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

to describe particle systems with many degrees of freedom. More specifically, the sets of classification rules will be described by Gibbs measures defined on parameter sets and depending on the observed sample value. A Gibbs measure is the special kind of probability measure used in statistical mechanics to describe the state of a particle system driven by a given energy function at some given temperature. Here, Gibbs measures will emerge as minimizers of the average loss value under entropy (or mutual information) constraints. Entropy itself, more precisely the Kullback divergence function between probability measures, will emerge in conjunction with the use of exponential deviation inequalities: indeed, the log-Laplace transform may be seen as the Legendre transform of the Kullback divergence function, as will be stated in Lemma 1.1.3 (page 4).

To fix notation, let $(X_i, Y_i)_{i=1}^N$ be the canonical process on $\Omega = (\mathcal{X} \times \mathcal{Y})^N$ (which means the coordinate process). Let the pattern space be provided with a sigma-algebra $\mathcal{B}$ turning it into a measurable space $(\mathcal{X}, \mathcal{B})$. On the finite label space $\mathcal{Y}$, we will consider the trivial algebra $\mathcal{B}'$ made of all its subsets. Let $\mathcal{M}_+^1\big[(\mathcal{X} \times \mathcal{Y})^N, (\mathcal{B} \otimes \mathcal{B}')^{\otimes N}\big]$ be our notation for the set of probability measures (i.e. of positive measures of total mass equal to 1) on the measurable space $\big[(\mathcal{X}\times\mathcal{Y})^N, (\mathcal{B}\times\mathcal{B}')^{\otimes N}\big]$. Once some probability distribution $\mathbb{P} \in \mathcal{M}_+^1\big[(\mathcal{X}\times\mathcal{Y})^N, (\mathcal{B}\otimes\mathcal{B}')^{\otimes N}\big]$ is chosen, it turns $(X_i, Y_i)_{i=1}^N$ into the canonical realization of a stochastic process modelling the observed sample (also called the training set). We will assume that $\mathbb{P} = \bigotimes_{i=1}^N P_i$, where for each $i = 1, \ldots, N$, $P_i \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}, \mathcal{B} \otimes \mathcal{B}')$, to reflect the assumption that we observe independent pairs of patterns and labels. We will also assume that we are provided with some indexed set of possible classification rules

$$\mathcal{R}_\Theta = \big\{ f_\theta : \mathcal{X} \to \mathcal{Y}; \theta \in \Theta \big\},$$

where $(\Theta, \mathcal{T})$ is some measurable index set. Assuming some indexation of the classification rules is just a matter of presentation. Although it leads to heavier notation, it allows us to integrate over the space of classification rules as well as over Ω, using the usual formalism of multiple integrals. For this matter, we will assume that $(\theta, x) \mapsto f_\theta(x) : (\Theta \times \mathcal{X}, \mathcal{B} \otimes \mathcal{T}) \to (\mathcal{Y}, \mathcal{B}')$ is a measurable function.

In many cases, as already mentioned, $\Theta = \bigcup_{m \in M} \Theta_m$ will be a finite (or more generally countable) union of subspaces, dividing the classification model $\mathcal{R}_\Theta = \bigcup_{m \in M} \mathcal{R}_{\Theta_m}$ into a union of sub-models. The importance of introducing such a structure has been put forward by V. Vapnik, as a way to avoid making strong hypotheses on the distribution $\mathbb{P}$ of the sample. If neither the distribution of the sample nor the set of classification rules were constrained, it is well known that no kind of statistical inference would be possible. Considering a family of sub-models is a way to provide for adaptive classification where the choice of the model depends on the observed sample. Restricting the set of classification rules is more realistic than restricting the distribution of patterns, since the classification rules are a processing tool left to the choice of the statistician, whereas the distribution of the patterns is not fully under his control, except for some planning of the learning experiment which may enforce some weak properties like independence, but not the precise shapes of the marginal distributions P_i which are as a rule unknown distributions on some high dimensional space.

In these notes, we will concentrate on general issues concerned with a natural measure of risk, namely the *expected error rate* of each classification rule f_θ, expressed as

$$R(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}\big[f_\theta(X_i) \neq Y_i\big]. \tag{0.1}$$

As this quantity is unobserved, we will be led to work with the corresponding *empirical error rate*

$$r(\theta,\omega)=\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\big[f_\theta(X_i)\neq Y_i\big]. \tag{0.2}$$

This does not mean that practical learning algorithms will always try to minimize this criterion. They often on the contrary try to minimize some other criterion which is linked with the structure of the problem and has some nice additional properties (like smoothness and convexity, for example). Nevertheless, and independently of the precise form of the estimator $\widehat{\theta}:\Omega\to\Theta$ under study, the analysis of $R(\widehat{\theta})$ is a natural question, and often corresponds to what is required in practice.

Answering this question is not straightforward because, although $R(\theta)$ is the expectation of $r(\theta)$, a sum of independent Bernoulli random variables, $R(\widehat{\theta})$ is not the expectation of $r(\widehat{\theta})$, because of the dependence of $\widehat{\theta}$ on the sample, and neither is $r(\widehat{\theta})$ a sum of independent random variables. To circumvent this unfortunate situation, some uniform control over the deviations of r from R is needed.

We will follow the PAC-Bayesian approach to this problem, originated in the machine learning community and pioneered by McAllester (1998, 1999). It can be seen as some variant of the more classical approach of M-estimators relying on empirical process theory — as described for instance in Van de Geer (2000).

It is built on some general principles:

- One idea is to embed the set of estimators of the type $\widehat{\theta}:\Omega\to\Theta$ into the larger set of regular conditional probability measures $\rho:\big(\Omega,(\mathcal{B}\otimes\mathcal{B}')^{\otimes N}\big)\to\mathcal{M}_+^1(\Theta,\mathcal{T})$. We will call these conditional probability measures *posterior distributions*, to follow standard terminology.
- A second idea is to measure the fluctuations of ρ with respect to the sample, using some prior distribution $\pi\in\mathcal{M}_+^1(\Theta,\mathcal{T})$, and the Kullback divergence function $\mathcal{K}(\rho,\pi)$. The expectation $\mathbb{P}\big\{\mathcal{K}(\rho,\pi)\big\}$ measures the randomness of ρ. The optimal choice of π would be $\mathbb{P}(\rho)$, resulting in a measure of the randomness of ρ equal to the mutual information between the sample and the estimated parameter drawn from ρ. Anyhow, since $\mathbb{P}(\rho)$ is usually not better known than $\mathbb{P}$, we will have to be content with some less concentrated prior distribution π, resulting in some looser measure of randomness, as shown by the identity $\mathbb{P}\big[\mathcal{K}(\rho,\pi)\big]=\mathbb{P}\big\{\mathcal{K}\big[\rho,\mathbb{P}(\rho)\big]\big\}+\mathcal{K}\big[\mathbb{P}(\rho),\pi\big]$.
- A third idea is to analyse the fluctuations of the random process $\theta\mapsto r(\theta)$ from its mean process $\theta\mapsto R(\theta)$ through the log-Laplace transform

$$-\frac{1}{\lambda}\log\left\{\iint\exp\big[-\lambda r(\theta,\omega)\big]\pi(d\theta)\mathbb{P}(d\omega)\right\},$$

  as would be done in statistical mechanics, where this is called the free energy. This transform is well suited to relate $\min_{\theta\in\Theta}r(\theta)$ to $\inf_{\theta\in\Theta}R(\theta)$, since for large enough values of the parameter λ, corresponding to low enough values of the temperature, the system has small fluctuations around its ground state.
- A fourth idea deals with localization. It consists of considering a prior distribution $\overline{\pi}$ depending on the unknown expected error rate function R. Thus some central result of the theory will consist in an empirical upper bound for $\mathcal{K}\big[\rho,\pi_{\exp(-\beta R)}\big]$, where $\pi_{\exp(-\beta R)}$, defined by its density

$$\frac{d}{d\pi}\big[\pi_{\exp(-\beta R)}\big]=\frac{\exp(-\beta R)}{\pi\big[\exp(-\beta R)\big]},$$

is a Gibbs distribution built from a known prior distribution $\pi \in \mathcal{M}_+^1(\Theta, \mathcal{T})$, some inverse temperature parameter $\beta \in \mathbb{R}_+$ and the expected error rate R. This bound will in particular be used when ρ is a posterior Gibbs distribution, of the form $\pi_{\exp(-\beta r)}$. The general idea will be to show that in the case when ρ is not too random, in the sense that it is possible to find a prior (that is non-random) distribution $\overline{\pi}$ such that $\mathcal{K}(\rho, \overline{\pi})$ is small, then $\rho(r)$ can be reliably taken for a good approximation of $\rho(R)$.

This monograph is divided into four chapters. The first deals with the inductive setting presented in these lines. The second is devoted to relative bounds. It shows that it is possible to obtain a tighter estimate of the mutual information between the sample and the estimated parameter by comparing prior and posterior Gibbs distributions. It shows how to use this idea to obtain adaptive model selection schemes under very weak hypotheses.

The third chapter introduces the *transductive* setting of V. Vapnik (Vapnik, 1998), which consists in comparing the performance of classification rules on the learning sample with their performance on a test sample instead of their average performance. The fourth one is a fast introduction to Support Vector Machines. It is the occasion to show the implications of the general results discussed in the three first chapters when some particular choice is made about the structure of the classification rules.

In the first chapter, two types of bounds are shown. *Empirical bounds* are useful to build, compare and select estimators. *Non random bounds* are useful to assess the speed of convergence of estimators, relating this speed to the behaviour of the Gibbs prior expected error rate $\beta \mapsto \pi_{\exp(-\beta R)}(R)$ and to covariance factors related to the margin assumption of Mammen and Tsybakov when a finer analysis is performed. We will proceed from the most straightforward bounds towards more elaborate ones, built to achieve a better asymptotic behaviour. In this course towards more sophisticated inequalities, we will introduce *local bounds* and *relative bounds*.

The study of relative bounds is expanded in the third chapter, where tighter comparisons between prior and posterior Gibbs distributions are proved. Theorems 2.1.3 (page 54) and 2.2.4 (page 72) present two ways of selecting some nearly optimal classification rule. They are both proved to be adaptive in all the parameters under Mammen and Tsybakov margin assumptions and parametric complexity assumptions. This is done in Corollary 2.1.17 (page 66) of Theorem 2.1.15 (page 65) and in Theorem 2.2.11 (page 88). In the first approach, the performance of a randomized estimator modelled by a posterior distribution is compared with the performance of a prior Gibbs distribution. In the second approach posterior distributions are directly compared between themselves (and leads to slightly stronger results, to the price of using a more complex algorithm). When there are more than one parametric model, it is appropriate to use also some *doubly localized scheme*: two step localization is presented for both approaches, in Theorems 2.3.2 (page 93) and 2.3.9 (page 107) and provides bounds with a decreased influence of the number of empirically inefficient models included in the selection scheme.

We would not like to induce the reader into thinking that the most sophisticated results presented in these first two chapters are necessarily the most useful ones, they are as a rule only more efficient *asymptotically*, whereas, being more involved, they use looser constants leading to less precision for small sample sizes. In practice whether a sample is to be considered small is a question of the ratio between the number of examples and the complexity (roughly speaking the number of parameters) of the model used for classification. Since our aim here is to describe methods

appropriate for complex data (images, speech, DNA, ...), we suspect that practitioners wanting to make use of our proposals will often be confronted with small sample sizes; thus we would advise them to try the simplest bounds first and only afterwards see whether the asymptotically better ones can bring some improvement.

We would also like to point out that the results of the first two chapters are not of a purely theoretical nature: posterior parameter distributions can indeed be computed effectively, using Monte Carlo techniques, and there is well-established know-how about these computations in Bayesian statistics. Moreover, non-randomized estimators of the classical form $\widehat{\theta} : \Omega \to \Theta$ can be efficiently approximated by posterior distributions $\rho : \Omega \to \mathcal{M}_+^1(\Theta)$ supported by a fairly narrow neighbourhood of $\widehat{\theta}$, more precisely a neighbourhood of the size of the typical fluctuations of $\widehat{\theta}$, so that this randomized approximation of $\widehat{\theta}$ will most of the time provide the same classification as $\widehat{\theta}$ itself, except for a small amount of dubious examples for which the classification provided by $\widehat{\theta}$ would anyway be unreliable. This is explained on page 7.

As already mentioned, the third chapter is about the *transductive setting*, that is about comparing the performance of estimators on a training set and on a test set. We show first that this comparison can be based on a set of exponential deviation inequalities which parallels the one used in the inductive case. This gives the opportunity to *transport* all the results obtained in the inductive case in a systematic way. In the transductive setting, the use of prior distributions can be extended to the use of *partially exchangeable posterior distributions* depending on the union of training and test patterns, bringing increased possibilities to adapt to the data and giving rise to such crucial notions of complexity as the Vapnik–Cervonenkis dimension.

Having done so, we more specifically focus on the *small sample case*, where local and relative bounds are not expected to be of great help. Introducing a fictitious (that is unobserved) shadow sample, we study Vapnik-type generalization bounds, showing how to tighten and extend them with some original ideas, like making no Gaussian approximation to the log-Laplace transform of Bernoulli random variables, using a shadow sample of arbitrary size. shrinking from the use of any symmetrization trick, and using a suitable subset of the group of permutations to cover the case of independent non-identically distributed data. The culminating result of the third chapter is Theorem 3.3.3 (page 125), subsequent bounds showing the separate influence of the above ideas and providing an easier comparison with Vapnik's original results. Vapnik-type generalization bounds have a broad applicability, not only through the concept of Vapnik–Cervonenkis dimension, but also through the use of compression schemes (Little et al., 1986), which are briefly described on page 117.

The beginning of the fourth chapter introduces Support Vector Machines, both in the separable and in the non-separable case (using the box constraint). We then describe different types of bounds. We start with compression scheme bounds, to proceed with margin bounds. We begin with transductive margin bounds, recalling on this occasion in Theorem 4.2.2 (page 144) the growth bound for a family of classification rules with given Vapnik–Cervonenkis dimension. In Theorem 4.2.4 (page 145) we give the usual estimate of the Vapnik–Cervonenkis dimension of a family of separating hyperplanes with a given transductive margin (we mean by this that the margin is computed on the union of the training and test sets). We present an original probabilistic proof inspired by a similar one from Cristianini et al. (2000), whereas other proofs available usually rely on the informal claim that

the simplex is the worst case. We end this short review of Support Vector Machines with a discussion of inductive margin bounds. Here the margin is computed on the training set only, and a more involved combinatorial lemma, due to Alon et al. (1997) and recalled in Lemma 4.2.6 (page 149) is used. We use this lemma and the results of the third chapter to establish a bound depending on the margin of the training set alone.

In appendix, we finally discuss the textbook example of classification by thresholding: in this setting, each classification rule is built by thresholding a series of measurements and taking a decision based on these thresholded values. This relatively simple example (which can be considered as an introduction to the more technical case of classification trees) can be used to give more flesh to the results of the first three chapters.

It is a pleasure to end this introduction with my greatest thanks to Anthony Davison, for his careful reading of the manuscript and his numerous suggestions.