10

METHODOLOGICAL ISSUES AND CASE STUDIES

> Truth is not as benevolent as its semblance is malicious.
> La Rochefoucauld

Earlier chapters of this book presented various learning algorithms, which estimate predictive models from available data. In addition to sound learning methods, real-life applications require a great deal of common sense and good engineering. Practitioners usually face the task of choosing the 'right' learning algorithm for their application. This is a non-trivial task, which requires the combined knowledge of an application domain *and* data modeling methodology. Few people have such knowledge, so data modelers need to interact closely with application domain experts. This initial stage is critical for the overall success, but it cannot be formalized. The interactions between experts are usually hindered by the terminological and conceptual barriers. That is, professionals from different fields simply do not speak the same

Predictive Learning
By Vladimir Cherkassky

language. Often, application domain experts have very naïve and optimistic expectations about data-analytic modeling. Such expectations are usually promoted by software companies marketing their tools. Similarly, machine learning researchers pursue their own agenda and invent new algorithms that have a shelf life of a few years.

This book started with a general description of the problem of extracting empirical knowledge from data in Chapter 1. As discussed in Section 1.4, data-analytic modeling involves three main steps:

1. Formalization of application problem;
2. Model estimation (from available data);
3. Interpretation and understanding of a data-analytic model.

Earlier technical chapters in this book focused on model estimation methods in Step 2. As explained in Chapters 1 and 2, this step can be viewed as a problem of function estimation from noisy samples. This learning problem is inherently difficult or ill-posed, so it is important to adopt a sound methodological framework. This book argues in favor of predictive learning or VC-theoretical framework, which is used to design and understand various learning methods. This chapter is concerned mainly with the informal Steps 1 and 3 in the data modeling process. These steps should be also related to the framework of predictive learning. Specifically,

- All learning methods follow certain learning problem settings, such as a standard inductive learning setting, or a particular nonstandard formulation. Formalization of application domain requirements in Step 1 leads to selection of an appropriate learning setting.
- Sound interpretation of a data-analytic model in Step 3 requires a clear understanding of the methodological assumptions used to estimate this model. This point becomes particularly important for high-dimensional problems.

We emphasize the relevance of predictive learning methodology for the whole data modeling process. Overall, this methodology adopts a cautious attitude towards data-analytic modeling. It is fairly easy to discover interesting models (patterns) from finite data, yet it may be quite difficult to estimate robust predictive models. In contrast, many practitioners and researchers often adopt a more opportunistic attitude, as evident in the term 'data mining'. The data mining methodology advocates applying various data modeling techniques to available data, and then reporting the best modeling results. Often, the learning algorithms are applied without clear specification of the learning problem setting. This may lead to an over-optimistic interpretation of the data modeling results.

Section 10.1 provides general guidance for the problem formalization and also argues that the predictive methodology can safeguard against several human modeling biases common in many application studies. Section 10.2 describes the problem of feature selection which is often regarded as a separate step preceding model estimation. Sections 10.3, 10.4 and 10.5 present application case studies. These studies illustrate several important practical design aspects of predictive learning, so they skip the details of learning algorithms. Section 10.6 elaborates on the philosophical aspects of predictive data modeling.

Relevant earlier parts of the book include: motivation for predictive approach in Chapter 1, experimental procedure for data-driven modeling (in Section 1.4), descriptions of standard inductive learning in Chapter 2 (Section 2.2) and non-standard learning formulations in Chapter 9. This material can be regarded as a prerequisite for this chapter.