1

INTRODUCTION

> Where observation is concerned, chance favors only the prepared mind.
> Louis Pasteur

This chapter describes the motivation and reasons for the growing interest in methods for learning (or estimation of empirical dependencies) from data and introduces informally some relevant terminology.

Section 1.1 points out that the problem of learning from data is just one part of the general experimental procedure used in different fields of science and engineering. This procedure is described in detail, with emphasis on the importance of other steps (preceding learning) for overall success. Two distinct goals of learning from data, predictive accuracy (generalization) and interpretation (explanation), are also discussed.

Section 1.2 discusses the relationship between statistical dependency and the notion of causality. It is pointed out that causality cannot be inferred from data analysis alone, but must be demonstrated by arguments outside the statistical analysis. Several examples are presented to support this point.

Section 1.3 describes different types of variables for representing the inputs and outputs of a learning system. These variable types are numeric, categorical, periodic, and ordinal.

Section 1.4 overviews several approaches for describing uncertainty. These include traditional (frequentist) probability corresponding to measurable frequencies,

Learning From Data: Concepts, Theory, and Methods, Second Edition
By Vladimir Cherkassky and Filip Mulier

Bayesian probability quantifying subjective belief, and fuzzy sets for characterization of event ambiguity. The distinction and similarity between these approaches are discussed. The difference between the probability as characterization of event randomness and fuzziness as characterization of the ambiguity of deterministic events is explained and illustrated by examples.

This book is mainly concerned with estimation of *predictive* models from data. This framework, called Predictive Learning, is formally introduced in Chapter 2. However, in many applications data-driven modeling pursues different goals (other than prediction). Several major data analytic methodologies are described and contrasted to Predictive Learning in Section 1.5.

1.1 LEARNING AND STATISTICAL ESTIMATION

Modern science and engineering are based on using *first-principle* models to describe physical, biological, and social systems. Such an approach starts with a basic scientific model (e.g., Newton's laws of mechanics or Maxwell's theory of electromagnetism) and then builds upon them various applications in mechanical engineering or electrical engineering. Under this approach, experimental data (measurements) are used to verify the underlying first-principle models and to estimate some of the model parameters that are difficult to measure directly. However, in many applications the underlying first principles are unknown or the systems under study are too complex to be mathematically described. Fortunately, with the growing use of computers and low-cost sensors for data collection, there is a great amount of data being generated by such systems. In the absence of first-principle models, such readily available data can be used to derive models by estimating useful relationships between a system's variables (i.e., unknown input–output dependencies). Thus, there is currently a paradigm shift from the classical modeling based on first principles to developing models from data.

The need for understanding large, complex, information-rich data sets is common to virtually all fields of business, science, and engineering. Some examples include medical diagnosis, handwritten character recognition, and time series prediction. In the business world, corporate and customer data are becoming recognized as a strategic asset. The ability to extract useful knowledge hidden in these data and to act on that knowledge is becoming increasingly important in today's competitive world.

Many recent approaches to developing models from data have been inspired by the learning capabilities of biological systems and, in particular, those of humans. In fact, biological systems learn to cope with the unknown statistical nature of the environment in a data-driven fashion. Babies are not aware of the laws of mechanics when they learn how to walk, and most adults drive a car without knowledge of the underlying laws of physics. Humans as well as animals also have superior pattern recognition capabilities for tasks such as face, voice, or smell recognition. People are not born with such capabilities, but learn them through

data-driven interaction with the environment. Usually humans cannot articulate the rules they use to recognize, for example, a face in a complex picture. The field of pattern recognition has a goal of building artificial pattern recognition systems that imitate human recognition capabilities. Pattern recognition systems are based on the principles of engineering and statistics rather than biology. There always has been an appeal to build pattern recognition systems that imitate human (or animal) brains. In the mid-1980s, this led to great enthusiasm about the so-called (artificial) neural networks. Even though most neural network models and applications have little in common with biological systems and are used for standard pattern recognition tasks, the biological terminology still remains, sometimes causing considerable confusion for newcomers from other fields. More recently, in the early 1990s, another biologically inspired group of learning methods known as fuzzy systems became popular. The focus of fuzzy systems is on highly interpretable representation of human application-domain knowledge based on the assertion that human reasoning is ''naturally'' performed using fuzzy rules. On the contrary, neural networks are mainly concerned with data-driven learning for good generalization. These two goals are combined in the so-called neurofuzzy systems.

The authors of this book do not think that biological analogy and terminology are of major significance for artificial learning systems. Instead, the book concentrates on using a statistical framework to describe modern methods for learning from data. In statistics, the task of predictive learning (from samples) is called statistical estimation. It amounts to estimating properties of some (unknown) statistical distribution from known samples or training data. Information contained in the training data (past experience) can be used to answer questions about future samples. Thus, we distinguish two stages in the operation of a learning system:

1. Learning/estimation (from training samples)
2. Operation/prediction, when predictions are made for future or test samples

This description assumes that both the training and test data are from the *same* underlying statistical distribution. In other words, this (unknown) distribution is fixed. Specific learning tasks include the following:

- Classification (pattern recognition) or estimation of class decision boundaries
- Regression: estimation of unknown real-valued function
- Probability density estimation (from samples)

A precise mathematical formulation of the learning problem is given in Chapter 2.

There are two common types of the learning problems discussed in this book, known as supervised learning and unsupervised learning. *Supervised* learning is used to estimate an unknown (input, output) mapping from known (input, output) samples. Classification and regression tasks fall into this group. The term ''supervised'' denotes the fact that output values for training samples are known (i.e., provided by a ''teacher'' or a system being modeled). Under the *unsupervised*

learning scheme, only input samples are given to a learning system, and there is no notion of the output during learning. The goal of unsupervised learning may be to approximate the probability distribution of the inputs or to discover ''natural'' structure (i.e., clusters) in the input data. In biological systems, low-level perception and recognition tasks are learned via unsupervised learning, whereas higher-level capabilities are usually acquired through supervised learning. For example, babies learn to recognize (''cluster'') familiar faces long before they can understand human speech. On the contrary, reading and writing skills cannot be acquired in unsupervised manner; they need to be taught. This observation suggests that biological unsupervised learning schemes are based on powerful internal structures (for optimal representation and processing of sensory data) developed through the years of evolution, in the process of adapting to the statistical nature of the environment. Hence, it may be beneficial to use biologically inspired structures for unsupervised learning in artificial learning systems. In fact, a well-known example of such an approach is the popular method known as the self-organizing map for unsupervised learning described in Chapter 6. Finally, it is worth noting here that the distinction between supervised and unsupervised learning is on the level of problem statement only. In fact, methods originally developed for supervised learning can be adapted for unsupervised learning tasks, and vice versa. Examples are given throughout the book.

It is important to realize that the problem of learning/estimation of dependencies from samples is only one part of the general experimental procedure used by scientists, engineers, medical doctors, social scientists, and others who apply statistical (neural network, machine learning, fuzzy, etc.) methods to draw conclusions from the data. The general experimental procedure adopted in classical statistics involves the following steps, adapted from Dowdy and Wearden (1991):

1. State the problem
2. Formulate the hypothesis
3. Design the experiment/generate the data
4. Collect the data and perform preprocessing
5. Estimate the model
6. Interpret the model/draw the conclusions

Even though the focus of this book is on step 5, it is just one step in the procedure. Good understanding of the whole procedure is important for any successful application. No matter how powerful the learning method used in step 5 is, the resulting model would not be valid if the data are not informative (i.e., gathered incorrectly) or the problem formulation is not (statistically) meaningful. For example, poor choice of the input and output variables (steps 1 and 2) and improperly chosen encoding/feature selection (step 4) may adversely affect learning/inference from data (step 5), or even make it impossible. Also, the type of inference procedure used in step 5 may be indirectly affected by the problem formulation in step 2, experiment design in step 3, and data collection/preprocessing in step 4.

Next, we briefly discuss each step in the above general procedure.

Step 1: Statement of the problem. Most data modeling studies are performed in a particular application domain. Hence, domain-specific knowledge and experience are usually necessary in order to come up with a meaningful problem statement. Unfortunately, many recent application studies tend to focus on the learning methods used (i.e., a neural network) at the expense of a clear problem statement.

Step 2: Hypothesis formulation. The hypothesis in this step specifies an unknown dependency, which is to be estimated from experimental data. At this step, a modeler usually specifies a set of input and output variables for the unknown dependency and (if possible) a general form of this dependency. There may be several hypotheses formulated for a single problem. Step 2 requires combined expertise of an application domain and of statistical modeling. In practice, it usually means close interaction between a modeler and application experts.

Step 3: Data generation/experiment design. This step is concerned with how the data are generated. There are two distinct possibilities. The first is when the data generation process is under control of a modeler—it is known as the *designed experiment* setting in statistics. The second is when the modeler cannot influence the data generation process—this is known as the *observational* setting. An observational setting, namely random data generation, is assumed in this book. We will also refer to a random distribution used to generate data (inputs) as a *sampling distribution*. Typically, the sampling distribution is not completely unknown and is implicit in the data collection procedure. It is important to understand how the data collection affects the sampling distribution because such a priori knowledge can be very useful for modeling and interpretation of modeling results. Further, it is important to make sure that past (training) data used for model estimation, and the future data used for prediction, come from the same (unknown) sampling distribution. If this is not the case, then (in most cases) predictive models estimated from the training data alone cannot be used for prediction with the future data.

Step 4: Data collection and preprocessing. This step has to do with both data collection and the subsequent preprocessing of data. In the observational setting, data are usually ''collected'' from the existing databases. Data preprocessing includes (at least) two common tasks: outlier detection/removal and data preprocessing/encoding/feature selection.

Outliers are unusual data values that are not consistent with most observations. Commonly, outliers are due to gross measurement errors, coding/recording errors, and abnormal cases. Such nonrepresentative samples can seriously affect the model produced later in step 5. There are two strategies for dealing with outliers: outlier detection and removal as a part of preprocessing, and development of robust modeling methods that are (by design) insensitive to outliers. Such robust statistical methods (Huber 1981)

are not discussed in this book. Note that there is a close connection between outlier detection (in step 4) and modeling (in step 5).

Data preprocessing includes several steps such as variable scaling and different types of encoding techniques. Such application-domain-specific encoding methods usually achieve dimensionality reduction by providing a small number of informative features for subsequent data modeling. Once again, preprocessing steps should not be considered completely independent from modeling (in step 5): There is usually a close connection between the two. For example, consider the task of variable scaling. The problem of scaling is due to the fact that different input variables have different natural scales, namely their own units of measurement. For some modeling methods (e.g., classification trees) this does not cause a problem, but other methods (e.g., distance-based methods) are very sensitive to the chosen scale of input variables. With such methods, a variable characterizing weight would have much larger influence when expressed in milligrams rather than in pounds. Hence, each input variable needs to be rescaled. Commonly, such rescaling is done independently for each variable; that is, each variable may be scaled by the standard deviation of its values. However, independent scaling of variables can lead to suboptimal representation for many learning methods.

Preprocessing/encoding step often includes selection of a small number of informative features from a high-dimensional data. This is known as *feature selection* in pattern recognition. It may be argued that good preprocessing/data encoding is the most important part in the whole procedure because it provides a small number of informative features, thus making the task of estimating dependency much simpler. Indeed, the success of many application studies is usually due to a clever preprocessing/data encoding scheme rather than to the learning method used. Generally, a good preprocessing method provides an optimal representation for a learning problem, by incorporating a priori knowledge in the form of application-specific encoding and feature selection.

Step 5: Model estimation. Each hypothesis in step 2 corresponds to unknown dependency between the input and output features representing appropriately encoded variables. These dependencies are quantified using available data and a priori knowledge about the problem. The main goal is to construct models for accurate prediction of future outputs from the (known) input values. The goal of predictive accuracy is also known as *generalization* capability in biologically inspired methods (i.e., neural networks). Traditional statistical methods typically use fixed parametric functions (usually *linear in parameters*) for modeling the dependencies. In contrast, more recent methods described in this book are based on much more flexible modeling assumptions that, in principle, enable estimating nonlinear dependencies of an arbitrary form.

Step 6: Interpretation of the model and drawing conclusions. In many cases, predictive models developed in step 5 need to be used for (human) decision making. Hence, such models need to be interpretable in order to be useful

because humans are not likely to base their decisions on complex ''black-box'' models. Note that the goals of accurate prediction and interpretation are rather different because interpretable models would be (necessarily) simple but accurate predictive models may be quite complex. The traditional statistical approach to this dilemma is to use highly interpretable (structured) parametric models for estimation in step 5. In contrast, modern approaches favor methods providing high prediction accuracy, and then view interpretation as a separate task.

Most of this book is on formal methods for estimating dependencies from data (i.e., step 5). However, other steps are equally important for an overall application success. Note that the steps preceding model estimation strongly depend on the application-domain knowledge. Hence, practical applications of learning methods require a *combination* of modeling expertise with application-domain knowledge. These issues are further explored in Section 2.3.4.

As steps 1–4 preceding model estimation are application domain dependent, they cannot be easily formalized, and they are beyond the scope of this book. For this reason, most examples in this book use simulated data sets, rather than real-life data.

Notwithstanding the goal of an accurate predictive model (step 5), most scientific research and practical applications of predictive learning also result in gaining better *understanding* of unknown dependencies (step 6). Such understanding can be useful for

- Gaining insights about the unknown system
- Understanding the limits of applicability of a given modeling method
- Identifying the most important (relevant) input variables that are responsible for the most variation of the output
- Making decisions based on the interpretation of the model.

It should be clear that for real-life applications, meaningful interpretation of the predictive learning model usually requires a good understanding of the issues and choices in steps 1–4 (preceding to the learning itself).

Finally, the interpretation formalism adopted in step 6 often depends on the target audience. For example, standard interpretation methods in statistics (i.e., analysis of variance decomposition) may not be familiar to an engineer who may instead prefer to use fuzzy rules for interpretation.

1.2 STATISTICAL DEPENDENCY AND CAUSALITY

Statistical inference and learning systems are concerned with estimating unknown dependencies hidden in the data, as shown in Fig. 1.1. This procedure corresponds to step 5 in the general procedure described in Section 1.1, but the input and output variables denote preprocessed features of step 4. The goal of predictive learning is

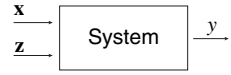

FIGURE 1.1 Real systems often have unobserved inputs **z**.

to estimate unknown dependency between the input $(\mathbf{x})$ and output (y) variables, from a set of past observations of $(\mathbf{x}, y)$ values. In Fig. 1.1, the other set of variables labeled **z** denotes all other factors that affect the outputs but whose values are not observed or controlled. For example, in manufacturing process control, the quality of the final product (output y) can be affected by nonobserved factors such as variations in the temperature/humidity of the environment or small variations in (human) operator actions. In the case of economic modeling based on the analysis of (past) economic data, nonobserved and noncontrolled variables include, for example, the black market economy, as well as quantities that are inherently difficult to measure, such as software productivity. Hence, the knowledge of observed input values $(\mathbf{x})$ does not uniquely specify the outputs (y). This uncertainty in the outputs reflects the lack of knowledge of the unobserved factors $(\mathbf{z})$, and it results in *statistical dependency* between the observed inputs and output(s). The effect of unobserved inputs can be characterized by a conditional probability distribution $p(y|\mathbf{x})$, which denotes the probability that y will occur given the input $\mathbf{x}$.

Sometimes the existence of statistical dependencies between system inputs and outputs (see Fig 1.1) is (erroneously) used to demonstrate cause-and-effect relationship between variables of interest. Such misinterpretation is especially common in social studies and political arguments. We will discuss the difference between statistical dependency and causality and show some examples. The main point is that *causality* cannot be inferred from data analysis *alone*; instead, it must be assumed or *demonstrated* by an argument outside the statistical analysis.

For example, consider (x, y) samples shown in Fig. 1.2. It is possible to interpret these data in a number of ways:

- Variables (x, y) are correlated
- Variable x statistically depends on y, that is, $x = g(y) + \text{error}$

Each formulation is based on different assumptions (about the nature of the data), and each would require different methods for dependency estimation. However,

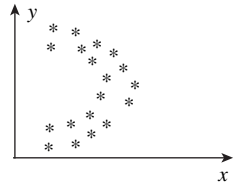

FIGURE 1.2 Scatterplot of two variables that have a statistical dependency.

statistical dependency does not imply causality. In fact, causality is not necessary for accurate estimation of the input–output dependency in either formulation. Meaningful interpretation of the input and output variables, in general, and specific assumptions about causality, in particular, should be made in step 1 or 2 of the general procedure discussed in Section 1.1. In some cases, these assumptions can be *supported* by the data, but they should never be deduced from the data alone.

Next, we consider several common instances of the learning problem shown in Fig. 1.1 along with their application-specific interpretation. For example, in manufacturing process control the causal relationship between controlled input variables and the output quality of the final product is based on understanding of the physical nature of the process. However, it does not make sense to claim causal relationship between person's height and weight, even though statistical dependency (correlation) between height and weight can be easily demonstrated from data. Similarly, it is well known that people in Florida are older (on average) than those in the rest of the United States. This observation does not imply, however, that the climate of Florida causes people to live longer (people just move there when they retire).

The next example is from a real-life study based on the statistical analysis of life expectancy for married versus single men. Results of this study can be summarized as follows: Married men live longer than single men. Does it imply that marriage is (causally) good for one's health; that is, does marriage increase life expectancy? Most likely not. It can be argued that males with physical problems and/or socially deviant patterns of behavior are less likely to get married, and this explains why married men live longer. If this explanation is true, the observed statistical dependency between the input (person's marriage status) and the output (life expectancy) is due to other (unobserved) factors such as person's health and social habits.

Another interesting example is medical diagnosis. Here the observed symptoms and/or test results (inputs $\mathbf{x}$) are used to diagnose (predict) the disease (output y). The predictive model in Fig. 1.1 gives the *inverse* causal relationship: It is the output (disease) that causes particular observed symptoms (input values).

We conclude that the task of learning/estimation of statistical dependency between (observed) inputs and outputs can occur in the following situations:

- Outputs causally depend on the (observed) inputs
- Inputs causally depend on the output(s)
- Input–output dependency is caused by other (unobserved) factors
- Input–output correlation is noncausal
- Any combination of them

Nevertheless, each possibility is specified by the arguments *outside* the data.

The preceding discussion has a negative bearing on naive approaches by some proponents of automatic data mining and knowledge discovery in databases. These approaches advocate the use of automatic tools for discovery of meaningful associations (dependencies) between variables in large databases. However, meaningful dependencies can be extracted from data only if the problem formulation is

meaningful, namely if it reflects a priori knowledge about the application domain. Such commonsense knowledge cannot be easily incorporated into general-purpose automatic knowledge discovery tools.

One situation when a causal relationship can be inferred from the data is when all relevant input factors (affecting the outputs) are observed and controlled in the formulation shown in Fig. 1.1. This is a rare situation for most applications of predictive learning and data mining. As a hypothetical example, consider again the life expectancy study. Let us assume that we can (magically) conduct a controlled experiment where the life expectancy is observed for the two groups of people identical in every (physical and social) respect, except that men in one group get married, and in the other stay single. Then, any different life expectancy in the two groups can be used to infer causality. Needless to say, such controlled experiments cannot be conducted for most social systems or physical systems of practical interest.

1.3 CHARACTERIZATION OF VARIABLES

Each of the input and output variables (or features) in Fig. 1.1 can be of several different types. The two most common types are *numeric* and *categorical*. Numeric type includes real-valued or integer variables (age, speed, length, etc.). A numeric feature has two important properties: Its values have an *order relation* and a *distance relation* defined for any two feature values. In contrast, categorical (or symbolic) variables have neither their order nor distance relation defined. The two values of a categorical variable can be either equal or unequal. Examples include eye color, sex, or country of citizenship. Categorical outputs in Fig. 1.1 occur quite often and represent a class of problems known as pattern recognition, classification, or discriminant analysis. Numeric (real-valued) outputs correspond to regression or (continuous) function estimation problems. Mathematical formulation for classification and regression problems is given in Chapter 2, and much of the book deals with approaches for solving these problems.

A categorical variable with two values can be converted, in principle, to a numeric binary variable with two values (0 or 1). A categorical variable with J values can be converted into J binary numeric variables, namely one binary variable for each categorical value. Representing a categorical variable by several binary variables is known as ''dummy variables'' encoding in statistics. In the neural network literature this method is known as 1-of-J encoding, indicating that each of the J binary variables encodes one feature value.

There are two other (less common) types of variables: periodic and ordinal. A *periodic* variable is a numeric variable for which the distance relation exists, but there is no order relation. Examples are day of the week, month, or year. An *ordinal* variable is a categorical variable for which an order relation is defined but no distance relation. Examples are gold, silver, and bronze medal positions in a sport competition or student ranking within a class. Typically, ordinal variables encode (map) a numeric variable onto a small set of *overlapping* intervals corresponding to

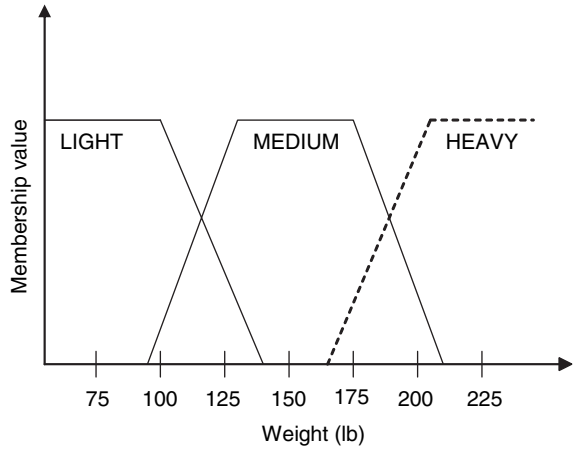

FIGURE 1.3 Membership functions corresponding to different fuzzy sets for the feature *weight*.

the values (labels) of an ordinal variable. Ordinal variables are closely related to linguistic or fuzzy variables commonly used in spoken English, for example, AGE (with values young, middle-aged, and old) and INCOME (with values low, middle-class, upper-middle-class, and rich). There are two reasons why the distance relation for the ordinal or fuzzy values is not defined. First, these values are often subjectively defined by humans in a particular context (hence known as linguistic values). For example, in a recent poll caused by the debate over changes in the U.S. tax code, families with an annual income between $40,000 and $50,000 classified incomes over $100,000 as rich, whereas families with an income of $100,000 defined themselves as middle-class. The second reason is that (even in a fixed context) there is usually no crisp boundary (distinction) between the two closest values. Instead, ordinal values denote overlapping sets. Figure 1.3 shows possible reasonable assignment values for an ordinal feature weight where, for example, the weight of 120 pounds can be encoded as both medium and light weight but with a different degree of membership. In other words, a single (numeric) input value can belong (simultaneously) to *several* values of an ordinal or fuzzy variable.

1.4 CHARACTERIZATION OF UNCERTAINTY

The main formalism adopted in this book (and most other sources) for describing uncertainty is based on the notions of probability and statistical distribution. Standard interpretation/definition of probability is given in terms of (measurable) frequencies, that is, a probability denotes the relative frequency of a random experiment with K possible outcomes, when the number of trials is very large (infinite). This traditional view is known as a *frequentist* interpretation. The $(\mathbf{x}, y)$ observations in the system shown in Fig. 1.1 are sampled from some (unknown) statistical

distribution, under the frequentist interpretation. Then, learning amounts to estimating parameters and/or structure of the unknown input–output dependency (usually related to the conditional probability $p(y|\mathbf{x})$) from the available data. This approach is introduced in Chapter 2, and most of the book describes concepts, theory, and methods based on this formulation. In this section, we briefly mention two other (alternative) ways of describing uncertainty.

Sometimes the frequentist interpretation does not make sense. For example, an economist predicting 80 percent chance of an interest rate cut in the near future does not really have in mind a random experiment repeated, say, 1000 times. In this case, the term probability is used to express a measure of *subjective degree of belief* in a particular outcome by an observer. Assuming events with disjoint outcomes (as in the frequentist interpretation), it is natural to encode subjective beliefs as real numbers between 0 and 1. The value of 1 indicates complete certainty that an event will occur, and 0 denotes complete certainty that an event will not occur. Then, such degrees of belief (provided they satisfy some natural consistency properties) can be viewed as conventional probabilities. This is known as the *Bayesian* interpretation of probabilities. The Bayesian interpretation is often used in statistical inference for specifying a priori knowledge (in the form of subjective prior probabilities) and combining this knowledge with available data via the Bayes theorem. The prior probability encodes our knowledge about the system before the data are known. This knowledge is encoded in the form of a prior probability distribution. The Bayes formula then provides a rule for updating prior probabilities after the data are known. This is known as Bayesian inference or the Bayesian inductive principle (discussed later in Section 2.3.3).

Note that probability is used to measure uncertainty in the *event outcome*. However, an event A itself can either occur or not. This is reflected in the probability identities:

$$P(A) + P(A^c) = 1, \qquad P(AA^c) = 0,$$

where A^c denotes a complement of A, namely $A^c =$ not A, and $P(A)$ denotes the probability that event A will occur.

These properties hold for both the frequentist and Bayesian views of probability. This view of uncertainty is applicable if an observer is capable of unambiguously recognizing occurrence of an event. For example, an ''interest rate cut'' is an unambiguous event. However, in many situations the events themselves occur to a certain subjective degree, and (useful) characterization of uncertainty amounts to specifying a degree of such partial occurrence. For example, consider a feature *weight* whose values light, medium, and heavy correspond to overlapping intervals as shown in Fig. 1.3. Then, it is possible to describe uncertainty of a statement like

Person weighing x pounds is HEAVY

by a number (between 0 and 1), and denoted as $\mu_H(x)$. This is known as a *fuzzy membership function*, and it is used to quantify the degree of subjective belief that the above statement is true, that a person belongs to a (fuzzy) set HEAVY. Ordinal values LIGHT, MEDIUM, and HEAVY are examples of the

fuzzy sets (values), and the membership function is used to specify the degree of partial membership (i.e., of a person weighing x pounds in a fuzzy set HEAVY). As the membership functions corresponding to different fuzzy sets can overlap (see Fig. 1.3), a person weighing 170 pounds belongs to two fuzzy sets, H(eavy) and M(edium), and the sum of the two membership functions does not have to add up to 1. Moreover, a person weighing 170 pounds can belong *simultaneously* to fuzzy set HEAVY and to its complement *not* HEAVY. This type of uncertainty cannot be properly handled using probabilistic characterization of uncertainty, where a person cannot be HEAVY and *not* HEAVY at the same time. A description of uncertainty related to partial membership is provided by fuzzy logic (Zadeh 1965; Zimmerman 1996).

A continuous fuzzy set (linguistic variable) A is specified by the fuzzy membership function $\mu_A(x)$ that gives partial degree of membership of an object x in A. The fuzzy membership function, by definition, has values in the interval $[0, 1]$, to denote partial membership. The value $\mu_A(x) = 0$ means that an object x is not a member of the set A, and the value 1 indicates that x entirely belongs to A.

It is usually assumed that an object is (uniquely) characterized by a scalar feature x, so the fuzzy membership function $\mu_A(x)$ effectively represents a univariate function such that $0 \leq \mu_A(x) \leq 1$. Figure 1.4 illustrates the difference between the fuzzy set (or partial membership) and the traditional "crisp" set membership using different ways to define the concept "boiling temperature" as a function of the water temperature. Note that ordinary (crisp) sets can be viewed as a special case of fuzzy sets with only two (allowed) membership values $\mu_A(x) = 1$ or $\mu_A(x) = 0$.

There are numerous proponents and opponents of the Bayesian and fuzzy characterization of uncertainty. As both the frequentist view and (subjective) Bayesian view of uncertainty can be described by the same axioms of probability, it has lead to the view (common among statisticians) that any type of uncertainty can be fully described by probability. That is, according to Lindley (1987), "probability is the

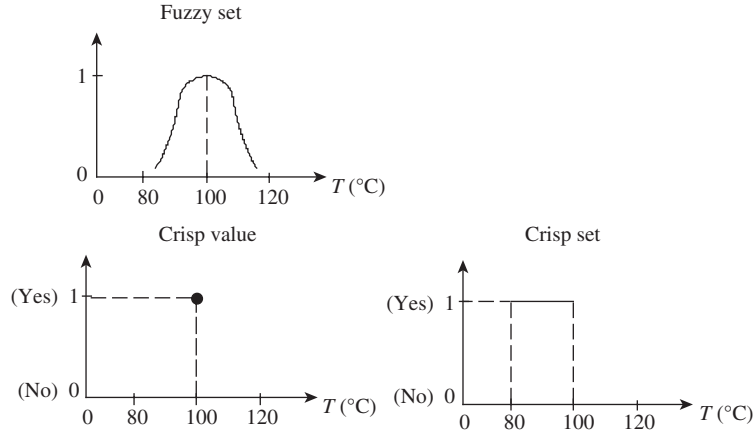

FIGURE 1.4 Fuzzy versus crisp definition of a boiling temperature.

only sensible description of uncertainty and is adequate for all problems involving uncertainty. All other methods are inadequate.'' However, probability describes *randomness*, that is, uncertainty of event occurrence. Fuzziness describes uncertainty related to event *ambiguity*, that is, the subjective degree to which an event occurs. This is an important distinction. Moreover, there are recent claims that probability theory is a special case of fuzzy theory (Kosko 1993).

In the practical context of learning systems, both Bayesian and fuzzy approaches are useful for specification of a priori knowledge about the unknown system. However, both approaches provide *subjective* (i.e., observer-dependent) characterization of uncertainty. Also, there are practical situations where multiple types of uncertainty (frequentist probability, Bayesian probability, and fuzzy) can be combined. For example, a statement ''there is an 80 percent chance of a happy marriage'' describes a (Bayesian) probability of a fuzzy event.

Finally, note that mathematical tools for describing uncertainty (i.e., probability theory and fuzzy logic) have been developed fairly recently, even though humans have dealt with uncertainty for thousands of years. In practice, uncertainty cannot be separated from the notion of *risk* and *risk taking*. In a way, predictive learning methods described in this book can be viewed as a general framework for *risk management*, using empirical models estimated from past data. This view is presented in the last chapter of this book.

1.5 PREDICTIVE LEARNING VERSUS OTHER DATA ANALYTICAL METHODOLOGIES

The growing uses of computers and database technology have resulted in the explosive growth of methods for learning (or estimating) useful models from data. Hence, a number of diverse methodologies have emerged to address this problem. These include approaches developed in classical statistics (multivariate regression/classification, Bayesian methods), engineering (statistical pattern recognition), signal processing, computer science (AI and machine learning), as well as many biologically inspired developments such as artificial neural networks, fuzzy logic, and genetic algorithms. Even though all these approaches often address similar problems, there is little agreement on the fundamental issues involved, and it leads to many heuristic techniques aimed at solving specific applications. In this section, we identify and contrast major methodologies for empirical learning that are often obscured by terminology and minor (technical) details in the implementation of learning algorithms.

At the present time, there are three distinct methodologies for estimating (learning) empirical models from data:

- *Statistical model estimation*, based on extending a classical statistical and function approximation framework (rooted in a density estimation approach) to developing flexible (adaptive) learning algorithms (Ripley 1995; Hastie et al. 2001).

- *Predictive learning*: This approach has originally been developed by practitioners in the field of artificial neural networks in the late 1980s (with no particular theoretical justification). Under this approach, the main focus is on estimating models with good generalization capability, as opposed to estimating ''true'' models under a statistical model estimation methodology. The theoretical framework for predictive learning called Statistical Learning Theory or Vapnik–Chervonenkis (VC) theory (Vapnik 1982) has been relatively unknown until the wide acceptance of its practical methodology called Support Vector Machines (SVMs) in late 1990s (Vapnik 1995). In this book, we use the terms VC theory and predictive learning interchangeably, to denote a methodology for estimating models from data.
- *Data mining*: This is a new practical methodology developed at the intersection of computer science (database technology), information retrieval, and statistics. The goal of data mining is sometimes stated generically as estimating ''useful'' models from data, and this includes, of course, predictive learning and statistical model estimation. However, in a more narrow sense, many data mining algorithms attempt to extract a subset of data samples (from a given large data set) with useful (or interesting) properties. This goal is conceptually similar to *exploratory data analysis* in statistics (Hand 1998; Hand et al. 2001), even though the practical issues are quite different due to huge data size that prevents manual exploration of data (commonly used by statisticians). There seems to be no generally accepted theoretical framework for data mining, so data mining algorithms are initially introduced (by practitioners) and then ''justified'' using formal arguments from statistics, predictive learning, and information retrieval.

There is a significant overlap between these methodologies, and many learning algorithms (developed in one field) have been universally accepted by practitioners in other fields. For example, classification and regression trees (CART) developed in statistics later became very popular in data mining. Likewise, SVMs, originally developed under the predictive learning framework (in VC theory), have been later used (and reformulated) under the statistical estimation framework, and also used in data mining applications. This may give a (misleading) impression that there are only superficial (terminological) differences between these methodologies. In order to understand their differences, we focus on the main assumptions underlying each approach.

Let us relate the three methodologies (statistical model estimation, predictive learning, and data mining) to the general experimental procedure for estimating empirical dependencies from data discussed in Section 1.1. The goal of any data-driven methodology is to estimate (learn) a *useful model* of the unknown system (see Fig. 1.1) from *available data*. We can clearly identify three distinct concepts that help to differentiate between learning methodologies:

1. ''*Useful*'' *model*: There are several commonly used criteria for ''usefulness.'' The first is the prediction accuracy (aka generalization), related to the

capability of the model (obtained using available or training data) to provide accurate estimates (predictions) for future data (from the same statistical population). The second criterion is accurate estimation of the "true" underlying model for data generation, that is, system identification (in Fig. 1.1). Note that correct system identification always implies accurate prediction (but the opposite is not true). The third criterion of the model's "usefulness" relates to its explanatory capabilities; that is, its ability to describe available data in a manner leading to better understanding or interpretation of available data. Note that the goal of obtaining good "descriptive" models is usually quite subjective, whereas the quality of "predictive" models (i.e., generalization) can be objectively evaluated, in principle, using independent (test) data. In the machine learning and neural network literature, predictive methods are also known as "supervised learning" because a predictive model has a unique "response" variable (being predicted by the model). In contrast, descriptive models are referred to as "unsupervised learning" because there is no predefined variable central to the model.

2. *Data set* (used for model estimation): Here we distinguish between the two possibilities. In predictive learning and statistical model estimation, the data set is given explicitly. In data mining, the data set (used for obtaining a useful model) often is not given but must be extracted from a large (given) data set. The term "data mining" suggests that one should search for this data set (with useful properties), which is hidden somewhere in available data.
3. *Formal problem statement* providing (assumed) statistical model for data generation and the goal of estimation (learning). Here we may have two possibilities. That is, when the problem statement is formally well defined and given a priori (i.e., *independent* of the learning algorithm). In predictive learning and statistical model estimation, the goal of learning can be formally stated, that is, there exist mathematical formulations of the learning problem (e.g., see Section 2.1). On the contrary, the field of data mining does not seem to have a single clearly defined formal problem statement because it is mainly concerned with exploratory data analysis.

The existence of the learning problem statement *separate* from the solution approach is critical for meaningful (scientific) comparisons between different learning methodologies. (It is impossible to rigorously compare the performance of methods if each is solving a different problem.) In the case of data mining, the lack of formal problem statement does not suggest that such methods are "inferior" to other approaches. On the contrary, successful applications of data mining to a specific problem may imply that existing learning problem formulations (adopted in predictive learning and statistical model estimation) may not be appropriate for certain data mining applications.

Next, we describe the three methodologies (statistical model estimation, predictive learning, and data mining), in terms of their learning problem statement and solution approaches.

Statistical model estimation is the use of a subset of a population (called a sample) to estimate an underlying statistical model, in order to make conclusions about the entire population (Petrucelli et al. 1999). Classical statistics assumes that the data are generated from some distribution with *known* parametric form, and the goal is to estimate certain properties (of this distribution) useful for specific applications (*problem setting*). Frequently, this goal is stated as density estimation. This goal is achieved by estimating parameters (of unknown distributions) using available data. This goal (probability density estimation) is achieved by maximum-likelihood methods (*solution approach*). The theoretical analysis underlying statistical inference relies heavily on parametric assumptions and asymptotic arguments (i.e., statistically ''optimal'' properties are proved in an asymptotic case when the sample size is large). For example, applying the maximum-likelihood approach to linear regression with normal independent and identically distributed (iid) noise leads to parameter estimation via least squares. In many applications, however, the goal of learning can be stated as obtaining models with good prediction (generalization) capabilities (for future samples). In this case, the approach based on density estimation/function approximation may be suboptimal because it may be possible to obtain good predictive models (reflecting certain properties of the unknown distributions), even when accurate estimation of densities is impossible (due to having only a finite amount of data). Unfortunately, the statistical methodology remains deeply rooted in density estimation/function approximation theoretical framework, which interprets the goal of learning as accurate estimation of the unknown system (in Fig. 1.1), or accurate estimation of the unknown statistical model for data generation, even when application requirements dictate a predictive learning setting. It may be argued that system identification or density estimation is not as prevalent today, because the ''system'' itself is too complex to be identified, and the data are often collected (recorded) automatically for purposes other than system identification. In such real-life applications, often the only meaningful goal is the prediction accuracy for future samples. This may be contrasted to a classical statistical setting where the data are manually collected on a one-time basis, typically under experimental design setting, and the goal is accurate estimation of a given prespecified parametric model.

Predictive learning methodology also has a goal of estimating a useful model using *available training data*. So the problem formulation is often similar to the one used under the statistical model estimation approach. However, the *goal of learning* is explicitly stated as obtaining a model with good prediction (generalization) capabilities for future (test) data. It can be easily shown that estimating a good predictive model is not equivalent to the problem of density estimation (with finite samples). Most practical implementations of predictive learning are based on the idea of obtaining a good predictive model via fitting a set of possible models (given a priori) to available (training) data, aka minimization of empirical risk. This approach has been theoretically described in VC learning theory, which provides general conditions under which various estimators (implementing empirical risk minimization) can generalize well. As noted earlier, VC theory is, in fact, a mathematical theory formally describing the predictive learning methodology.

Historically, many practical predictive learning algorithms (such as neural networks) have been originally introduced by practitioners, but later have been "explained" or "justified" by researchers using statistical model estimation (i.e., density estimation) arguments. Often this leads to certain confusion because such an interpretation creates a (false) impression that the methodology itself (the goal of learning) is based on statistical model estimation. Note that by choosing a simpler but more appropriate problem statement (i.e., estimating relevant properties of unknown distributions under the predictive learning approach), it is possible to make some gains on the inherent stumbling blocks of statistical model estimation (curse of dimensionality, dealing with finite samples, etc.). Bayesian approaches in statistical model estimation can be viewed as an alternative approach to this issue because they try to fix statistical model estimation by including information outside of the data to improve on these stumbling blocks.

Data mining methodology is a diverse field that includes many methods developed under statistical model estimation and predictive learning. There exist two classes of data mining techniques, that is, methods aimed at building "global" models (describing all available data) and "local" models describing some (unspecified) portion of available data (Hand 1998, 1999). According to this taxonomy, "global" data mining methods are (conceptually) identical to methods developed under predictive learning or statistical model estimation. On the contrary, methods for obtaining "local" models aim at discovering "interesting" models for (unspecified) subsets of available data. This is clearly an ill-posed problem, and any meaningful solution will require either (1) exact specification of the portion of the data for which a model is sought or (2) specification of the model that describes the (unknown) subset of available data. Of course, the former leads again to the predictive learning or the statistical model estimation paradigm, and only the latter represents a new learning paradigm. Hence, the data mining paradigm amounts to selecting a portion of data samples (from a given data set) that have certain predefined properties. This paradigm covers a wide range of problems (i.e., data segmentation), and it can also be related to information retrieval, where the "useful" information is specified by its "predefined properties."

This book describes learning (estimation) methods using mainly the predictive learning methodology following concepts developed in VC learning theory. Detailed comparisons between the predictive learning and statistical model estimation paradigms are presented in Sections 3.4.5, 4.5 and 9.9.