5 Elements of Statistical Learning Theory

We now give a more complete exposition of the ideas of statistical learning theory, which we briefly touched on in Chapter 1. We mentioned previously that in order to learn from a small training set, we should try to *explain* the data with a model of *small* capacity; we have not yet justified *why* this is the case, however. This is the main goal of the present chapter.

Overview

We start by revisiting the difference between risk minimization and empirical risk minimization, and illustrating some common pitfalls in machine learning, such as overfitting and training on the test set (Section 5.1). We explain that the motivation for empirical risk minimization is the law of large numbers, but that the classical version of this law is not sufficient for our purposes (Section 5.2). Thus, we need to introduce the statistical notion of *consistency* (Section 5.3). It turns out that consistency of learning algorithms amounts to a law of large numbers, which holds uniformly over all functions that the learning machine can implement (Section 5.4). This crucial insight, due to Vapnik and Chervonenkis, focuses our attention on the set of attainable functions; this set must be restricted in order to have any hope of succeeding. Section 5.5 states probabilistic bounds on the risk of learning machines, and summarizes different ways of characterizing precisely how the set of functions can be restricted. This leads to the notion of *capacity concepts*, which gives us the main ingredients of the typical generalization error bound of statistical learning theory. We do not indulge in a complete treatment; rather, we try to give the main insights to provide the reader with some intuition as to how the different pieces of the puzzle fit together. We end with a section showing an example application of risk bounds for model selection (Section 5.6).

Prerequisites

The chapter attempts to present the material in a fairly non-technical manner, providing intuition wherever possible. Given the nature of the subject matter, however, a limited amount of mathematical background is required. The reader who is not familiar with basic probability theory should first read Section B.1.

5.1 Introduction

Let us start with an example. We consider a regression estimation problem. Suppose we are given empirical observations,

$$(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \mathcal{Y}, \tag{5.1}$$