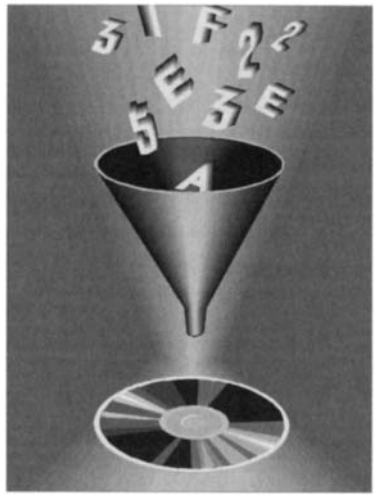
Chapter 1

Probabilistic Informatics



Informatics programs in the United States go back at least to the 1980s when Stanford University offered a Ph.D. in medical informatics. Since that time, a number of informatics programs in other disciplines have emerged at universities throughout the United States. These programs go by various names, including bioinformatics, medical informatics, chemical informatics, music informatics, marketing informatics, etc. What do these programs have in common? To answer that question we must articulate what we mean by the term "informatics." Since other disciplines are usually referenced when we discuss informatics, some define informatics as the application of information technology in the context of another field. However, such a definition does not really tell us the focus of informatics itself. First, we explain what we mean by the term informatics. Then we discuss why we have chosen to concentrate on the probabilistic approach in this book. Finally, we provide an outline of the material that will be covered in the rest of the book.

1.1 What Is Informatics?

In much of western Europe, informatics has come to mean the rough translation of the English "computer science," which is the discipline that studies computable processes. Certainly, there is overlap between computer science programs and informatics programs, but they are not the same. Informatics programs ordinarily investigate subjects such as biology and medicine, whereas computer science programs do not. So the European definition does not suffice for the way the word is currently used in the United States.

To gain insight into the meaning of informatics, let us consider the suffix "-ics," which means the science, art, or study of some entity. For example, "linguistics" is the study of the nature of language, "economics" is the study of the production and distribution of goods, and "photonics" is the study of electromagnetic energy whose basic unit is the photon. Given this, informatics should be the study of information. Indeed, WordNet 2.1 defines informatics as "the science concerned with gathering, manipulating, storing, retrieving and classifying recorded information." To proceed from this definition we need to define the word "information." Most dictionary definitions do not help as far as giving us anything concrete. That is, they define information either as knowledge or as a collection of data, which means we are left with the situation of determining the meaning of knowledge and data. To arrive at a concrete definition of informatics, let's define data, information, and knowledge first.

By **datum** we mean a character string that can be recognized as a unit. For example, the nucleotide G in the nucleotide sequence GATC is a datum, the field "cancer" in a record in a medical data base is a datum, and the field "Gone with the Wind" in a movie data base is a datum. Note that a single character, a word, or a group of words can be a datum depending on the particular application. **Data** then are more than one datum. By **information** we mean the meaning given to data. For example, in a medical data base the data "Joe Smith" and "cancer" in the same record mean that Joe Smith has cancer. By **knowledge** we mean dicta which enable us to infer new information from existing information. For example, suppose we have the following item of knowledge (dictum):[1]

IF the stem of the plant is woody
AND the position is upright
AND there is one main trunk
THEN the plant is a tree.

Suppose further that I am looking at a plant in my backyard and I observe that its stem is woody, its position is upright, and it has one main trunk. Then using the above knowledge item, we can deduce the new information that the plant in my backyard is a tree.

Finally, we define **informatics** as the discipline that applies the methodologies of science and engineering to information. It concerns organizing data

[1] Such an item of knowledge would be part of a rule-based expert system.

into information, learning knowledge from information, learning new information from existing information and knowledge, and making decisions based on the knowledge and information learned. We use engineering to develop the algorithms that learn knowledge from information and that learn information from information and knowledge. We use science to test the accuracy of these algorithms.

Next, we show several examples that illustrate how informatics pertains to other disciplines.

Example 1.1 *(**medical informatics**) Suppose we have a large data file of patients records as follows:*

Patient	Smoking History	Bronchitis	Lung Cancer	Fatigue	Positive Chest X-Ray
1	yes	yes	yes	no	yes
2	no	no	no	no	no
3	no	no	yes	yes	no
⋮	⋮	⋮	⋮	⋮	⋮
10,000	yes	no	no	no	no

*From the **information** in this data file we can use the methodologies of informatics to obtain **knowledge** such as "25% of people with smoking history have bronchitis" and "60% of people with lung cancer have positive chest X-rays." Then from this knowledge and the information that "Joe Smith has a smoking history and a positive chest X-ray" we can use the methodologies of informatics to obtain the new **information** that "there is a 5% chance Joe Smith also has lung cancer."*

Example 1.2 *(**bioinformatics**) Suppose we have long homologous DNA sequences from the human, the chimpanzee, the gorilla, the orangutan, and the rhesus monkey. From this **information** we can use the methodologies of informatics to obtain the new **information** that it is most probable that the human and the chimpanzee are the most closely related of the five species.*

Example 1.3 *(**marketing informatics**) Suppose we have a large data file of movie ratings as follows:*

Person	Aviator	Shall We Dance	Dirty Dancing	Vanity Fair
1	1	5	5	4
2	5	1	1	2
3	4	1	2	1
4	2	5	4	5
⋮	⋮	⋮	⋮	⋮
10,000	1	4	5	5

This means, for example, that Person 1 rated Aviator the lowest (1) and Shall We Dance the highest (5). From the information in this data file, we can develop

a knowledge system that will enable us to estimate how an individual will rate a particular movie. For example, suppose Kathy Black rates Aviator as 1, Shall We Dance as 5, and Dirty Dancing as 5. The system could estimate how Kathy will rate Vanity Fair. Just by eyeballing the data in the five records shown, we see that Kathy's ratings on the first three movies are similar to those of Persons 1, 4, and 5. Since they all rated Vanity Fair high, based on these five records, we would suspect Kathy would rate it high. An informatics algorithm can formalize a way to make these predictions. This task of predicting the utility of an item to a particular user based on the utilities assigned by other users is called ***collaborative filtering****.*

In this book we concentrate on two related areas of informatics, namely financial informatics and marketing informatics. **Financial informatics** involves applying the methods of informatics to the management of money and other assets. In particular, it concerns determining the risk involved in some financial venture. As an example, we might develop a tool to improve portfolio risk analysis. **Marketing informatics** involves applying the methods of informatics to promoting and selling products or services. For example, we might determine which advertisements should be presented to a given Web user based on that user's navigation pattern.

Before ending this section, let's discuss the relationship between informatics and the relatively new expression "'data mining." The term data mining can be traced back to the *First International Conference on Knowledge Discovery and Data Mining* (KDD-95) in 1995. Briefly, **data mining** is the process of extrapolating unknown knowledge from a large amount of observational data. Recall that we said informatics concerns (1) organizing data into information, (2) learning knowledge from information, (3) learning new information from existing information and knowledge, and (4) making decisions based on the knowledge and information learned. So, technically speaking, data mining is a subfield of informatics that includes only the first two of these procedures. However, both terms are still evolving, and some individuals use data mining to refer to all four procedures.

1.2 Probabilistic Informatics

As can be seen in Examples 1.1, 1.2, and 1.3, the knowledge we use to process information often does not consist of IF-THEN rules, such as the one concerning plants discussed earlier. Rather, we only know relationships such as "smoking makes lung cancer more likely." Similarly, our conclusions are uncertain. For example, we feel it is most likely that the closest living relative of the human is the chimpanzee, but we are not certain of this. So ordinarily we must reason under uncertainty when handling information and knowledge. In the 1960s and 1970s a number of new formalisms for handling uncertainty were developed, including certainty factors, the Dempster-Shafer Theory of Evidence, fuzzy logic, and fuzzy set theory. Probability theory has a long history of representing uncertainty in a formal axiomatic way. Neapolitan [1990] contrasts the various

approaches and argues for the use of probability theory.[2] We will not present that argument here. Rather, we accept probability theory as being the way to handle uncertainty and explain why we choose to describe informatics algorithms that use the model-based probabilistic approach.

A **heuristic algorithm** uses a commonsense rule to solve a problem. Ordinarily, heuristic algorithms have no theoretical basis and therefore do not enable us to prove results based on assumptions concerning a system. An example of a heuristic algorithm is the one developed for collaborative filtering in Chapter 11, Section 11.1.

An **abstract model** is a theoretical construct that represents a physical process with a set of variables and a set of quantitative relationships (axioms) among them. We use models so we can reason within an idealized framework and thereby make predictions/determinations about a system. We can mathematically prove these predictions/determinations are "correct," but they are correct only to the extent that the model accurately represents the system. A **model-based algorithm** therefore makes predictions/determinations within the framework of some model. Algorithms that make predictions/determinations within the framework of probability theory are model-based algorithms. We can prove results concerning these algorithms based on the axioms of probability theory, which are discussed in Chapter 2. We concentrate on such algorithms in this book. In particular, we present algorithms that use Bayesian networks to reason within the framework of probability theory.

1.3 Outline of This Book

In Part I we cover the basics of Bayesian networks and decision analysis. Chapter 2 reviews the probability and statistics necessary to understanding the remainder of the book. In Chapter 3 we present Bayesian networks, which are graphical structures that represent the probabilistic relationships among many related variables. Bayesian networks have become one of the most prominent architectures for representing multivariate probability distributions and enabling probabilistic inference using such distributions. Chapter 4 shows how we can learn Bayesian networks from data. A Bayesian network augmented with a value node and decision nodes is called an influence diagram. We can use an influence diagram to recommend a decision based on the uncertain relationships among the variables and the preferences of the user. The field that investigates such decisions is called decision analysis. Chapter 5 introduces decision analysis, while Chapter 6 covers further topics in decision analysis. Once you have completed Part I, you should have a basic understanding of how Bayesian networks and decision analysis can be used to represent and solve real-world problems. Parts II and III then cover applications to specific problems. Part II covers financial applications. Specifically, Chapter 7 presents the basics of investment science and develops a Bayesian network for portfolio risk analysis. In Chapter

[2] Fuzzy set theory and fuzzy logic model a different class of problems than probability theory and therefore complement probability theory rather than compete with it. See [Zadeh, 1995] or [Neapolitan, 1992].

8 we discuss the modeling of real options, which concerns decisions a company must make as to what projects it should pursue. Chapter 9 covers venture capital decision making, which is the process of deciding whether to invest money in a start-up company. In Chapter 10 we show an application to bankruptcy prediction. Finally, Part III contains chapters on two of the most important areas of marketing. First, Chapter 11 shows an application to collaborative filtering/market basket analysis. These disciplines concern determining what products an individual might prefer based on how the user feels about other products. Second, Chapter 12 presents an application to targeted advertising, which is the process of identifying those customers to whom advertisements should be sent.