2

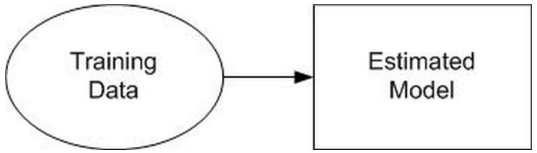
BASIC LEARNING APPROACHES AND COMPLEXITY CONTROL

Making predictions is hard – especially about the future.
Yogi Berra

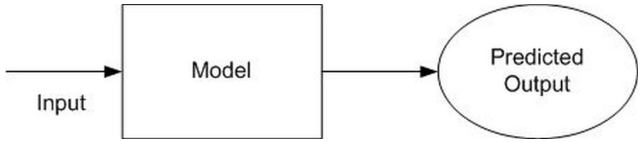
This chapter introduces terminology used in predictive learning, and describes basic learning approaches. These learning approaches follow the interpretation of learning as function estimation from noisy samples, i.e., estimating dependency between several input variables (denoted as input vector $\mathbf{x}$) and an output (or response) y. This estimation is based on past observations of (input, output) samples, known as training data $(\mathbf{x}_i, y_i)$ $i = 1, 2, ...n$. The estimated function (or model) is then used for predicting output values for new (test) inputs. This setting, called *inductive learning*, can be interpreted as an inductive-deductive reasoning process (shown in Fig. 2.1), where:

- *induction step* corresponds to estimation of a general rule (function) from specific instances (or past observations);
- *deduction step* corresponds to prediction of future instances using the general rule.

Predictive Learning
By Vladimir Cherkassky



(a) Induction ~ function estimation from data



(b) Deduction ~ prediction for new inputs

FIGURE 2.1 Learning or function estimation as an induction-deduction process.

The inductive learning setting is used for most machine learning, statistical and data mining algorithms. This chapter describes different types of inductive learning problems called *learning tasks*, and introduces relevant concepts and terminology.

Section 2.1 provides characterization of the training data used for learning. This includes: encoding of input and output variables, data scaling and pre-processing, corresponding to Step 4 in the general experimental procedure in Section 1.4. Section 2.2 describes three common learning tasks: classification, regression and clustering. Section 2.3 presents basic modeling approaches: parametric modeling, non-parametric (or local estimation) methods and data reduction methods. These basic learning approaches form a foundation for more sophisticated learning algorithms described later in the book. Section 2.4 relates generalization (or prediction error) to model complexity control. It also describes a practical approach to complexity control via resampling. Finally, Section 2.5 describes an application example illustrating resampling for complexity control and estimation of prediction error of a learning method.

This chapter introduces important concepts, such as learning task, prediction error, model complexity, complexity control and resampling, using examples and informal arguments. A more formal (mathematical) description of these concepts is given later in Chapter 4. Hence, this

chapter, along with Chapter 4, forms a conceptual basis for understanding various learning algorithms described later in this book.