9

NONSTANDARD LEARNING APPROACHES

Science starts from problems, and not from observations.

Karl Popper

All learning methods presented so far in this book assume a *standard inductive learning* formulation, where the goal is to estimate a predictive model from finite training data. That is, learning from data amounts to estimation of a function of input variables (denoted as vector $\mathbf{x}$). This estimation is based on past observations of (input, output) samples, or training data $(\mathbf{x}_i, y_i)$ $i = 1,2,...n$. The function (or model) is then used for prediction of output values for new (test) inputs $\hat{y} = f(\mathbf{x})$. This setting can be also interpreted as an inductive-deductive reasoning process (as shown in Fig. 2.1), and its philosophical interpretations were discussed in Chapter 3.

While this standard inductive learning setting is quite general and commonly adopted in most machine learning, statistical and data mining algorithms, it should not be taken for granted. As argued in Section 2.2, the main assumptions (behind this setting) may not hold for some applications. Also, modern applications often involve learning with sparse high-dimensional data, where the number of samples is smaller than the number of input features. In micro-array data analysis, technologies have been designed to measure the gene expression levels of tens of thousands of genes in a single experiment. However, the sample size in each data set is typically small, ranging from tens to low hundreds. Similarly, in brain imaging studies the dimensionality of the input data vector is much larger than the sample size. Such sparse high-dimensional problems are very ill-posed and represent new challenges for learning methods. There are two main approaches for improving generalization with high-dimensional data. The traditional approach is to adopt the standard inductive learning setting and introduce a priori knowledge about the properties of application data or the properties of a 'good' model. Typical examples include:

- preprocessing and feature extraction techniques that incorporate application-domain knowledge into the selection of a small number of informative features;
- selection of good kernels in SVM methods;
- specification of 'informative' prior distributions in statistical (Bayesian) methods.

Another approach is to adopt *non-standard learning settings*, where a priori knowledge is used to modify the learning problem formulation itself. This chapter emphasizes the importance of the learning setting, vs. the notion of a learning method. The learning method is a constructive procedure (algorithm) for implementing a particular learning setting. For example, all statistical, neural network and machine learning methods for classification and regression, presented in Chapters 2 – 8, implement standard inductive learning formulation.

In order to illustrate several non-standard methodologies, consider the task of hand-written digit recognition. For simplicity, assume a binary classification problem of discriminating between digits '5' and '8'. Under the standard inductive learning setting, one has to estimate a binary classifier from available labeled examples of handwritten digits '5' and '8'. Then the prediction accuracy of a classifier is measured using an independent test set. Consider the following modifications or extensions of this setting:

1) Assume that the *unlabeled test data* is also available during training. Then incorporating this unlabeled data into the process of learning can improve the prediction accuracy of a classifier. This leads to

Predictive Learning
By Vladimir Cherkassky

transductive and semi-supervised learning settings, which are discussed in Section 9.1.

2) Assume that additional data in the form of 'other digits' is available during training. These are unlabeled digits other than labeled training samples '5' and '8'. Clearly, these examples of 'other digits' contain some additional information about handwritten digits, however, they do not belong to the same distribution as training data. These unlabeled samples are called the Universum. Including these Universum samples into learning can potentially improve generalization. This new learning setting, known as Learning through Contradiction, or Universum learning, is discussed in Section 9.2.
3) Assume that labeled training data (e.g., digits '5' and '8') is provided by t different persons. So the training data may be partitioned into t groups, accordingly. This additional information can be used to improve generalization performance. One possibility is to assume that both the training and test data are generated by t persons, and that the group label is known for both training and test data. This leads to estimation of t related classifiers, which is known as Multi-Task Learning (MTL), as discussed in Section 9.4.
4) Another possibility is that the additional group label information is *known only* for the training data (but not known for test inputs). In this case, the goal of learning is to estimate a single classifier. However, utilizing additional information (about groups in the training data) can improve generalization. This leads to a setting called Learning Using Privileged Information (LUPI) discussed in Section 9.3. In general, the LUPI approach assumes that training data is provided in the form $(\mathbf{x}_i, \mathbf{x}_i^*, y_i)\ i = 1,2,...n$, where $\mathbf{x}_i$ denotes the input features (e.g., pixels in the character recognition problem), and $\mathbf{x}_i^*$ denotes additional information (in our example, a person or group label). This additional information is called *hidden* or *privileged*, because it is available only for training samples, so that a classification model $\hat{y} = f(\mathbf{x})$ is estimated in the original input space.

Such non-standard learning settings reflect properties of real-life applications, and can result in improved generalization, relative to standard inductive learning. However, these new methodologies are more complex, and their advantages and limitations are not well understood. In the following sections, we provide detailed descriptions of several important non-standard learning settings, followed by their

mathematical optimization formulations. These new optimization formulations can be viewed as extensions of standard support vector machines, so earlier material on SVM classifiers in Chapter 7 is a necessary prerequisite. The SVM-style framework is naturally adopted for describing new learning settings for both conceptual and technical reasons. Conceptually, these new learning formulations illustrate how additional a priori knowledge can be used to specify a new type of structure (or complexity ordering) on a set of admissible models. This follows the general Structural Risk Minimization (SRM) approach in VC-theory. Technically, additional a priori knowledge is encoded as additional constraints on the complexity of admissible models. Hence, all optimization formulations for non-standard settings in Sections 9.1-9.4 include the usual margin-based SVM capacity control, and also some additional terms and constraints. These extra terms reflect specific properties of new formulations, and result in additional capacity control. Optimization software for solving some of these new SVM-like formulations may require only minor modifications of existing software for standard SVM.

Finally, Section 9.5 presents summary and philosophical interpretation of non-standard learning formulations. Section 9.6 contains bibliographic notes and pointers for further reading.