Introduction

Among the possible approaches to pattern recognition, statistical learning theory has received a lot of attention in the last few years. Although a realistic pattern recognition scheme involves data pre-processing and post-processing that need a theory of their own, a central role is often played by some kind of supervised learning algorithm. This central building block is the subject we are going to analyse in these notes.

Accordingly, we assume that we have prepared in some way or another a *sample* of N labelled patterns $(X_i, Y_i)_{i=1}^N$, where X_i ranges in some pattern space $\mathcal{X}$ and Y_i ranges in some finite label set $\mathcal{Y}$. We also assume that we have devised our experiment in such a way that the couples of random variables (X_i, Y_i) are independent (but not necessarily equidistributed). Here, randomness should be understood to come from the way the statistician has planned his experiment. He may for instance have drawn the X_is at random from some larger population of patterns the algorithm is meant to be applied to in a second stage. The labels Y_i may have been set with the help of some external expertise (which may itself be faulty or contain some amount of randomness, so we do not assume that Y_i is a function of X_i, and allow the couple of random variables (X_i, Y_i) to follow any kind of joint distribution). In practice, patterns will be extracted from some high dimensional and highly structured data, such as digital images, speech signals, DNA sequences, etc. We will not discuss this pre-processing stage here, although it poses crucial problems dealing with segmentation and the choice of a representation. The aim of supervised classification is to choose some classification rule $f : \mathcal{X} \to \mathcal{Y}$ which predicts Y from X making as few mistakes as possible on average.

The choice of f will be driven by a suitable use of the information provided by the sample $(X_i, Y_i)_{i=1}^N$ on the joint distribution of X and Y. Moreover, considering all the possible measurable functions f from $\mathcal{X}$ to $\mathcal{Y}$ would not be feasible in practice and maybe more importantly not well founded from a statistical point of view, at least as soon as the pattern space $\mathcal{X}$ is large and little is known in advance about the joint distribution of patterns X and labels Y. Therefore, we will consider parametrized subsets of classification rules $\{f_\theta : \mathcal{X} \to \mathcal{Y}\,;\, \theta \in \Theta_m\}$, $m \in M$, which may be grouped to form a big parameter set $\Theta = \bigcup_{m \in M} \Theta_m$.

The subject of this monograph is to introduce to statistical learning theory, and more precisely to the theory of supervised classification, a number of technical tools akin to statistical mechanics and information theory, dealing with the concepts of entropy and temperature. A central task will in particular be to control the mutual information between an estimated parameter and the observed sample. The focus will not be directly on the description of the data to be classified, but on the description of the classification rules. As we want to deal with high dimensional data, we will be bound to consider high dimensional sets of candidate classification rules, and will analyse them with tools very similar to those used in statistical mechanics