case tend to be a complete initial segment of the order. Thus, if the order on the variables is $\alpha_1 < \alpha_2 < \ldots < \alpha_n$, then precisely the variables $\alpha_1, \alpha_2, \ldots, \alpha_i$ are known, for some i. Notice that this is the opposite state of affairs to that when the variables are ordered according to causation. This is why Monte Carlo methods tend to be efficient for inferential representations, but not for descriptive ones.

If the known variables do constitute an initial segment of the order, then Monte Carlo propagation is particularly simple and effective. In order to sample the distribution over the unknown variables, conditioned on the values of the known ones, we start by setting all known variables to their respective values (0 or 1). We then turn to the next variable in the order and compute the probability that it has value 1. Suppose the probability is p. We then assign the value 1 to this variable randomly with probability p, and value 0 with probability $1 - p$. We repeat this step for each successive variable in turn. This entire procedure corresponds to one simulation. We repeat many simulations (say 1000), and count the relative frequency with which each of the unknown variables is assigned value 1.

Example

Continuing the previous example, suppose we observe that $\lambda = 0$ (the lights do not work) and $\varepsilon = 0$ (the engine does not start). We set these variables to their respective values.

$$\begin{aligned} \lambda &:= 0 \\ \varepsilon &:= 0 \end{aligned}$$

The next variable in the order is β. We use Rules 7.23, 7.24 and 7.25 to determine the probability $p(\mathrm{B} \mid \overline{\mathrm{L}}, \overline{\mathrm{E}})$. Rule 7.23 'fires' (i.e. its antecedent evaluates to 1), but Rules 7.24 and 7.25 do not (i.e. their antecedents evaluate to 0). Thus,

$$\begin{aligned} \mathrm{cert}\,(\mathrm{B} \mid \overline{\mathrm{L}}, \overline{\mathrm{E}}) &= -0.932 \oplus 0 \oplus 0 \\ &= -0.932 \end{aligned}$$

So, applying Equation 7.12 (Page 63),

$$\begin{aligned} p(\mathrm{B} \mid \overline{\mathrm{L}}, \overline{\mathrm{E}}) &= (-0.932 + 1)/2 \\ &= 0.034 \end{aligned}$$

which corresponds to the value given in Table 7.1 (Page 60).

Now we draw a random number from the rectangular distribution over the interval $[0, 1)$. Using a random number generator, let us suppose we obtain 0.663. This is not strictly less than 0.034 so we set β to value 0.

$$\beta := 0$$

The next variable in the order is γ. Only Rule 7.26 fires for γ. Thus,

$$\mathrm{cert}\,(\mathrm{C} \mid \overline{\mathrm{E}}, \overline{\mathrm{B}}) = 0.951$$

and

$$p(\mathrm{C} \mid \overline{\mathrm{E}}, \overline{\mathrm{B}}) = 0.975$$

Again we have simply computed the relevant entry in Table 7.1. Suppose the next random number is 0.102. This is strictly less than 0.975, so we set γ to value 1.

$$\gamma := 1$$

Finally, regarding α, only Rules 7.30 and 7.32 fire. Rule 7.31 does not. Therefore,

$$\begin{aligned} \mathrm{cert}\,(\mathrm{A} \mid \overline{\mathrm{B}}, \mathrm{C}) &= 0.402 \oplus -0.678 \\ &= -0.379 \end{aligned}$$

and

$$p(\mathrm{A} \mid \overline{\mathrm{B}}, \mathrm{C}) = 0.310$$

Suppose the next random number is 0.721. Since this is not strictly less than 0.310, so we set α to value 0.

$$\alpha := 0$$

Therefore, at the end of the first simulation, $\beta = 0$, $\gamma = 1$ and $\alpha = 0$.

Shown in Table 7.2 are the results of an actual experiment in which this procedure was repeated many times. As the number of runs increases, the relative frequencies approach the actual conditional probabilities given $\lambda = 0$ and $\varepsilon = 0$.

Table 7.2: Frequencies with which variables were assigned value 1 during increasing numbers of simulations. The last column shows the actual conditional probabilities.

VARIABLE	RUNS			PROB
	100	1000	10000	
β	3	34	342	0.034
γ	96	969	9689	0.969
α	35	346	3434	0.343

This algorithm is both efficient and universally applicable. Notice that it is unaffected by the degree of connectivity of the graph.

Exercise 7.1 *Suppose we intend to use Monte Carlo simulation to approximate the conditional probability associated with a particular unknown variable. If we require the estimate to be correct to two decimal places at a confidence level of 95%, how many simulations do we need to perform?*

7.4 Inferential versus Causal Representations

As we have seen, when knowledge is represented as a set of inference rules, propagation of evidence is computationally more tractable than when knowledge is represented as a descriptive causal graph, unless that graph is sparse.

These are not the only reasons, however, for preferring inference rules as a representation.

7.4.1 Insufficiency of Causation

In practice, training data are generally conditioned on events which are observable effects rather than underlying causes. For example, a medical database is usually conditioned on the event 'the patient seeks medical advice' or some specialization thereof (e.g. 'the patient presents to hospital with acute abdominal pain'). Clearly this excludes persons who have no symptoms, which in turn makes causally unrelated events spuriously dependent.

This means, when constructing a descriptive causal graph, it is not safe to assume that the parents of a given variable are only those which are its direct causes. For example, we had to include an arc from 'Alternator is ok' to 'Carburettor is ok' in Figure 6.1 (Page 45) even though there is no direct causal link. Perhaps we should also have included an arc from 'Battery is charged' for the same reason: if the battery becomes old it may no longer hold a charge, thus making starting difficult. However, causally unrelated and independent faults tend to be mutually exclusive amongst those vehicles showing signs of trouble. Incompleteness of the graph means that the joint distribution it specifies does not correspond exactly to the population, no matter how accurately the individual entries in the conditional probability tables are estimated.

7.4.2 Scarcity of Training Data

When constructing a large expert system, perhaps one encompassing many rare diseases for example, it may be found that the available training data are insufficient for estimating all the required conditional probabilities. There are two ways to proceed.

Numerical Stability

One solution is to make use of other sources of information, such as published results of relevant studies, or subjective estimates elicited from experts. Selection of the kind described above, however, complicates this process; it is difficult to be sure that numerical estimates obtained from one population apply to another one selected in a different way. For example, are probability estimates derived from patients referred to hospital with acute abdominal pain compatible with those for patients who consult their general practitioner with the same symptom?

However, Dawid has shown that conditional probabilities of diseases given symptoms remain stable while those of symptoms given diseases vary according to the way a population is selected [Daw76]. Therefore, we might expect any selection bias in the training data to have less effect on certainty factors derived for inference rules, than on conditional probability tables for a descriptive causal graph.

Variable Reduction

Another solution is to try to reduce the number of parameters to estimate. Suppose our expert system includes three rare symptoms σ_1, σ_2 and σ_3, any one of which is evidence for a particular disease δ. If we adopt a descriptive representation, we will need to specify a conditional probability table for each of the three symptoms. However, if we choose to write inference rules instead, then if data are scarce a single rule will suffice:

$$\sigma_1 \vee \sigma_2 \vee \sigma_3 \Rightarrow^c \delta$$

This reqnires estimation of only one certainty factor (c), and so allows data regarding the three symptoms to be pooled.

7.4.3 Explanations

Lastly, explanations are more easily generated from inference rules than from descriptive knowledge representations. This is because explanations must justify conclusions in terms of the given observations, and this corresponds directly to the orientation of the knowledge expressed in the form of inference rules.

Chapter 8

Alternative Calculi of Uncertainty

Most of the expert systems described in earlier chapters have used probability theory to model uncertainty. However, alternative formalisms have been proposed and are being developed to address perceived weaknesses of probability theory as a calculus of uncertainty. This has caused a certain polarization of opinion, and has led to some friction between proponents of different methods - see the discussion following [Spi84] for example. Nevertheless, we describe here two formalisms, 'approximate reasoning' based on fuzzy sets, and the Dempster-Shafer theory of evidence, which have aroused a great deal of interest and which are relevant to expert systems.

8.1 Fuzzy Sets

Expert opinion is often used as a source of knowledge for expert systems, yet it tends to be imprecise. Recall an example from Chapter 4 (Page 24): most clinicians would readily assert that

inflammation of an abdominal organ usually causes local pain.

Clearly this is an important fact that ought to be useful diagnostically, yet it is imprecise; what exactly is meant by 'inflammation', 'usually', 'local' and 'pain'? The study of fuzzy sets [Zad65] is motivated by the desire to model concepts such as these which are inherently vague.

8.1.1 Paradoxes of Gradual Change

Consider what it is to be bald. Choose an arbitrary bald man; in general, he is not completely hairless, but has noticeably fewer hairs than normal.

Now suppose he grows precisely one additional hair: clearly he is still bald. However, if we continue this chain of reasoning, repeatedly postulating the growth of a single further hair, we will eventually conclude that this individual remains bald no matter how many extra hairs he grows (paradox of *falakros*, 'the bald man' [Car69]). Similar paradoxes derive from heaps of objects that remain heaps even after a single object is removed, and from large numbers that remain large even after they are decremented by one.

The source of the paradox is that the concept 'bald' ('heap', or 'large') is inherently vague, and can be made precise only by arbitrary definition (e.g. 'A person is bald precisely when he has fewer than 10,000 scalp hairs.'). Expressed another way, the set of all bald persons is not precisely defined; it is *fuzzy*. Paradoxes such as those of falakros can be avoided by formal reasoning in terms of fuzzy sets [Gog69].

8.1.2 A Representation for Fuzzy Sets

Crisp Sets

A (conventional) set whose membership is clearly defined is said to be *crisp*. Any crisp set A of type $\mathbf{P}\,\alpha$ is uniquely represented by its characteristic function μ_A which maps each element of α to 1 if it is in set A or 0 otherwise.

$$\begin{aligned} &\mu_A : \alpha \to \{0,1\} \\ &\forall u : \alpha \bullet \mu_A(u) = \begin{cases} 1 & u \in A \\ 0 & u \notin A \end{cases} \end{aligned} \tag{8.1}$$

For example, suppose α is the set of outcomes of rolling a die, and A is the set of even scores.

$$\begin{aligned} \alpha &\mathrel{\widehat{=}} \{1,2,3,4,5,6\} \\ A &\mathrel{\widehat{=}} \{2,4,6\} \end{aligned}$$

Then the characteristic function of A is

$$\mu_A = \{1 \mapsto 0,\ 2 \mapsto 1,\ 3 \mapsto 0,\ 4 \mapsto 1,\ 5 \mapsto 0,\ 6 \mapsto 1\}$$

Fuzzy Sets

Fuzzy sets are represented by generalizing the notion of a characteristic function to allow continuous grades of membership. Thus, in general,

$$\mu_A : \alpha \to [0,1] \tag{8.2}$$

Thus, elements can be only partial members of a fuzzy set. (Notice that crisp sets are just special cases of fuzzy ones.)

For example, regarding dice, suppose HIGH is the (fuzzy) set of outcomes which are 'high' scores. One possible characterization of this set is

$$\mu_{\text{HIGH}} \;\widehat{=}\; \{1 \mapsto 0.0,\ 2 \mapsto 0.0,\ 3 \mapsto 0.1,\ 4 \mapsto 0.5,\ 5 \mapsto 0.9,\ 6 \mapsto 1.0\} \quad (8.3)$$

One possible interpretation for the degree of membership of an element u to a fuzzy set A is the proportion of persons who would agree that u is a member of A.

Exercise 8.1 *Let $\emptyset$ and U be, respectively, the empty and the universal fuzzy subsets of the (crisp) set ω. Define $\emptyset$ and U by means of their characteristic functions. Also, define '?', the fuzziest of all subsets of ω.*

Exercise 8.2 *Explain how the use of fuzzy sets can avoid the paradox of Epimenides.*

8.1.3 Operations on Fuzzy Sets

Union and Intersection

Consider now what it means to take the union and intersection of fuzzy sets. Assume that A, B and C are fuzzy subsets of a universal set ω, of which u and v are arbitrary members.

Operations on fuzzy sets should preserve the familiar properties in the case that the operands are crisp.

$$\mu_A(u) = 0 \wedge \mu_B(u) = 0 \;\Rightarrow\; \mu_{A\cup B}(u) = 0 \wedge \mu_{A\cap B}(u) = 0 \quad (8.4)$$
$$\mu_A(u) = 0 \wedge \mu_B(u) = 1 \;\Rightarrow\; \mu_{A\cup B}(u) = 1 \wedge \mu_{A\cap B}(u) = 0 \quad (8.5)$$
$$\mu_A(u) = 1 \wedge \mu_B(u) = 1 \;\Rightarrow\; \mu_{A\cup B}(u) = 1 \wedge \mu_{A\cap B}(u) = 1 \quad (8.6)$$

Furthermore, the degree of membership of u to the union of A and B should be no less than its degree of membership to either set.

$$\mu_{A\cup B}(u) \geq (\mu_A(u) \sqcup \mu_B(u)) \quad (8.7)$$

where $\sqcup$ denotes the infix binary operator 'maximum'. Similarly, the degree of membership of u to the intersection of A and B should be no more than its degree of membership to either.

$$\mu_{A\cap B}(u) \leq (\mu_A(u) \sqcap \mu_B(u)) \quad (8.8)$$

where $\sqcap$ denotes 'minimum'. Also, even when extended to fuzzy sets, the operations of union and intersection should have their usual algebraic properties of associativity, commutativity, idempotency and distributivity.

$$(A \cup B) \cup C = A \sqcup (B \cup C) \tag{8.9}$$
$$(A \cap B) \cap C = A \cap (B \cap C) \tag{8.10}$$
$$A \cup B = B \cup A \tag{8.11}$$
$$A \cap B = B \cap A \tag{8.12}$$
$$A \cup A = A \tag{8.13}$$
$$A \cap A = A \tag{8.14}$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \tag{8.15}$$

If also we require that $\mu_{A \cup B}$ and $\mu_{A \cap B}$ are continuous and non-decreasing with respect to μ_A and μ_B, then it follows [Bel73] that

$$\mu_{A \cup B}(u) = \mu_A(u) \sqcup \mu_B(u) \tag{8.16}$$

and

$$\mu_{A \cap B}(u) = \mu_A(u) \sqcap \mu_B(u) \tag{8.17}$$

Complement

Clearly complementation should reverse the ordering of the degree of membership of an element to two sets.

$$\mu_A(u) > \mu_B(u) \Rightarrow \mu_{\overline{A}}(u) < \mu_{\overline{B}}(u) \tag{8.18}$$

Furthermore, complementation should be its own inverse.

$$\overline{\overline{A}} = A \tag{8.19}$$

Lastly, if we also require that the effect of complementation on the degrees of membership is symmetric

$$\mu_A(u) + \mu_A(v) = 1 \Rightarrow \mu_{\overline{A}}(u) + \mu_{\overline{A}}(v) = 1 \tag{8.20}$$

then it follows [Gai76] that

$$\mu_{\overline{A}}(u) = 1 - \mu_A(u) \tag{8.21}$$

Exercise 8.3 *Using the definitions of union, intersection and complementation given by Equations 8.16, 8.17 and 8.21, show that De Morgan's Laws hold for fuzzy sets.*

8.1.4 Linguistic Hedges

Zadeh [Zad72] introduced the notion of *linguistic hedges* (e.g. 'very', 'more or less', 'not very') as modifiers of fuzzy sets. The hedge 'very' is defined as replacing the degree of membership of an element by its square.

$$\mu_{\text{very}A}(u) \;\widehat{=}\; (\mu_A(u))^2 \tag{8.22}$$

The inverse, 'fairly', is defined as replacing the degree of membership of an element by its square-root.

$$\mu_{\text{fairly}A}(u) \;\widehat{=}\; \sqrt{\mu_A(u)} \tag{8.23}$$

These operators are said to correspond reasonably well to normal usage of the terms; although, in one study, 'very' seemed to be more of a horizontal translator than a power function [Her76]. Nevertheless, the operators provide a consistent way of constructing more complicated fuzzy sets from fundamental ones.

Example

For example, returning to the previous example of a die, we could define MOD to be the set of 'moderately high' scores, with the assumption that 'moderately high' means 'high, but not very high'. Recall (Equation 8.3) that we have chosen to characterize HIGH by

$$\mu_{\text{HIGH}} = \{1 \mapsto 0.0,\ 2 \mapsto 0.0,\ 3 \mapsto 0.1,\ 4 \mapsto 0.5,\ 5 \mapsto 0.9,\ 6 \mapsto 1.0\}$$

Squaring,

$$\mu_{\text{very HIGH}} = \{1 \mapsto 0.0,\ 2 \mapsto 0.0,\ 3 \mapsto 0.01,\ 4 \mapsto 0.25,\ 5 \mapsto 0.81,\ 6 \mapsto 1.0\}$$

Taking the complement,

$$\mu_{\text{not very HIGH}} = \{1 \mapsto 1.0,\ 2 \mapsto 1.0,\ 3 \mapsto 0.99,\ 4 \mapsto 0.75,\ 5 \mapsto 0.19,\ 6 \mapsto 0.0\}$$

And, taking the intersection with H itself,

$$\mu_{\text{HIGH}\cap(\text{not very HIGH})} = \mu_M = \{1 \mapsto 0.0,\ 2 \mapsto 0.0,\ 3 \mapsto 0.1,\ 4 \mapsto 0.5,\ 5 \mapsto 0.19,\ 6 \mapsto 0.0\}$$

8.1.5 Fuzzy Inference

Fuzzy Relations

A *fuzzy relation* R is a fuzzy subset of $\alpha \times \beta$, where α and β are the domain and range types of R, respectively. The degree of membership $\mu_R(u, v)$ of a pair (u, v) to R is the degree to which R relates u to v.

For example, regarding scores obtained by rolling a die, let LOTLESS be the relation 'is a lot lower than'. Shown below is one possible characterization of this relation.

$$
\begin{array}{l}
\mu_{\text{LOTLESS}} \triangleq \\
\{(1,1)\mapsto 0.0, (1,2)\mapsto 0.1, (1,3)\mapsto 0.5, (1,4)\mapsto 0.9, (1,5)\mapsto 1.0, (1,6)\mapsto 1.0, \\
\ (2,1)\mapsto 0.0, (2,2)\mapsto 0.0, (2,3)\mapsto 0.1, (2,4)\mapsto 0.5, (2,5)\mapsto 0.9, (2,6)\mapsto 1.0, \\
\ (3,1)\mapsto 0.0, (3,2)\mapsto 0.0, (3,3)\mapsto 0.0, (3,4)\mapsto 0.1, (3,5)\mapsto 0.2, (3,6)\mapsto 0.7, \\
\ (4,1)\mapsto 0.0, (4,2)\mapsto 0.0, (4,3)\mapsto 0.0, (4,4)\mapsto 0.0, (4,5)\mapsto 0.1, (4,6)\mapsto 0.3, \\
\ (5,1)\mapsto 0.0, (5,2)\mapsto 0.0, (5,3)\mapsto 0.0, (5,4)\mapsto 0.0, (5,5)\mapsto 0.0, (5,6)\mapsto 0.1, \\
\ (6,1)\mapsto 0.0, (6,2)\mapsto 0.0, (6,3)\mapsto 0.0, (6,4)\mapsto 0.0, (6,5)\mapsto 0.0, (6,6)\mapsto 0.0\ \}
\end{array}
$$

In general, if we know that two variables $x : \alpha$ and $y : \beta$ are related by a relation R, where $\mu_R : (\alpha \times \beta) \to [0, 1]$, and we learn the actual value of x, then we can infer that y lies in the (fuzzy) image of x through R.

For example, suppose x and y are the scores obtained on two consecutive rolls of a die, and we are told that x 'is a lot lower than' y. If then we learn that x is actually 2, adopting the characterization of LOTLESS given above we can conclude that y is a member of the set B where

$$
\begin{array}{rcl}
\mu_B & = & \lambda v : \beta \bullet \mu_{\text{LOTLESS}}(2, v) \\
& = & \{1 \mapsto 0.0,\ 2 \mapsto 0.0,\ 3 \mapsto 0.1,\ 4 \mapsto 0.5,\ 5 \mapsto 0.9,\ 6 \mapsto 1.0\} \\
& = & \mu_{\text{HIGH}}
\end{array}
$$

So we conclude that y is 'high'.

The Compositional Rule of Fuzzy Inference

In the case that all we know of x is, say, that it is either 2 or 3, we obtain two alternative fuzzy restrictions on y, either of which may be appropriate. Naturally, therefore, the resultant set is given by the fuzzy uuion of the two. This principle readily extends to larger (crisp) sets of possible values of x.

Zadeh [Zad73] generalized this principle to the case when the value of x is known only fuzzily. This is expressed as the 'Compositional Rule of Fuzzy

Inference'.

$$\begin{array}{l} x \text{ is } A \\ (x,y) \text{ is } R \\ \hline\hline \\ y \text{ is } A \circ R \end{array}$$

where $A \circ R$ denotes the 'composition' of A with R (i.e. the image of A through R). This is defined

$$\mu_{A \circ R} \mathrel{\widehat{=}} \lambda v : \beta \bullet \bigsqcup_{u:\alpha} (\mu_A(u) \sqcap \mu_R(u,v)) \tag{8.24}$$

where

$$\begin{array}{rcl} \mu_A & : & \alpha \to [0,1] \\ \mu_R & : & (\alpha \times \beta) \to [0,1] \end{array}$$

(The principle of fuzzy inference readily extends to the case where the value of a variable is determined from the values of $n-1$ other variables through a n-ary relation.)

Exercise 8.4 *Suppose that a die is rolled twice. The first score is 'not very high', and, even worse, the second 'is a lot lower than' the first! Use the Compositional Rule of Fuzzy Inference to calculate the (fuzzy) set of possible scores that might have been obtained on the second roll. (Use HIGH and LOTLESS.)*

8.1.6 Production Rules

It may not always be feasible to specify fuzzy relations by explicit enumeration. Production rules provide a convenient shorthand. Although by no means the only way of deriving a fuzzy relation from a production rule, the simplest way is to take the cartesian product of the antecedent and conclusion. Thus, we interpret a rule

$$x \text{ is } A \Rightarrow y \text{ is } B$$

as the proposition

$$(x,y) \text{ is } A \times B$$

The cartesian product of two fuzzy sets is defined by

$$\mu_{A \times B} = \lambda u : \alpha; v : \beta \bullet \mu_A(u) \sqcap \mu_B(v) \tag{8.25}$$

where A is a fuzzy subset of α, and B is a fuzzy subset of β. In the case that we have several rules relating x and y, we simply take the fuzzy union of the corresponding relations.

Exercise 8.5 *Suppose x and y are the scores obtained on two rolls of a die. The assertion that x and y are 'about the same' can be expressed as three production rules.*

$$\begin{aligned} x \text{ is } low &\Rightarrow y \text{ is } low \\ x \text{ is } moderate &\Rightarrow y \text{ is } moderate \\ x \text{ is } high &\Rightarrow y \text{ is } high \end{aligned}$$

Assume that 'high' corresponds to the set HIGH defined above (Equation 8.3, Page 72), and that 'low' corresponds to the set LOW defined below. Take 'moderate' here to mean 'not low and not high'. Reduce the set of three production rules to a single fuzzy relation by taking the union of the corresponding cartesian products. Is this an accurate characterization of 'is about the same'?

$$\mu_{\text{LOW}} \widehat{=} \{1 \mapsto 1.0,\ 2 \mapsto 0.9,\ 3 \mapsto 0.5,\ 4 \mapsto 0.1,\ 5 \mapsto 0.0,\ 6 \mapsto 0.0\}$$

Exercise 8.6 *Continuing Exercise 8.5, use the Compositional Rule of Fuzzy Inference to determine the value of y if x is 'fairly low'.*

8.1.7 Fuzzy Inference and Medical Diagnosis

Fuzzy sets have been claimed by some (e.g. [Adl85]) as 'highly suitable for the formalization of medical processes and concepts'. Others disagree. For example, De Dombal pointed out that the obvious remedy to the vagueness of clinical terminology is to make the terminology more precise [Dom78]. Although some medical expert systems have been built which employ fuzzy sets (for example, [Adl85, Fie90]), the 'min-max' operations that are an integral part of fuzzy reasoning seem inappropriate: medical diagnosis involves accumulation and weighing of evidence. The multiplication and addition operations of probability theory seem intuitively more correct.

8.2 Dempster-Shafer Theory of Evidence

8.2.1 Some Difficulties with Probability Theory

Dempster-Shafer theory [Dem67, Sha76] directly addresses two problematic aspects of the use of probability theory to model belief: the representation of ignorance, and the separation of belief in competing hypotheses. According to probability theory and the 'Principle of Indifference', if we have no reason to choose between two mutually exclusive events then both are assigned equal prior probabilities. No distinction, therefore, is made between

complete ignorance about the relative likelihood of two mutually exclusive events, and *secure knowledge* that the two events are eqnally probable.

Furthermore, it is a consequence of the axioms of probability theory that given any event E, the probability $p(\overline{E})$ is $1 - p(E)$. This means that any evidence for E is necessarily evidence against its complement $\overline{E}$, yet often this seems counter-intuitive. For example, fever is evidence for measles, yet it is also evidence *for* (rather than *against*) the alternative diagnosis of influenza.

8.2.2 Mass Functions

According to Dempster-Shafer theory, rather than assign probability *mass* to individual sample points alone, we can distribute the total mass amongst all subsets of the sample space. Thus $m(E)$ denotes the amount of probability mass that we are prepared to associate with event E, but not with any proper subset of E, on the strength of the available evidence. A *probability mass* function m therefore has the following properties. (In Dempster-Shafer theory Θ is used rather than Ω to denote the sample space, and we follow that convention.)

$$m : \mathbf{P}\,\Theta \rightarrow [0, 1]$$

$$m(\{\}) = 0 \tag{8.26}$$

$$\sum_{E \subseteq \Theta} m(E) = 1 \tag{8.27}$$

Belief

The total probability we have committed to event E (or, in subjective terms, our current belief 'bel(E)' that event E has occurred) is given by the sum of the probability mass associated with all subsets of E.

$$\text{bel}(E) = \sum_{F \subseteq E} m(F) \tag{8.28}$$

Thus the constraint that the probabilities of an event and its complement must sum to 1 has been relaxed to the following.

$$\text{bel}(E) + \text{bel}(\overline{E}) \leq 1 \tag{8.29}$$

As one would expect, we are always certain that the universal event Θ has occurred. This follows directly from Equations 8.27:

$$\text{bel}(\Theta) = 1 \tag{8.30}$$

Some Extreme Mass Functions

In the case m_0 that all the probability mass is assigned to Θ, we have no belief in any more refined event than the universal event itself. This represents complete ignorance.

$$m_0(\mathrm{E}) = \begin{cases} 1 & \mathrm{E} = \Theta \\ 0 & \mathrm{E} \neq \Theta \end{cases} \tag{8.31}$$

Whereas, if all the probability mass is assigned to singleton events (m_1)

$$m_1(\mathrm{E}) = \begin{cases} p(\mathrm{E}) & \#\mathrm{E} = 1 \\ 0 & \#\mathrm{E} \neq 1 \end{cases} \tag{8.32}$$

then the belief in any event is identical to the probability of that event. So a probability function is just a particular kind of belief function.

8.2.3 Dempster's Rule of Combination

Consider now how beliefs based on two sources of evidence can be combined. Suppose a patient has either measles (M), influenza (I) or some other infectious disease (O).

$$\Theta = \{M, I, O\}$$

Furthermore, suppose that, taken individually, two items of clinical evidence induce the mass functions m_1 and m_2 shown below. (Only events with non-zero probability mass have been included in the table.)

			m_2			
			$\{M,I\}$	0.80	$\{I\}$	0.20
m_1	$\{M\}$	0.40	$\{M\}$	0.32	$\{\}$	0.08
	$\{M,O\}$	0.50	$\{M\}$	0.40	$\{\}$	0.10
	$\{M,I,O\}$	0.10	$\{M,I\}$	0.08	$\{I\}$	0.02

The only way that, say, event $\{M, I\}$ can occur is if two events occur simultaneously whose intersection is $\{M, I\}$: in this case, $\{M, I, O\}$ and $\{M, I\}$ with probability masses 0.10 and 0.80, respectively. Therefore, assuming independence, the combined probability mass of $\{M, I\}$ is 0.10×0.80. When an event (such as $\{M\}$ in this case) can occur in more than one way, the sum of the products is calculated; thus the combined probability associated with $\{M\}$ is $0.32 + 0.40$.

There is a difficulty, however, when the intersection of the respective events is empty. A total probability mass of 0.18 is apparently associated

in the example above with the impossible (empty) event. The solution is to set this to zero, and redistribute the probability mass amongst the possible events by normalization (division by 0.82). The combined probability mass function is as follows, all other events mapping to zero probability, where $\oplus$ denotes combination of two mass functions. Values are shown to two decimal places only.

$$\begin{array}{lll} m_1 \oplus m_2 \ (\{M\}) & = & 0.88 \\ m_1 \oplus m_2 \ (\{I\}) & = & 0.02 \\ m_1 \oplus m_2 \ (\{M, I\}) & = & 0.10 \end{array}$$

The general form of Dempster's Rule of Combination is

$$m_1 \oplus m_2(\mathrm{E}) \ \hat{=} \ \begin{cases} 0 & \mathrm{E} = \{\} \\ \dfrac{\sum_{\mathrm{F} \cap \mathrm{G} = \mathrm{E}} m_1(\mathrm{F}) m_2(\mathrm{G})}{\sum_{\mathrm{F} \cap \mathrm{G} \neq \{\}} m_1(\mathrm{F}) m_2(\mathrm{G})} & \mathrm{E} \neq \{\} \end{cases} \tag{8.33}$$

Exercise 8.7 *You are playing a game of ludo. Your opponent rolls the die and then seems very pleased with himself indeed. This is strange because the score on the die looks like only a five, but you can't be 100% sure without your specs.*

Judging your opponent's reaction, you assign subjective probability to the possible events according to the following mass function (m_1).

$$\begin{array}{lll} m_1(\{4,5,6\}) & = & 0.1 \\ m_1(\{5,6\}) & = & 0.2 \\ m_1(\{6\}) & = & 0.7 \end{array}$$

While, the blurred appearance of the die suggests mass function m_2.

$$\begin{array}{lll} m_2(\{4,5,6\}) & = & 0.2 \\ m_2(\{5\}) & = & 0.8 \end{array}$$

Using Dempster's Rule of Combination, calculate how strongly you believe the opponent's score is five, when taking both pieces of evidence into account. Also, how certain are you that the score was more than four?

Chapter 9

Testing and Evaluation of Decision Aids

9.1 Evaluation

Once a new expert system has been designed and implemented, the next stage is to evaluate its performance. In many applications, the user interface is important; it may ultimately determine the acceptability of the system. More fundamental however, and the subject of this chapter, is the ability of the system to arrive at the correct diagnosis and to give the right advice. How then should the diagnostic accuracy of an expert system be assessed? (We answer this question specifically in relation to medical expert systems, although the principles generalize to many other applications.)

9.1.1 Test Data

Retrospective vs Prospective

For training and test purposes we require a collection of case descriptions specifying both the symptoms and the true diagnoses for a random set of patients. There are two ways of collecting such data:

1. *Retrospectively* - Case notes are retrieved from hospital archives, and the relevant information is transcribed onto structured forms.

2. *Prospectively* - Doctors are asked to fill in structured forms themselves at the time patients are seen.

Retrospective data are easily collected, but tend to be of poor quality. Handwritten entries in case notes are often ambiguous, and sometimes illegible. There is a tendency to record only positive findings and key negative

ones. This means that it is often hard to tell whether a symptom or sign was truly absent, or simply not looked for.

Prospective data, on the other hand, are generally of a similar quality to the data that would be entered into the computer if the expert system were in routine use. Nevertheless, it may not be worth the trouble of collecting such data unless the expert system has already been tested with retrospective data, and has shown promise.

Avoiding Bias

When assessing performance, two sources of bias should be avoided. Firstly, the *training set*[1] and the *test set*[2] should be random samples from the *same* population. If not, then misleadingly poor performance figures may be obtained.

Secondly, the training set and the test set should not intersect (except by chance). If training cases are used to test the performance of the system then performance may be deceptively optimistic.

'Leaving-One-Out' Method Often only a limited number (n) of cases are available to the system developer, and the numbers become too small if the set is partitioned into training cases and test cases. If training entails only calculation of numerical parameters, and is computationally efficient, then the 'leaving-one-out' method is applicable. This entails using each case in turn as a test case, and training the expert system afresh each time on the remaining $(n-1)$ cases.

9.1.2 Trial Design

When evaluating a system, the results are more easily interpreted if they can be compared with those of familiar standards such as Bayes' theorem and the unaided clinician himself. The significance of such results can be better assessed also if the computer and the clinician are compared with respect to the same test cases; this allows paired rank tests of statistical significance to be applied.

In some applications, the 'true' diagnosis may he unclear and open to debate. This is the case, for example, regarding selection of antimicrobial therapy. In order to overcome this difficulty, the treatment recommendations of MYCIN (Chapter 4, Page 30) were compared with those of eight

[1] (the set of cases used to derive statistical parameters for the system and to optimize its performance)

[2] (the set of cases used to test the system)

clinicians by a panel of expert judges [Yu79]. The judges were blinded as to which was the computer's advice and which were the clinicians'. In 35% of assessments, MYCIN's recommendation was considered 'unacceptable', but this was marginally better than any of the eight clinicians!

Although an expert system is likely to be useful if it gets the diagnosis right more often than the unaided clinician, the real purpose is to assist the clinician himself to achieve a higher diagnostic accuracy. This is a more difficult hypothesis to test, and where it has been tested there have been some surprising results.

In a multicentre trial of the Leeds program for the diagnosis of abdominal pain (Chapter 2, Page 11), not only did the clinicians' diagnostic accuracy rise from 46% to more than 65% when the computer system was introduced, but a real improvement in patient management was observed [Ada86]. For example, approximately 278 unnecessary operations were avoided during the trial period, and savings in NHS resources amounting to £20m were achieved.

However, introduction of a computer system not only makes available to the clinician an interpretation of his own findings, but also requires that the clinician use a structured data-collection form. This discipline itself is likely to lead to an improvement in diagnostic accuracy. When structured forms alone were used, diagnostic accuracy was found to be about 57% [Ada86]. Furthermore, when clinicians were also given regular feedback about their own performance, they achieved the same diagnostic accuracy after three months as those using the computer program. It is therefore far from clear what contribution, if any, the computer is making [Sut89]. One would hope that this question can be eventually resolved by developing more accurate programs whose contribution is greater and more easily measurable.

9.2 Performance Parameters

Let us now look more closely at the various performance parameters that we can measure.

9.2.1 Diagnostic Accuracy

The *diagnostic accuracy* of an expert system (or clinician, flowchart *etc.*) is the proportion of cases it correctly diagnoses. While this is useful as a single numerical parameter of overall performance, it is generally helpful to know which diseases the system identifies well, and which it identifies poorly.

Consider a single arbitrary disease δ. A case which has δ is said to be a *true positive* if δ is diagnosed by the system, otherwise a *false negative*.

Conversely, a case which does not have δ is said to be a *false positive* if δ is diagnosed, otherwise a *true negative*.

The *sensitivity* of an expert system to a particular disease δ is the proportion of cases which have δ, that the system correctly diagnoses. Conversely, the *specificity* is the proportion of cases which do not have δ that the system correctly diagnoses as not having δ.

For example, Vastola [Vas73] described a flowchart-style program called ASSIGN for deciding whether or not patients have a neurological disorder requiring referral to a neurologist. The program was tested on 308 patients attending a neurological clinic, and ASSIGNS's decision was compared with that of a physician, whose decision was assumed to be correct. The following results were obtained.

183 patients	(TP)	Correctly referred
57 patients	(FP)	Unnecessarily referred
58 patients	(TN)	Correctly discharged
10 patients	(FN)	Wrongly discharged
308 patients		Total

Therefore,

$$\begin{aligned} \text{Sensitivity} &= \text{TP}/(\text{TP} + \text{FN}) = 183/(183+10) \approx 0.95 \\ \text{Specificity} &= \text{TN}/(\text{TN} + \text{FP}) = 58/(58+57) \approx 0.50 \end{aligned}$$

Thus while ASSIGN is quite sensitive (i.e. it correctly refers most cases that require referral), it is not very specific (i.e. it is not very good at identifying the patients who can be safely discharged).

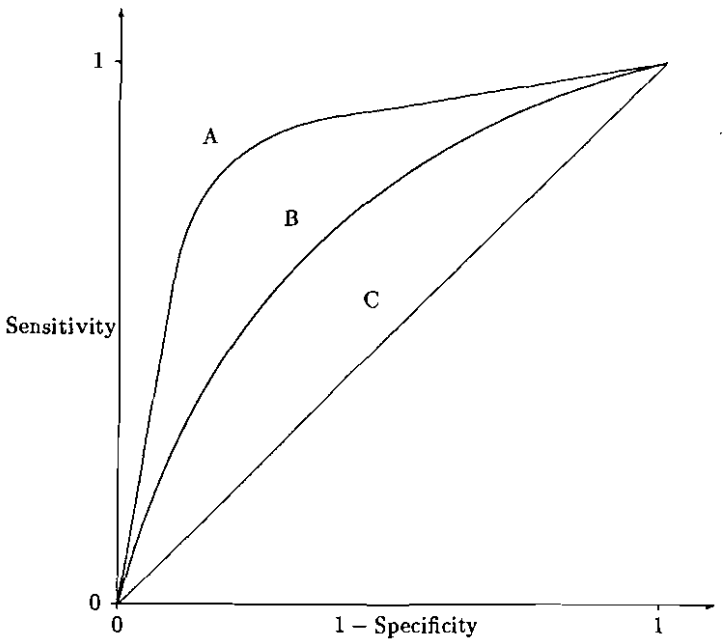
9.2.2 ROC Curves

Generally, expert systems do not make categorical decisions in the way that ASSIGN does. Instead, they calculate a numerical measure of support (e.g. the conditional probability) for a particular hypothesis. A decision as to whether or not to accept the hypothesis is taken (either by the clinician or the expert system itself) by comparing the numerical measure of support for the hypothesis with a pre-determined threshold. By lowering this threshold, the sensitivity of the system can be increased, but only at the expense of a decrease in the specificity.

A graph of sensitivity against the complement of the specificity, as the threshold is altered, provides a way of comparing on a common scale, diverse kinds of expert system designed for the same discrimination task. The

graph (see Figure 9.1) is called the *relative operating characteristic* (ROC) of the system [Swe88]. The area under the curve provides a single numerical measure of discrimination; a value of 1 denotes perfect discrimination, and a value of 0.5 denotes zero discrimination.

Figure 9.1: ROC curves for three expert systems (A, B, C) regarding a particular binary decision problem. System A provides the most discriminatory power, and system C provides none at all.



Exercise 9.1 *Test the flowchart shown in Figure 3.1 (Page 21) on the random sample of 20 cases given in Table 9.1 (Page 87). Sketch the ROC curve, and use it to decide whether the flowchart is a better discriminant than another system which is known to have a sensitivity of 0.90 and a specificity of 0.95.*

9.2.3 Discriminant Matrices

Once a decision threshold has been selected, a clear and comprehensive way of summarizing an expert system's performance with a range of alternative diagnoses is by means of a *diseriminant matrix*. A discriminant matrix itemizes the total number of test cases which actually have disease δ_i but were diagnosed as having disease δ_j, for every i and j.

For example, suppose an expert system is designed to classify patients into exactly one of three categories (A, B and C). Table 9.2 presents some hypothetical test results in the form of a discrimination matrix.

This provides all the information necessary to calculate the sensitivity and specificity of the expert system to each of the disorders. Take diagnosis A, for example:

$$\text{Sensitivity to A} = \frac{23}{23+1+1} = 0.92$$

$$\text{Specificity to A} = \frac{167+18+22+166}{15+167+18+12+22+166} \approx 0.93$$

Reliability

There is another parameter, however, that we have not considered: *reliability*. How reliable is the computer diagnosis 'A' in the above example (Table 9.2)? In other words, when the computer asserts that the diagnosis is A, what is the probability that the computer is correct? This too is easily determined from the discrimination matrix.

$$\text{Reliability of A} = 23/(23+15+12) = 0.46$$

Notice that while the expert system is very sensitive and specific to A, the computer's diagnosis of A is quite unreliable and unsafe.

Exercise 9.2 *From Table 9.2, calculate the sensitivity, specificity and reliability of the expert system with respect to diagnoses 'B' and 'C'.*

Table 9.1: Test data for Exercise 9.1.

Case	δ	σ_1	σ_2	σ_3	σ_4
1	1	1	1	1	0
2	0	0	0	0	1
3	1	1	0	0	1
4	1	1	0	1	1
5	0	0	1	1	0
6	0	1	0	1	0
7	1	1	0	1	1
8	1	1	0	0	1
9	1	1	1	0	0
10	0	1	0	1	1
11	0	1	0	1	1
12	1	1	1	1	0
13	1	1	0	0	0
14	0	1	1	0	1
15	1	1	0	1	0
16	0	1	0	0	0
17	0	1	0	1	0
18	0	1	1	1	1
19	0	0	0	1	0
20	1	1	1	0	1

Table 9.2: Discrimination matrix for diagnoses A, B and C.

		Computer's diagnosis		
		A	B	C
True diagnosis	A	23	1	1
	B	15	167	18
	C	12	22	166

Bibliography

[Ada76] **Adams JB.** A probability model of medical reasoning and the MYCIN model. *Mathematical Biosciences* (1976) **32** 177-86.

[Ada86] **Adams ID, Chan M, Clifford PC, Cooke WM, Dallos V, De Dombal FT, Edwards MH, Hancock DM, Hewett DJ, McIntyre N, Somerville PG, Spiegelhalter DJ, Wellwood J, Wilson DH.** Computer-aided diagnosis of acute abdominal pain: a multicentre study. *British Medical Journal* (1986) **293** 800-4.

[Adl85] **Adlassnig KP.** A fuzzy logical model of computer-assisted medical diagnosis. *Methods of Information in Medicine* (1980) **19** *iii* 141-8.

[Aik83a] **Aikens JS, Kunz JC, Fallat RJ.** PUFF: an expert system for interpretation of pulmonary function data. *Computers and Biomedical Research* (1983) **16** 199-208.

[Aik83b] **Aikens JS.** Prototypical knowledge for expert systems. *Artificial Intelligence* (1983) **20** 163-210.

[Ait76] **Aitchison J, Aitken CGG.** Multivariate binary discrimination by the kernel method. *Biometrika* (1976) **63** *iii* 413-20.

[And82] **Anderson JA.** Logistic discrimination, **in Krishnaiah PR, Kanal LN (eds).** *Handbook of Statistics 2.* North-Holland: Amsterdam, New York, Oxford (1982) 169-91.

[And87] **Andreassen SM, Woldbye M, Falck B, Andersen SK.** MUNIN - a causal probabilistic network for interpretation of electromyographic findings. *Proceedings of the* 10^{th} *International Joint Conference on Artificial Intelligence* (1987) Milan, Italy.

[Bel73] **Bellman RE, Giertz M.** On the analytic formalism of the theory of fuzzy sets. *Information Sciences* (1973) **5** 149-56.

[Ble72] **Bleisch HL.** Computer-based consultation. *American Journal of Medicine* (1972) **53** 285-91.

[Car69] **Cargile J.** The sorites paradox. *British Journal for Philosophy of Science* (1969) **20** 193-202.

[Cha89] **Chard T.** The effect of dependence on the performance of Bayes' theorem: an evaluation using computer simulation. *Computer Methods and Programs in Biomedicine* (1989) **29** *i* 15-9.

[Clo81] **Clocksin WF, Mellish CS.** *Programming in Prolog.* Springer-Verlag: Berlin, Heidelberg, New York (1981).

[Coo89] **Cooper GF.** Current research directions in the development of expert systems based on belief networks. *Applied Stochastic Models and Data Analysis* (1989) **5** *i* 39-52.

[Cor86] **Corlett RA, Todd SJ.** A Monte Carlo approach to uncertain inference, **in** *Artificial Intelligence and its Applications.* **Cohn AG, Thomas JR (eds).** John Wiley & Sons: Chichester (1986) 127-37.

[Cro72] **Croft DJ.** Is computerized diagnosis possible? *Computers and Biomedical Research* (1972) **5** 351-67.

[Daw76] **Dawid AP.** Properties of diagnostic data distributions. *Biometrics* (1976) **32** 647-58.

[Dem67] **Dempster AP.** Upper and lower probabilities induced by a multi-valued mapping. *Annals of Mathematical Statistics* (1967) **38** *ii* 325-39.

[Dom72] **De Dombal FT, Leaper DJ, Staniland JR, McCann AP, Horrocks JC.** Computer-aided diagnosis of acute abdominal pain. *British Medical Journal* (1972) **2** 9-13.

[Dom78] **De Dombal FT.** Medical diagnosis from a clinician's point of view. *Methods of Information in Medicine* (1978) **17** *i* 28-35.

[Dom80] **De Dombal FT.** *Diagnosis of Acute Abdominal Pain.* Churchill Livingstone: Edinburgh, New York, London (1980).

[Dom84] **De Dombal FT.** Clinical decision making and the computer: consultant, expert, or just another test? *British Journal of Healthcare Computing* (1984) **1** 7-14.

[Dud79] **Duda R, Gaschnig J, Hart P.** Model design in the PROSPECTOR consultant system for mineral exploration, **in Michie D (ed).** *Expert Systems in the Micro-Electronic Age.* Edinburgh University Press: Edinburgh (1979) 153-67.

[Edw70] **Edwards DAW.** Flow charts, diagnostic keys, and algorithms in the diagnosis of dysphagia. *Scottish Medical Journal* (1970) **15** *x* 378-85.

[Fie90] **Fieschi M.** *Artificial Intelligence in Medicine.* Chapman and Hall: London, New York, Tokyo, Melbourne, Madras (1990).

[Fox83] **Fox JR, Alvey P.** Computer-assisted medical decision-making. *British Medical Journal* (1983) **287** 742-6.

[Fry78] **Fryback DG.** Bayes' theorem and conditional nonindependence of data in medical diagnosis. *Computers and Biomedical Research* (1978) **11** 423-34.

[Gai76] **Gaines BR.** Foundations of fuzzy reasoning. *International Journal of Man-Machine Studies* (1976) **8** *vi* 623-68.

[Gil73] **Gill PW, Leaper DJ, Guillou PJ, Staniland JR, Horrocks JC, De Dombal FT.** Observer variation in clinical diagnosis - a computer-aided assessment of its magnitude and importance in 552 patients with abdominal pain. *Methods of Information in Medicine* (1973) **12** *ii* 108-13.

[Gog69] **Goguen JA.** The logic of inexact concepts. *Synthese* (1969) **19** 325-73.

[Gol80] **Golumbic MC.** *Algorithmic Graph Theory and Perfect Graphs.* Academic Press: London (1980).

[Goo85] **Goodall A.** *The Guide to Expert Systems.* Learned Information: Oxford, New Jersey (1985).

[Haj85] **Hájek P.** Combining functions for certainty degrees in consulting systems. *International Journal of Man-Machine Studies.* (1985) **22** 59-76.

[Hec86] **Heckerman D.** Probabilistic interpretations for MYCIN's certainty factors, **in Kanal LN, Lemmer JF (eds).** *Uncertainty in Artificial Intelligence.* North-Holland: Amsterdam, New York, Oxford, Tokyo (1986).

[Her76] **Hersh HM, Caramazza A.** A fuzzy set approach to modifiers and vagueness in natural language. *Journal of Experimental Psychology, General* (1976) **105** *iii* 254-76.

[Jac86] **Jackson P.** *Introduction to Expert Systems.* Addison-Wesley: Reading Mass. (1985).

[Jas87] **Jaspers RBM, van der Helm FCT.** Computer-aided diagnosis and treatment of brachial plexus injuries. *Lecture Notes in Medical Informatics* (1987) **33** 237-46.

[Kni85] **Knill-Jones RP.** A formal approach to symptoms in dyspepsia. *Clinics in Gastroenterology* (1985) **14** *iii* 517-29.

[Lau88] **Lauritzen SL, Spiegelhalter DJ.** Local computation with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society (Series B)* (1988) **50** *ii* 157-224.

[Lip58] **Lipkin ML, Hardy JD.** Mechanical correlation of data in differential diagnosis of hematological diseases. *Journal of the American Medical Association* (1958) **166** *ii* 113-25.

[Lip87] **Lippmann RP.** An introduction to computing with neural nets. *IEEE Acoustics, Speech and Signal Processing Society Magazine* (1987) *April* 4-22.

[Lud83] **Ludwig D, Heilbronn D.** The design and testing of a new approach to computer-aided differential diagnosis. *Methods of Information in Medicine* (1983) **22** *iii* 156-66.

[McD82] **McDermott J.** R1: A rule-based configurer of computer systems. *Artificial Intelligence* (1982) **19** *i* 39-88.

[Mic89] **Michie D.** Personal Models of Rationality. *Journal of Statistical Planning and Inference* (1990) **25** 381-99.

[Mil82] **Miller RA, Pople HE, Myers JD.** INTERNIST-1, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine* (1982) **307** *viii* 468-76.

[Nas54] **Nash FA.** Differential diagnosis: an apparatus to assist the logical faculties. *Lancet* (1954) 1 874-5.

[Nea89] **Neapolitan RE.** *Probabilistic Reasoning in Expert Systems.* John Wiley: New York, Chichester, Brisbane, Toronto, Singapore (1989).

[Nor75a] **Norusis MJ, Jacquez JA.** Diagnosis I. Symptom nonindependence in mathematical models for diagnosis. *Computers and Biomedical Research* (1975) **8** 156-72.

[Nor75b] **Norusis MJ, Jacquez JA.** Diagnosis II. Diagnostic models based on attribute clusters: a proposal and comparisons. *Computers and Biomedical Research* (1975) **8** 173-88.

[Pat82] **Patil RS, Szolovits P, Schwartz WB.** Modeling knowledge of the patient in acid-base and electrolyte disorders, in **Szolovits P (ed).** *AAAS Selected Symposium 51.* Bowker Publishing Company: Epping (1982) 191-226.

[Pau76] **Pauker SG, Gorry GA, Kassirer JP, Schwarz WB.** Towards the simulation of clinical cognition. Taking a present illness by computer. *American Journal of Medicine* (1976) **60** 981-96.

[Pea86] **Pearl J.** Fusion, propagation, and structuring in belief networks. *Artificial Intelligence* (1986) **29** 241-88.

[Pea87] **Pearl J.** Evidential reasoning using stochastic simulation of causal models. *Artificial Intelligence* (1987) **32** 245-57.

[Pen87] **Peng Y, Reggia JA.** A probabilistic causal model for diagnostic problem solving - Part I: integrating symbolic causal inference with numeric probabilistic inference. *IEEE Transactions on Systems, Man and Cybernetics* (1987) **17** *ii* 146-62.

[Pop85] **Pople HE.** Evolution of an expert system: from Internist to Caduceus, in **De Lotto I, Stefanelli M (eds).** *Artificial*

Intelligence in Medicine. North-Holland: Amsterdam, New York, Oxford, Tokyo (1985) 179-208.

[Rum86] **Rumelhart DE, Hinton GE, Williams RJ.** Learning representations by back-propagating errors. *Nature* (1986) **323** *ix* 533-6.

[San85] **Sandell HSH, Bourne JR.** Expert systems in medicine: a biomedical engineering perspective. *Critical Reviews in Biomedical Engineering* (1985) **12** 95-129.

[Ser85] **Séroussi B, The ARC & AURC Cooperative Group.** Comparison of several discrimination methods: application to the acute abdomen pain diagnosis. *Lecture Notes in Medical Informatics* (1985) **28** 12-8.

[Ser86] **Séroussi B, The ARC & AURC Cooperative Group.** Computer-aided diagnosis of acute abdominal pain when taking into account interactions. *Methods of Information in Medicine* (1986) **25** *iv* 194-8.

[Sey90] **Seymour DG, Green M, Vaz FG.** Making better decisions: construction of clinical scoring systems by the Spiegelhalter-Knill-Jones approach. *British Medical Journal* (1990) **300** 223-6.

[Sha76] **Shafer G.** *A Mathematical Theory of Evidence*. Princeton University Press: Princeton, New Jersey (1976).

[Sho76] **Shortliffe EH.** *Computer-based medical consultations: MYCIN*. American Elsevier: New York (1976).

[Sho81] **Shortliffe EH, Scott AC, Bischoff MB, Cambell AB, van Melle W, Jacobs CD.** ONCOCIN: an expert system for oncology protocol management. *Proceedings of the* 7^{th} *International Joint Conference on Artificial Intelligence* (1981) 876-81.

[Spi84] **Spiegelhalter DJ, Knill-Jones RP.** Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *Journal of the Royal Statistical Society (Series A)* (1984) **147** *i* 35-77.

[Sta86] **Stanfill C, Waltz D.** Toward memory-based reasoning. *Communications of the ACM* (1986) **29** *xii* 1213-28.

[Sut89] **Sutton GC.** Computer-aided diagnosis: a review. *British Journal of Surgery* (1989) **76** *i* 82-5.

[Swe88] **Swets JA.** Measuring the accuracy of diagnostic systems. *Science* (1988) **240** 1285-93.

[Tar84] **Tarjan RE, Yannakakis MY.** Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM Journal on Computing* (1984) **13** *iii* 566-79.

[Tea81] **Teach RL, Shortliffe EH.** An analysis of physician attitudes regarding computer-based clinical consultation systems. *Computers and Biomedical Research* (1981) **14** *vi* 542-58.

[Tit80] **Titterington DM.** A comparative study of kernel-based density estimates for categorical data. *Technometrics* (1980) **22** *ii* 259-68.

[Tut86] **Tutz G.** An alternative choice of smoothing for kernel-based density estimates in discrete discriminant analysis. *Biometrika* (1986) **73** *ii* 405-11.

[Vas73] **Vastola EF.** Assign: an automated screening system in general neurology. *Computers in Biology and Medicine* (1973) **3** *ii* 107-9.

[Wei78] **Weiss SM, Kulikowski CA, Amarel S, Safir A.** A model-based method for computer-aided medical decision-making. *Artificial Intelligence* (1978) **11** 145-72.

[Win84] **Winter RM, Baraitser M, Douglas JM.** A computerized data base for the diagnosis of rare dysmorphic syndromes. *Journal of Medical Genetics* (1984) **21** *ii* 121-3.

[Yu79] **Yu VL, Fagan LM, Wraith SM, Clancey WJ, Scott AC, Hannigan J, Blum RL, Buchanan BG, Cohen SN.** Antimicrobial selection by computer: a blinded evaln-ation by infectious disease experts. *Journal of the American Medical Association* (1979) **242** 1279-82.

[Zad65] **Zadeh LA.** Fuzzy sets. *Information and Control* (1965) **8** 338-53.

[Zad72] **Zadeh LA.** A fuzzy-set-theoretic interpretation of linguistic hedges. *Journal of Cybernetics* (1972) **2** 4-34.

[Zad73] **Zadeh LA.** Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Transactions on Systems, Man and Cybernetics* (1973) **3** *i* 28-44.

[Zen75] **Zentgraf R.** A note on Lancaster's definition of higher-order interactions. *Biometrika* (1975) **62** *ii* 375-78.