4
STATISTICAL LEARNING THEORY

So, so you think you could tell, Heaven from Hell, blue skies from pain?
Pink Floyd

This chapter describes Statistical Learning Theory, or Vapnik-Chervonenkis (VC) theory which provides mathematical and conceptual basis for predictive learning. The VC theory provides conditions under which learning methods based on the notion of data fitting, or empirical risk minimization (ERM), can generalize for future samples. We have already informally introduced such learning approaches based on data fitting, in Chapter 2, and also demonstrated the effect of model complexity on generalization.

Statistical Learning Theory (SLT) formally describes the inductive learning problem setting, and gives precise mathematical meaning to the notions of prediction risk, empirical risk and model complexity. SLT provides analysis of the ERM inductive principle, leading to *generalization bounds*, which relate unknown prediction risk to known empirical risk and a new measure of model complexity called the VC-dimension. From these generalization bounds, the theory then develops a new inductive principle, called Structural Risk Minimization (SRM), which provides theoretical justification for model complexity control. This chapter does not contain mathematical derivations and proofs. Instead, we present the main concepts and results developed in VC-theory, and try to relate these concepts to:

- philosophical ideas introduced in Chapter 3,
- practical aspects of machine learning methods, such as model selection.

Understanding of a theoretical framework is very important for practitioners who apply machine learning methods to their application data. This is because all learning problems are inherently *ill-posed*. So generalization from finite samples depends on a priori knowledge and various assumptions. A learning theory provides a theoretical framework for solving ill-posed estimation problems. Most statistical or machine learning algorithms used for model estimation usually follow a theoretical framework. There is no algorithm that provides superior generalization performance for all data sets. A sound learning theory helps to understand why learning methods can generalize well, and improve their performance. Further, clear theoretical framework helps to understand the limitations of various learning methods described in later chapters.

This chapter describes abstract mathematical concepts that some readers may find difficult to follow. So it is not for casual reading. Also, the material presented in this chapter is closely related with ideas and examples in Chapter 2 that can be regarded as prerequisite.

Predictive Learning
By Vladimir Cherkassky