On the Flexibility of Theoretical Models for Pattern Recognition

Daniil Riabko

Thesis submitted to the University of London
for the degree of Doctor of Philosophy

Royal Holloway
University of London

Department of Computer Science,
Royal Holloway, University of London,
Egham Hill, Egham, TW20 0EX,
United Kingdom
e-mail: daniil@cs.rhul.ac.uk

April, 2005

Abstract

This thesis is devoted to relaxing certain theoretical assumptions in pattern recognition models. In pattern recognition a predictor is trying to guess a discrete label of some object (usually a real vector), based on given examples of object-label pairs.

Pattern recognition was developed to serve a growing number of practical applications; many of these applications still do not fit any theoretical model. The aim of this work is to make existing learning algorithms justifiably applicable to a wider range of practical tasks, by proving that certain theoretical assumptions can be relaxed without significant loss in performance.

The first assumption this work aims to relax is that examples are independent and identically distributed (i.i.d.); this condition is traditionally imposed in the majority of works on pattern recognition, but is often violated in applications. It turns out that many results of pattern recognition theory carry over a weaker assumption. Namely, under the assumption that objects are conditionally independent and identically distributed given their labels, while the rate of occurrence of each label should be above some positive threshold. A broad class of predictors is found which retain their performance under the conditional i.i.d. assumption.

The second assumption concerns the online scenario, according to which the recognition process consists of a sequence of trials: an object, a prediction, the correct label, another object, etc. While this scenario is convenient for theoretical studies, in reality the correct answers are usually not available after each prediction. A more general online scenario is suggested, which allows correct answers to be given with delays and not at every trial. The new scenario is studied on symmetric predictors, mainly concentrating on Transductive Confidence Machines. Some sufficient conditions are found under which a predictor's error rates remain the same in the new online scenario as in the traditional one.

Contents

List of statements

the size of a cell tends to zero and the number of examples in a cell tends to infinity in probability.

Theorem 3.2 62

Provides some sufficient conditions on the parameters of the deterministic Weak Teachers scenario under which a region predictor retains (from the pure on-line scenario) its rate of uncertain predictions or a TCM remains well-calibrated.

Theorem 3.3 65

Provides some sufficient conditions on the parameters of the randomised Weak Teachers scenario under which a region predictor retains its rate of uncertain predictions (from the pure on-line scenario) or a TCM remains well-calibrated.

Acknowledgements

First I would like to thank my supervisors Alex Gammerman and Vladimir Vovk. I am grateful to Royal Holloway, University of London for funding my Ph.D. studies. Special thanks to David Lindsay for his numerous useful comments made while reading draft versions of the thesis.

Chapter 1

Introduction

Motivation

The task of pattern recognition concerns the prediction of an unknown label of some observation (or object). For instance, the object can be an image of a hand-written letter, in which case the label is the actual letter represented by this image. Other examples include DNA sequence identification, diagnosis of an illness based on a set of symptoms, speech recognition, and many others. In mathematical formulation a label is a member of a finite set while an object is a finite-dimensional real vector.

The development of pattern recognition in its modern form probably began with the works of Rosenblatt on perceptron in the late 1950s. Since that time numerous approaches to pattern recognition and related learning problems have been developed. Of these we will mention Statistical Learning Theory of Vapnik and Chervonenkis, classical non-parametric rules such as nearest neighbours and partitioning rules, studied by Stone, Cover and many others, a recent approach proposed by Vovk and Gammerman called Conformal Prediction, and Neural Networks.

An abundance of practical applications is what attracts researchers to studying pattern recognition tasks, but it is also what makes perhaps any theoretical model deficient. Thus, many applications, even such classical as hand-written text recognition, do not comply with some theoretical assumptions traditionally imposed on data in the literature.

This thesis is devoted to proving that certain theoretical assumptions can be relaxed without significant loss in performance, thus making pattern recognition methods justifiably applicable to a wider range of practical tasks.

Before proceeding with a detailed explanation we need some more formal definitions.

Pattern recognition models

The formal model used most widely can be briefly introduced as follows. The objects $x \in \mathbf{X}$ are drawn independently and identically distributed (i.i.d.) according to some unknown (but fixed) probability distribution $P(x)$. The labels $y \in \mathbf{Y}$ are given for each object according to some, also unknown but fixed, function $\eta(x)$. Often a more general situation is considered, in which the labels are drawn according to some probability distribution $P(y|x)$, i.e. each object can have more than one possible label. The space $\mathbf{Y}$ of labels is assumed to be finite (often binary). The task is to construct the best predictor for the labels, based on the data observed, i.e. actually to "learn" $\eta(x)$.

This task is usually considered in either of the following two settings. In off-line setting a (finite) set of examples is divided into two finite subsets, the training set and the testing set. A predictor is constructed based on the training set and then is used to classify the objects from the testing set.

In the on-line setting a predictor starts by classifying the first object with zero knowledge; then it is given the correct label and (having "learned" this information) proceeds with classifying the second object, the second correct label is given, and so on.

In either of the settings, a predictor is just a function (more formally, a family of functions indexed by n) which takes as its arguments a n-tuple of object-label examples (training examples) and on object, and returns the label for this object.

Learning conditionally i.i.d. data

The first theoretical assumption we aim to relax is one of the central in the model: the assumption that examples are independent and identically distributed. We show that many pattern recognition methods work nearly as well under weaker assumptions. Namely, under the assumption that objects are conditionally independent given labels.

First consider the following example. Suppose we are trying to recognise a hand-written text. Obviously, letters in the text are dependent (for example, we strongly expect to meet "u" after "q"). This seemingly implies that pattern recognition can not be applied to this task, which is, however, one of their classical applications.

It appears that the only required assumptions on the distribution of examples are as follows. First, that the dependence between objects is only that between their labels; in other words, the type of object-label dependence remains constant over time. In our example, an image of a letter which in the beginning of the text denotes, say, "a", later on in the text will not be interpreted as, say, "e". Second, for each label, the distribution of corresponding objects does not change. In the hand–written text example this means that e.g. hand–writing style is not changing. Finally, the third assumption is that the rate of occurrence of each label should keep above some positive threshold. In the above example, the rate of occurrence of each letter should be, say, between 1% and 99% of all letters, with some feasible probability (depending on the size of the text).

These intuitive ideas lead us to the following model (to which we refer as "the conditional model"). The labels $y \in \mathbf{Y}$ are drawn according to some unknown (but fixed) distribution over the set of all infinite sequences of labels. There can be any type of dependence between labels; moreover, we can assume that we are dealing with any (fixed) combinatorial sequence of labels. However, in this sequence the rate of occurrence of each label should keep above some positive threshold. For each label y the corresponding object $x \in \mathbf{X}$ is generated according to some (unknown but fixed) probability distribution $P(x|y)$. All the rest is as in the i.i.d. model.

The main difference from the i.i.d. model is in that in the conditional model we made the distribution of labels primal; having done that we can relax the requirement of independence of objects to the conditional independence, and replace the i.i.d. assumption about the distribution of labels with the only assumption that the rate of occurrence of each label does not tend to zero.

One of the main criteria in estimating how well a predictor works is the probability of its error. In this work we provide a tool for obtaining estimations of probability of error of a predictor in the conditional model from an estimation of the probability of error in the i.i.d. model. The only assumption on a predictor under which the new estimations are of the same order is what we call *tolerance to data*: in any large dataset there is no small subset which significantly affects the probability of error. This property should also hold with respect to permutations. This assumption on a predictor should be valid in the i.i.d. model. Thus, the results achieved in the i.i.d. model can be extended to the conditional model; this concerns distribution–free results as well as distribution–specific, results on the performance on finite samples as well as asymptotic results.

The general theorems about extending results concerning performance of a predictor to the conditional model are illustrated on two classes of predictors.

First, we use some results of Vapnik-Chervonenkis theory to estimate performance in the conditional model (on finite amount of data) of predictors minimising empirical risk. We obtain new bounds on the probability of error for such predictors in the conditional model, which are of the same order as the known bounds in the i.i.d. model. We also obtain some strong consistency results for predictors minimising empirical risk over classes with growing VC dimension.

Second, we extend weak consistency results concerning partitioning and nearest neighbour estimates from the i.i.d. model to the conditional model. That is, we show that for the nearest neighbour predictors and certain partitioning rules the probability of error in the conditional model tends to zero.

Various attempts to relax the i.i.d. assumption on learning tasks have been taken in the literature. Some of these approaches involve the ideas of conditional independence, wheres others are concerned with such probabilistic models as Markov chains, stationarity of the sequence of examples, and so on. See section "Related work" for an overview.

Online prediction with Weak Teachers

The second assumption we aim to relax concerns the on-line scenario for pattern recognition. As noted earlier, in the on-line learning scenario examples arrive one by one: object, prediction, the correct label, next object, and so on. Examples are assumed to be independent and identically distributed.

While the online scenario is convenient for theoretical studies, in practice, however, rarely does one immediately obtain the true label for every object (otherwise the prediction is not needed). In practice, the true label for an object is usually given with some delay, if it is given at all. In this work we modify the online scenario to cover all such cases.

We suggest the following modified scenario for online prediction, which we call *online prediction with weak teachers*. At each trial we are given an object, and at some trials we are also given the correct label for one of the previous objects, so that the predictor can use this data afterwards. Thus, some labels may never be revealed while others may be revealed with some delay. Examples are assumed to be i.i.d., as in the pure on-line scenario. The trials at and for which true labels are given are chosen independently of data. We consider two types of what we call learning rules: deterministic and randomised; in the first case the steps on which labels are revealed, as well as the numbers of actual labels are supposed to be pre-defined, while in the second case they are drawn according to some probability distribution.

The suggested scenarios are studied on so-called region predictors, which instead of a single label can output a set of labels as a prediction. We particularly concentrate on the class of region predictors called Transductive Confidence Machines (TCMs), or conformal predictors.

Transductive Confidence Machines, proposed and developed to a great extend by Vovk, Gammerman and others, is a way of constructing predictors from machine-learning algorithms. One of the advantages of a TCM is that it is always *well-calibrated*: the number of errors it makes up to trial n divided by n tends to δ almost surely, where $\delta \in (0, 1)$ is any pre-specified "significance level". Moreover, the probability of error at each trial is δ and errors are made independently at different trials.

We find sufficient conditions on the amount of information revealed to a predictor up to the nth trial under which it has the same rate of errors; or, for the case of a TCM, remains well-calibrated and has the same asymptotic rate of uncertain predictions.

The main results obtained in this direction can be illustrated by two simple examples. Deterministic case: suppose only every kth label is revealed to a predictor, and even this is done with a delay of l, where k and l are positive integer constants. Randomised case: suppose that each label is revealed with some probability $p < 1$, and this is done with some (random) delay bounded by an integer l. In both described cases the asymptotic rates of error and uncertain predictions will not suffer (in particular, if the predictor is a TCM then it will remain well-calibrated).

List of principal results

The contribution of this work can be summarised as follows.

Conditionally i.i.d. data:

- A general model for pattern recognition (called conditional model) which extends the traditionally used one is proposed.
- A set of general theoretical tools is developed for extending results obtained under the traditional (i.i.d.) model to the conditional model.
- As an application of the general results, a new bound on the probability of error for predictors minimising empirical risk in the conditional model is obtained.
- As another application, some classical weak consistency results for the nearest neighbour and partitioning estimates are generalised to the case of conditional model.

Weak Teachers:

- A new scenario (called weak teachers) for pattern recognition is proposed which generalises the on-line scenario to a more realistic case.
- Some sufficient conditions are found under which a predictor (a region predictor) achieves the same rate of errors (uncertain predictions) in the new scenario as in the pure on-line scenario.
- Some sufficient conditions are found under which a TCM remains well-calibrated in the new scenario.

List of author's publication related to the thesis

- D. Ryabko, Online Learning of Conditionally I.I.D. Data. *In: R. Greiner and D. Schuurmans, Proceedings of the 21st International Conference on Machine Learning, Banff, Canada,* pp. 727–734, 2004

- D. Ryabko, Application of Classical Nonparametric Predictors to Learning Conditionally I.I.D. Data. *In: S. Ben-David, J. Case, A. Maruoka, Proceedings of 15th International Conference on Algorithmic Learning Theory, Padova, Italy; Springer LNAI 3244,* pp. 171–180, 2004.

- D. Ryabko, Pattern Recognition for Conditionally Independent Data. Theoretical Computer Science, conditionally accepted (currently under the second review).

- The results of Chapter 3 are appearing as a part of:
  V. Vovk, A. Gammerman, G. Shafer. *Algorithmic Learning in a Random World*, Springer, 2004.

- D. Ryabko, V.Vovk, A.Gammerman Online region prediction with real teachers. Technical Report CSD-TR-03-09, Royal Holloway University of London Computer Science Dept., 2003

Related work

Many monographs are devoted to pattern recognition and related learning problems; of these we mention here [45] for overview of statistical learning theory (Vapnik-Chervonenkis theory), [12, 17] for extensive overview of nonparametric methods (the latter presenting a more practical approach), [54] for the theory of conformal prediction (TCMs), [38] for an overview on neural networks, [50, 24] for PAC theory.

The reader is referred to the above referenced sources for extensive overviews, while this section is devoted to the observation of work related to the main topics of the thesis: attempts to relaxing the i.i.d. assumption in pattern recognition, and on-line learning.

Relaxing the i.i.d. assumption

Individual sequence prediction

One approach in which conditional independence is used was developed in [25], [26]. The authors study nearest neighbours and kernel estimators for the task of regression estimation with continuous regression function (in [26] some Lipschitz conditions are also assumed). The task of regression estimation is similar to pattern recognition, except for that labels can range over the set of real numbers, rather than over a finite set. The probabilistic assumption considered by the authors is that labels are conditionally independent given their objects (each label y_i is drawn according to an unknown conditional distribution $P(y|x_i)$), while objects form any individual sequence.

Observe that this probabilistic assumption is strictly weaker than ours; what allows the authors to consider arbitrarily distributed objects is the assumption that the regression function $E(y|x)$ is continuous. (In the pattern recognition task which we consider the regression function is discontinuous in all non-trivial cases.) Note also the generality of the referred approach: both nearest neighbours and kernels estimators are considered; it is also natural to expect that similar results can be obtained for appropriate partitioning rules.

A similar approach is considered in [32], where a regression estimation scheme is proposed which is consistent for any individual stable sequence of object-label pairs — placing no probabilistic assumptions at all. (This approach is also applied to the density estimation task in [35].) Here the conditional independence of labels is replaced with a certain type of stability. More importantly, the authors also assume that there is a known upper bound on the variation of regression function (which plays the role of Lipschitz conditions of the previous approach). The regression estimation scheme considered is a partitioning rule.

Again, the probabilistic assumption is obviously weaker that ours but the assumption on variation of regression function makes the learning task completely different.

As the previous example illustrates, non-trivial consistency results can often be obtained without any probabilistic assumptions at all, at the cost of different restrictions on the nature of pattern recognition problem. As we are more interested in strict generalisations, while in such cases various non-trivial restrictions are also adopted, we mention only one more example.

This classical example concerns perceptron. Suppose that objects lie in a bounded subset of $\mathbb{R}^d$ and are separable by a hyperplane with a margin. The Novikoff Theorem [37] says that in this case, for any individual sequence of examples the perceptron finds a separating hyperplane, i.e. achieves zero probability of error. Moreover, a bound on the number of iterations made by the perceptron algorithm is provided.

Markov processes

Another attempt of relaxing the i.i.d. assumption had been taken in [20] and [1], where a generalisation of PAC approach to Markov chains is considered.

In the PAC approach (following Vapnik-Chervonenkis theory) a class of decision rules (functions from $\mathbf{X}$ to $\mathbf{Y} = \{0, 1\}$) $\mathcal{C}$ is considered, and the goal is to find the best rule in $\mathcal{C}$ for solving the current pattern recognition task. Often, and in particular in the referred works, it is also assumed that the optimal rule belongs to $\mathcal{C}$. The probabilistic assumption made is that the random process takes the form of a Markov chain with finite or countable state space.

Under these assumptions the authors obtain finite-step estimates of the probability of error.

Stationary processes and sequence prediction

As one of the most general probabilistic assumptions which still models many real-life situations, stationary processes are in the focus of attention of many researchers. However, usually learning tasks other than pattern recognition are considered under this assumption. It is mainly the task of sequential prediction, that is, predicting the next element of a series of "labels" (without objects) given a finite or infinite past, that is being investigated. The task of predicting a label given its object and the past, i.e. learning object-label dependence, is considered as an extension of the task of learning label-label dependence: the objects are thought of as "side information". Still, many ideas are shared between this task and the task of pattern recognition.

Thus, in [3, 31] the authors study the task of predicting a stationary sequence, and also consider the generalisation of this task to the task of regression estimation and pattern recognition. Under the assumption that the joint distribution of objects and labels is stationary and ergodic, the authors construct a weakly consistent predictor, based on partitioning estimates.

For these learning problems no strong consistency or finite-step results can possibly be obtained. As it was shown in [41], this limitation concerns even the discrete case of the task of predicting stationary task series. On limitations to classification from stationary processes see also [34].

Another track of research which concerns predicting a sequence of labels with side information is based on applying Hidden Markov Models (see e.g. [39]) and some their generalisations. Usually additional probabilistic constraints (of different degree) are imposed on the distribution of examples, often resulting in methods applicable to specific practical tasks. Here we mention Maximal Entropy Markov Models [30], Conditional Random Fields [27] and Hidden Markov Support Vector Machines [2]. Such approaches naturally require new (at least adapted) recognition methods, and in this sense they are different from our approach which aims to extend applicability of existing algorithms.

Transductive Confidence Machines

Transductive Confidence Machines (TCMs, or conformal transducers), first introduced in [53] in 1999, now represent a very general branch of machine learning. Here we provide a brief survey of the features of TCMs important for our work; for a complete overview of this theory see [54].

First of all, a TCM is a region predictor, which means that it can output a prediction not as a single label, but as a set of labels. A prediction is correct if it contains the correct label.

One of the advantages of a TCM is that it is well-calibrated: the number of (region) errors it makes up to trial n divided by n tends to δ almost surely, where $\delta \in (0, 1)$ is any pre-specified "significance level" (see [51]). Moreover, the probability of error at each trial is δ and errors are made independently at different trials.

The way in which a TCM actually makes its predictions is not fixed; that is, a TCM is not some fixed predictor, but a way of constructing region predictors from machine-learning algorithms. A learning algorithm is implemented into a TCM via a so-called individual strangeness measure. If underlying learning algorithm is bad (for the problem at hand) then a TCM will output many "uncertain" predictions, i.e. predictions consisting of more then one label. However, it will still remain well-calibrated: the total rate of errors is kept at a pre-defined level.

Thus, a TCM allows to make a trade-off between erroneous and uncertainty; and to measure the performance of such algorithms, two rates must be evaluated: the rate of errors and the rate of uncertain predictions. As it was already mentioned, it is customary to measure the performance of prediction of a TCM is by rates (of errors and uncertain predictions) rather than by probabilities. This theoretical measure of performance, introduced for learning algorithms in [41], in pattern recognition literature is used perhaps mainly in connection with TCMs. However, it is this measure of performance that allows us to consider delayed or omitted labels (i.e. imperfect or "weak" teachers) in the on-line scenario. Indeed, the probability of error remains constant over delays, hence nothing can be said about the influence of delays on the performance thus measured. On the other hand, rates of erroneous provide an appropriate tool for estimating the role of delayed or omitted labels.

The referred approaches being relatively new, there is not yet much work done in studying generalisations of the on-line scenario. However, our "Weak Teachers" model has already gained some attention. Thus, in [36] the authors establish some necessary and sufficient conditions for a TCM in the Weak Teachers scenario to remain well-calibrated in probability (rather then almost surely as in this study).

Synopsis

Chapter 2 is devoted to relaxing the i.i.d. assumption in pattern recognition. In this chapter, the first section (**Section 2.1**) introduces the basic definitions and two pattern recognition models: the traditional model, which we call "i.i.d. model", and the new one, which is called "conditional model".

In the i.i.d. model the objects x_i, $i \in \mathbb{N}$ are generated independently and identically distributed according to some (unknown but fixed) distribution on the object space $\mathbf{X} = \mathbb{R}^d$, while labels $y_i \in \mathbf{Y} = \{0, 1\}$ are defined by some (also unknown but fixed) function $\eta : \mathbf{X} \to \mathbf{Y}$. A predictor is required to estimate η based on some examples $x_1, y_, \dots, x_n, y_n$.

In the conditional model the assumption that examples are generated i.i.d. is replaced by the following two conditions (the symbol $\mathbf{P}$ is used for the distribution generating examples). First, for any $n \in \mathbb{N}$ and for any measurable set A in $\mathbf{X}$

$$\mathbf{P}(X_n \in A \mid Y_n, X_1, Y_1, \dots, X_{n-1}, Y_{n-1}) = \mathbf{P}(X_n \in A \mid Y_n)$$

(upper case letters X and Y are used for random variables). Second, for any $y \in \mathbf{Y}$, for any $n_1, n_2 \in \mathbb{N}$ and for any measurable set A in $\mathbf{X}$

$$\mathbf{P}(X_{n_1} \in A \mid Y_{n_1} = y) = \mathbf{P}(X_{n_2} \in A \mid Y_{n_2} = y)$$

These conditions mean that objects are independent given their labels, and that the object-label dependence does not change in time. To obtain new estimates of the probability of error some assumptions on the frequencies of labels are also needed. Namely, it is required that the frequency of occurrence of each label stays (with high probability) in some interval $[\delta, 1 - \delta]$, $\delta > 0$. For the finite-step results this assumption is captured in the constants C_n (see below), while for the asymptotic results the following condition is introduced:

$$\lim_{n\to\infty} \mathbf{P}(p(n) \in [\delta, 1-\delta]) = 1,$$

for some δ, $0 < \delta < 1/2$, where $p(n) := \frac{1}{n}\#\{i \le n : Y_i = 0\}$.

The general results which allow to convert estimates on the probability of error obtained in the i.i.d. model to the conditional model are also presented in **Section 2.1**.

The main result of this section is a theorem, which provides estimates of the probability of error of a predictor in the conditional model via the estimates of its probability of error in the i.i.d. model and its tolerance to data. Tolerance to data measures the response of a predictor to small changes in a training sample. The new estimates also depend on constants C_n which are related to the frequencies of occurrence of the labels, and are defined as follows. Fix some $\delta \in (0, 1/2]$. Define $C_n := \mathbf{P}(\delta \le |p(n)| \le 1-\delta)$ for each $n \in \mathbb{N}$.

The referred theorem is followed by a corollary which provides a general tool for obtaining weak consistency results in the conditional model.

In **Section 2.2** some bounds on the performance of predictors minimising empirical risk are obtained, using the results of the previous section. The new bounds are of the same order as the corresponding bounds known in the statistical learning theory for the classical (i.i.d.) model.

The new results of the section are preceeded by a brief introduction to the theory of empirical risk minimisation.

Let $\mathrm{err}_n(\Gamma, \mathbf{P})$ denote the probability of error of a predictor Γ on the step $n+1$ (it also depends on the $x_1, \dots, x_n$).

The main result of Section 2.2 is the following theorem.

Theorem. *Let $\mathcal{C}$ be a class of decision functions and let Γ be a predictor which for each $n \in \mathbb{N}$ minimises empirical error over $\mathcal{C}$ on the observed examples $x_1, y_1, \dots, x_n, y_n$. Fix some $\delta \in (0, 1/2]$. Assume $n > 4/\varepsilon^2$. We* *have*

$$\mathbf{P}(\mathrm{err}_n(\Gamma, \mathbf{P}) > \varepsilon) \leq I_{2\,\mathrm{err}(\varphi_{P_{1/2}}, P_{1/2}) > \varepsilon/2} + 32\mathcal{S}(\mathcal{C}, n)e^{-2^{-15}n\delta^2\varepsilon^2} + (1 - C_n),$$

where $\mathcal{S}(\mathcal{C}, n)$ *is the* n*th shatter coefficient of the class* $\mathcal{C}$*. If in addition* $\eta \in \mathcal{C}$ *then*

$$\mathbf{P}(\mathrm{err}_n(\Gamma, \mathbf{P}) > \varepsilon) \leq 8\mathcal{S}(\mathcal{C}, n)e^{-n\delta\varepsilon/48} + (1 - C_n).$$

Two corollaries of this theorem are derived. These corollaries provide a strong consistency result for predictors minimising empirical risk over a series of classes with growing VC dimension, and an application of this result to the classes of neural networks.

In **Section 2.3** the results of Section 2.1 are used to obtain weak consistency results for the nearest neighbours and partitioning rules. First these rules are defined, and the classical results concerning their consistency in the i.i.d. model are presented. The consistency results are then generalised to the conditional model as follows.

Theorem. *Let* Γ *be the nearest neighbour classifier. Let* $\mathbf{P}$ *be some distribution on* $\mathbf{X}^\infty$ *satisfying the assumptions of the conditional model such that rates of occurrence of labels are bounded from below. Then*

$$E(\mathrm{err}_n(\Gamma, \mathbf{P})) \to 0.$$

Theorem. *Let* Γ *be a partitioning predictor such that the diameter of a cell tends to zero and the number of examples in a cell tends to infinity in probability for any distribution generating i.i.d. examples. Then*

$$E(\mathrm{err}_n(\Gamma, \mathbf{P})) \to 0.$$

for any distribution $\mathbf{P}$ *satisfying the assumptions of the conditional model such that rates of occurrence of labels are bounded from below.*

Section 2.4 provides a discussion of the assumptions made in the new model and of the conditions of the main theorems. It is shown (and illustrated by counterexamples) that no assumption can be withdrawn.

Chapter 3 is devoted to a generalisation of the traditional on-line learning scenario for pattern recognition. In this chapter we work with a generalised version of predictors, so-called region predictors, which are allowed to output predictions consisting of several labels. The performance of such predictors is measured not only by rates of errors, but also by rates of uncertain predictions, i.e. predictions consisting of more than one label.

In **Section 3.1** these and related notions are defined, and the class of predictors called Transductive Confidence Machines (TCMs) is introduced. TCMs have many useful properties, one of which is that a TCM is always well calibrated, i.e. achieves in asymptotic any pre-specified level of erroneousness.

In the (strict) on-line scenario for pattern recognition a predictor starts by classifying the first object with zero knowledge, then it is given the correct label, and with this information it classifies the second object; it is given the second label, etc.

In **Section 3.2** a new scenario for pattern recognition is proposed, which allows situations in which some correct labels are delayed or not given at all. The new scenario is introduced in two versions: deterministic and randomised.

The deterministic scenario is defined as follows. A sequence $\mathfrak{L} = \big((l_i, k_i) : i \in \mathbb{N}\big)$, where $(l_i, k_i) \in \mathbb{N} \times \mathbb{N}$, $i \in \mathbb{N}$ is called a *(deterministic) learning rule* if $k_i \leq n_i = \sum_{j=1}^{i} l_j$ for all $i \in N$, and $i \neq j$ implies $k_i \neq k_j$ for all $i, j \in \mathbb{N}$. The symbol N denotes the set $\{n_1, n_2, \dots\}$. In this definition, the numbers l_i specify the delays with which true labels are disclosed, so that n_i are the numbers of trials on which the labels are disclosed, while k_i are the numbers of actual labels.

Thus, at the end of each trial $n \in N$ a predictor "learns" the label y_n if $n \in N$ and "learns" nothing otherwise. The sequence $(l_i : i \in \mathbb{N})$ specifies the intervals in which labels are given, while the sequence $(k_i : i \in \mathbb{N})$ specifies the labels given on corresponding steps.

Several examples of learning rules are considered: slow teacher (each label is delayed), lazy teacher (a fraction of labels is omitted), and some others. The main result of this section concerning deterministic learning rules provides some sufficient conditions on the parameters of the deterministic scenario under which a region predictor retains (from the pure on-line scenario) its rate of uncertain predictions or a TCM remains well-calibrated.

Theorem. *Let $\Gamma_{1-\delta}$ be a symmetric (region) predictor, let $\mathfrak{L}$ be a deterministic learning rule, and let $\Gamma^{\mathfrak{L}}_{1-\delta}$ be the $\mathfrak{L}$-taught version of $\Gamma_{1-\delta}$. The following statements hold for any probability distribution P^∞ generating the examples.*

- *If $\Gamma_{1-\delta}$ is a TCM and*
$$\sum_{i=2}^{\infty} \frac{l_i^2}{n_i^2} < \infty$$
then $\Gamma^{\mathfrak{L}}_{1-\delta}$ is well calibrated.
- *If for some $l \in \mathbb{N}$, $l_i = l$ from some i on, then $\overline{\mathrm{U}}(\Gamma^{\mathfrak{L}}_{1-\delta}, P) = \overline{\mathrm{U}}(\Gamma_{1-\delta}, P)$ and $\overline{\mathrm{Er}}(\Gamma^{\mathfrak{L}}_{1-\delta}, P) = \overline{\mathrm{Er}}(\Gamma_{1-\delta}, P)$.*

In this theorem, the intervals $\overline{\mathrm{U}}$ and $\overline{\mathrm{Er}}$ are the intervals which in asymptotic contain the rate of uncertain prediction and the rate of errors correspondingly; $1-\delta$ is a pre-specified level of confidence.

In Subsection 3.2.2, a randomised version of learning rules is introduced. It allows to consider such examples as Bernoulli teacher (each label is given with a certain fixed probability, independently of each other), Poisson teacher (the delays with which labels are given are distributed according to the Poisson distribution), and others. An analogue of the latter theorem for the case of randomised learning rules is provided.

Section 3.3 discusses the conditions of the theorems, illustrating the main results with some examples in which the conditions are satisfied. Also in this section the necessity of certain conditions is discussed.

Finally, **Chapter 4** presents a conclusion and possible directions for the future research.

Chapter 2

Learning conditionally i.i.d. data

2.1 Definitions and general results

In this section we describe the commonly used i.i.d. model and introduce our generalisation: the conditional model. General results are presented, which allow to obtain estimates of performance of a predictor in the conditional model based on estimates achieved in the i.i.d. model. This section also introduces basic notations for spaces, random variables, predictors, etc. used in the succeeding sections.

2.1.1 The i.i.d. model

Consider a sequence of *examples*

$$(x_1, y_1), (x_2, y_2), \dots;$$

each example $z_i := (x_i, y_i)$ consists of an *object* $x_i \in \mathbf{X}$ and a *label* $y_i := \eta(x_i) \in \mathbf{Y}$, where $\mathbf{X}$ is a measurable space called an *object space*, $\mathbf{Y} := \{0, 1\}$ is called a *label space* and $\eta : \mathbf{X} \to \mathbf{Y}$ is some deterministic function. For simplicity we made the assumption that the space $\mathbf{Y}$ is binary, but all results easily extend to the case of any finite space $\mathbf{Y}$. The notation $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ is used for the measurable space of examples. Objects are drawn according to some probability distribution $\mathbf{P}$ on $\mathbf{X}^\infty$ (and labels are defined by η).

The notation $\mathbf{P}$ is used for distributions on $\mathbf{X}^\infty$ while the symbol P is reserved for distributions on $\mathbf{X}$. In the latter case P^∞ denotes the i.i.d. distribution on $\mathbf{X}^\infty$ generated by P. Letters x, y, z will be used for elements of spaces $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ correspondingly, while letters X, Y, Z are reserved for random variables on these spaces.

The traditional assumption about the distribution $\mathbf{P}$ generating objects is that examples are independently and identically distributed (i.i.d.) according to some distribution P on $\mathbf{X}$ (i.e. $\mathbf{P} = P^\infty$).

2.1.2 The conditional model

We introduce a modified model for pattern recognition task which we call *conditional model*. In this model, we replace the assumption that examples are independent with the following two conditions.

First, for any $n \in \mathbb{N}$ and for any measurable set A in $\mathbf{X}$

$$\mathbf{P}(X_n \in A \mid Y_n, X_1, Y_1, \dots, X_{n-1}, Y_{n-1}) = \mathbf{P}(X_n \in A \mid Y_n) \tag{2.1}$$

(i.e. some versions of conditional probabilities coincide). This condition resembles the Markov condition, with the help of which it can be understood more easily. Markov condition requires that each object depends on the past only through its immediate predecessor. The condition (2.1) says that each object depends on the past only through its label.

Second, for any $y \in \mathbf{Y}$, for any $n_1, n_2 \in \mathbb{N}$ and for any measurable set A in $\mathbf{X}$

$$\mathbf{P}(X_{n_1} \in A \mid Y_{n_1} = y) = \mathbf{P}(X_{n_2} \in A \mid Y_{n_2} = y) \tag{2.2}$$

(i.e. the process is uniform in time; (2.1) allows dependence on n).

Note that the first condition means that objects are conditionally independent given labels (on conditional independence see [10, 11]). Under the conditions (2.1) and (2.2) we say that *objects are conditionally independent and identically distributed* (conditionally i.i.d).

For each $y \in \mathbf{Y}$ denote by P_y the conditional distributions $\mathbf{P}(X_n \mid Y_n = y)$ (it does not depend on n by (2.2)). Clearly, the distributions P_0 and P_1 define some distribution P on $\mathbf{X}$ up to a parameter $p = P(y = 1) \in [0, 1]$. That is, define $P_p(A) := pP_1(A) + (1 - p)P_0(A)$ for any measurable set $A \subset \mathbf{X}$ and for each $p \in [0, 1]$. Thus with each distribution $\mathbf{P}$ satisfying the assumptions (2.1) and (2.2) we will associate a family of distributions P_p, $p \in [0, 1]$.

The assumptions of the conditional model can be also interpreted as follows. Assume that we have some individual sequence $(y_n)_{n\in\mathbb{N}}$ of labels and two probability distributions P_0 and P_1 on $\mathbf{X}$, such that there exist sets X_0 and X_1 in $\mathbf{X}$ such that $P_1(X_1) = P_0(X_0) = 1$ and $P_0(X_1) = P_1(X_0) = 0$ (i.e. X_0 and X_1 define some function η). Each example $x_n \in \mathbf{X}$ is drawn according to the distribution P_{y_n}; examples are drawn independently of each other.

2.1.3 Predictors: erroneousness and stability

In this section we define the notion of predictor and its probability of error. We also define what we call *tolerance to data* of a predictor, a tool for estimating the sensitivity of a predictor to small changes in training data. A brief overview is provided drawing attention to similar notions in pattern recognition literature.

A *predictor* is a measurable function

$$\Gamma(x_1, y_1, \dots, x_n, y_n, x_{n+1})$$

taking values in $\mathbf{Y}$. Let $\Gamma_n := \Gamma(x_1, y_1, \dots, x_n, y_n, x_{n+1})$.

The probability of error of a predictor Γ on each step n is defined as

$$\mathrm{err}_n(\Gamma, \mathbf{P}, z_1, \dots, z_n) := \mathbf{P}\bigl\{(x, y) \in \mathbf{Z} \bigm| y \neq \Gamma_n(z_1, \dots, z_n, x)\bigr\}$$

We will sometimes omit some of the arguments of err_n when it can cause no confusion; in particular, we will often use a short notation

$$\mathbf{P}(\mathrm{err}_n(\Gamma, Z_1, \dots, Z_n) > \varepsilon)$$

and an even shorter one $\mathbf{P}(\mathrm{err}_n(\Gamma) > \varepsilon)$ in place of

$$\mathbf{P}\bigl\{z_1, \dots, z_n : \mathrm{err}_n(\Gamma, \mathbf{P}, z_1, \dots, z_n) > \varepsilon\bigr\}.$$

For a pair of distributions P_0 and P_1 and any $\delta \in (0, 1/2)$ set

$$\triangledown_\delta(P_0, P_1, n, \varepsilon) := \sup_{p \in [\delta, 1-\delta]} P_p^\infty(\mathrm{err}_n(\Gamma) > \varepsilon) \tag{2.3}$$

For a predictor Γ and a distribution P on $\mathbf{X}$ define

$$\Delta(P, n, z_1, \dots, z_n) := \max_{j \leq \varkappa_n;\ \pi : \{1,\dots,n\} \to \{1,\dots,n\}} \bigl| \mathrm{err}_n(\Gamma, z_1, \dots, z_n) - \\ \mathrm{err}_{n-j}(\Gamma, z_{\pi(1)}, \dots, z_{\pi(n-j)})\bigr|;$$

where $\varkappa_n := \sqrt{n \log n}$ (see the end of Section 2.4 for the discussion of the choice of the constants $\varkappa_n$). Again, the variables $z_1, \dots, z_n$ will be sometimes omitted when it can cause no confusion. Define

$$\Delta(P, n, \varepsilon) := P^n\bigl(\Delta(P, n, Z_1, \dots, Z_n) > \varepsilon\bigr) \tag{2.4}$$

for any $n \in \mathbb{N}$, any $\varepsilon > 0$. Furthermore, for a pair of distributions P_0 and P_1 and any $\delta \in (0, 1/2)$ define the *tolerance to data* of a predictor as

$$\Delta_\delta(P_0, P_1, n, \varepsilon) := \sup_{p \in [\delta, 1-\delta]} \Delta(P_p, n, \varepsilon).$$

Tolerance to data means, in effect, that in any typical large portion of data there is no small portion that changes significantly the probability of error. This property should also hold with respect to permutations.

We will also use another version of tolerance to data, in which instead of removing (at most) $\varkappa_n$ examples from the training data $(z_1,\dots,z_n)$ we replace them with an arbitrary sample $z'_j,\dots,z'_n$ consistent with η:

$$\bar\Delta(P,z_1,\dots,z_n) := \sup_{j<\varkappa_n;\pi_1,\pi_2:\{1,\dots,n\}\to\{1,\dots,n\};z'_{n-j},\dots,z'_n}$$

$$|\operatorname{err}_n(\Gamma,P^\infty,z_1,\dots,z_n)-\operatorname{err}_n(\Gamma,P^\infty,\zeta_1,\dots,\zeta_n)|,$$

where $\zeta_{\pi_1(i)} := z_{\pi_2(i)}$ if $i<n-j$ and $\zeta_{\pi_1(i)} := z'_i$ otherwise; the maximum is taken over all z'_i, $n-j<i\le n$ consistent with η. Define

$$\bar\Delta(P,n,\varepsilon) := P^n\big(\bar\Delta(P,n,Z_1,\dots,Z_n)>\varepsilon\big)$$

and

$$\bar\Delta_\delta(P_0,P_1,n,\varepsilon) := \sup_{p\in[\delta,1-\delta]}\bar\Delta(P_p,n,\varepsilon).$$

The same notational convention will be applied to Δ and $\bar\Delta$ as to err_n.

Various notions similar to tolerance to data have been studied in the literature. Perhaps first they appeared in connection with deleted and condensed estimates [40, 15, 16]. These notion were later called hypothesis stability in [23]. Hypothesis stability measures the difference in probability of error when one point from the training sample is removed. Several other versions of stability (such as pointwise hypothesis stability, error stability, uniform stability) have been later introduced in [6, 23]; in [6] also an extensive overview is provided.

Naturally, such notions arise when there is a need to study the behaviour of a predictor when some of the training examples are removed. These notions are much similar to what we call tolerance to data, only we are interested in the maximal deviation of probability of error (when some portion of training data is removed or changed) while usually it is the average or minimal deviations that are estimated.

Since the term "stability" is already overburdened with sense, a different term ("tolerance to data") is chosen in the present work.

2.1.4 General results

A predictor developed to work in the off-line setting should be, loosely speaking, tolerant to small changes in a training sample. The next theorem shows under which conditions this property of a predictor can be utilised.

Theorem 2.1. *Suppose that a distribution* $\mathbf{P}$ *generating examples is such that the objects are conditionally i.i.d, i.e.* $\mathbf{P}$ *satisfies (2.1) and (2.2). Fix some* $\delta \in (0, 1/2]$*, let* $p(n) := \frac{1}{n}\#\{i \le n : Y_i = 1\}$ *and* $C_n := \mathbf{P}(\delta \le p(n) \le 1-\delta)$ *for each* $n \in \mathbb{N}$*. Let also* $\alpha_n := \frac{1}{1-1/\sqrt{n}}$*. For any predictor* Γ *and any* $\varepsilon > 0$ *we have*

$$\mathbf{P}(\mathrm{err}_n(\Gamma) > \varepsilon) \le C_n^{-1}\alpha_n\big(\triangledown_\delta(P_0, P_1, n + \varkappa_n, \delta\varepsilon/2) \\ + \Delta_\delta(P_0, P_1, n + \varkappa_n, \delta\varepsilon/2)\big) + (1 - C_n), \quad (2.5)$$

and

$$\mathbf{P}(\mathrm{err}_n(\Gamma) > \varepsilon) \le C_n^{-1}\alpha_n\big(\triangledown_\delta(P_0, P_1, n, \delta\varepsilon/2) \\ + \bar{\Delta}_\delta(P_0, P_1, n, \delta\varepsilon/2)\big) + (1 - C_n). \quad (2.6)$$

The proofs for this section can be found in Section 2.5.

The theorem says that if we know with some confidence C_n that the rate of occurrence of each label is not less than some (small) δ, then having bounds on the error rate of a predictor in the i.i.d. model we can obtain bounds on its error rate in the conditional model. In other words, suppose that we believe that objects are conditionally i.i.d and know with some confidence that each label does not cease to occur; then if we want to obtain estimates of the performance of a predictor we need to estimate its performance for the case of i.i.d examples and to obtain bounds on its tolerance to data, also for the case of i.i.d examples.

Thus we have a tool for estimating the performance of a predictor on each *finite* step n. In Section 2.2 we will show how this results can be applied to predictors minimising empirical risk. However, if we are only interested in asymptotic results the formulations can be simplified.

For any sequence of examples $z_1, z_2, \ldots$ define $p(n) := \frac{1}{n}\#\{i \le n : y_i = 1\}$. Consider the following asymptotic condition on the frequencies of labels. We say that the *rates of occurrence of labels are bounded from below* if there exist such δ, $0 < \delta < 1/2$ that

$$\lim_{n\to\infty} \mathbf{P}(p(n) \in [\delta, 1-\delta]) = 1. \tag{2.7}$$

As the condition (2.7) means $C_n \to 1$ we can derive from Theorem 2.1 the following corollary.

Corollary 2.2. *Suppose that a distribution* $\mathbf{P}$ *satisfies (2.1), (2.2), and (2.7) for some* $\delta \in (0, 1/2]$. *Let* Γ *be such a predictor that*

$$\lim_{n\to\infty} \nabla_\delta(P_0, P_1, n, \varepsilon) = 0 \tag{2.8}$$

and either

$$\lim_{n\to\infty} \Delta_\delta(P_0, P_1, n, \varepsilon) = 0 \tag{2.9}$$

or

$$\lim_{n\to\infty} \bar{\Delta}_\delta(P_0, P_1, n, \varepsilon) = 0 \tag{2.10}$$

for any $\varepsilon > 0$. *Then*

$$E(\mathrm{err}_n(\Gamma, \mathbf{P}, Z_1, \ldots, Z_n)) \to 0.$$

In Section 2.3 we show how this statement can be applied to prove weak consistence of some classical nonparametric predictors in the conditional model.

2.2 Application to Empirical Risk Minimisation

In this section we show how to estimate the performance of a predictor minimising empirical risk (over or certain class of functions) using Theorem 2.1. To do this we estimate tolerance to data of such predictors, using some results of Vapnik-Chervonenkis theory. In the first subsection some basic definitions and results are assembled which are required for the next subsection, in which our generalisation is presented.

2.2.1 Empirical risk minimisation

Empirical risk minimisation theory was developed by Vapnik and Chervonenkis in 1970s [47, 48]. For more recent overviews see [45, 49] and also [12], Chapter 12.

The main idea is to search for the solution to pattern recognition problem in two stages. On the first stage an appropriate class $\mathcal{C} \subset \mathcal{P}(\mathbf{X})$ of decision functions is selected (for example, on the basis of some prior knowledge). Then, the available data (the training data) is used to select a function from this class. Empirical risk of a decision function is the number of examples from the training set which it fails to classify correctly. It is shown that, provided the class $\mathcal{C}$ is not too big, an optimal solution is to select a function which minimises empirical risk. Now we proceed with a more formal exposition.

Let $\mathbf{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$ and let $\mathcal{C}$ be a class of measurable functions of the form $\varphi : \mathbf{X} \to \mathbf{Y} = \{0, 1\}$, called *decision functions.* For a probability distribution P on $\mathbf{X}$ define $\operatorname{err}(P, \varphi) := P(\varphi(X_i) \neq Y_i)$. If the examples are generated i.i.d. according to some distribution P, the aim is to find a function φ from $\mathcal{C}$ for which $\mathrm{err}(P,\varphi)$ is minimal:

$$\varphi_P = \mathrm{argmin}_{\varphi\in\mathcal{C}}\,\mathrm{err}(P,\varphi).$$

In the theory of empirical risk minimisation this function is approximated by the function

$$\varphi_n^* := \mathrm{argmin}_{\varphi\in\mathcal{C}}\,\overline{\mathrm{err}}_n(\varphi)$$

where $\overline{\mathrm{err}}_n(\varphi) := \sum_{i=1}^n I_{\varphi(X_i)\neq Y_i}$ is the empirical error functional, based on a sample (X_i,Y_i), $i=1,\dots,n$. Thus, $\Gamma(z_1,\dots,z_n,x_{n+1}) := \varphi_n^*(x_{n+1})$ is a predictor minimising empirical risk over the class of functions $\mathcal{C}$.

It is important to measure the capacity of the class $\mathcal{C}$; roughly speaking, if $\mathcal{C}$ is small enough then empirical error minimisation works. The capacity is measured by Vapnik-Chervonenkis dimension (VC-dimension) which is defined as follows. Define the n-th shatter coefficient of the class $\mathcal{C}$ as

$$\mathcal{S}(\mathcal{C},n) := \max_{A=\{x_1,\dots,x_n\}\subset\mathbf{X}} \#\{C\cap A : C\in\mathcal{C}\}.$$

In words, the shatter coefficient is the maximal number of different subsets of n points that can be picked out by the sets from $\mathcal{C}$. Clearly, $\mathcal{S}(\mathcal{C},n)\le 2^n$, and if $\mathcal{S}(\mathcal{C},k) < 2^k$ for some k, then $\mathcal{S}(\mathcal{C},n) < 2^n$ for all $n>k$. The first such k is called VC-dimension:

$$V(\mathcal{C}) := \max\{n\in\mathbb{N} : \mathcal{S}(\mathcal{C},n) = 2^n\}.$$

If $V(\mathcal{C})<\infty$ then the following bound is valid [46]

$$\mathcal{S}(\mathcal{C},n) \le \sum_{i=0}^{V(\mathcal{C})} \left(\begin{array}{c} n \\ i \end{array}\right) \le (n+1)^{V(\mathcal{C})}. \qquad (2.11)$$

Trivially, the VC dimension of the class of all subsets of $\mathbf{X}$ is infinite, as is the dimension of the class of finite unions of all intervals in $\mathbb{R}$. On the other hand, VC dimension of the class of intervals in $\mathbb{R}$ is 2. The following example (which will be used later) is provided in [4]. Let $\mathcal{C}^k$ be the class of all neural networks with k nodes in the hidden layer and the threshold sigmoid. Let $\mathbf{X} = \mathbb{R}^d$. Then

$$V(\mathcal{C}^k) \leq (2kd + 4k + 2) \log_2(e(kd + 2k + 1)). \tag{2.12}$$

One of the basic results of Vapnik-Chervonenkis theory is the estimation of the difference of probabilities of error between the best possible function in the class (φ_P) and the function which minimises empirical error (see [45] or [12], Theorem 12.6):

$$P\big(\operatorname{err}_n(\Gamma, P) - \operatorname{err}(\varphi_P, P) > \varepsilon\big) \leq 8\mathcal{S}(\mathcal{C}, n)e^{-n\varepsilon^2/128}, \tag{2.13}$$

Thus,

$$P(\operatorname{err}_n(\Gamma, P) > \varepsilon) \leq I_{\operatorname{err}(\varphi_P,P)>\varepsilon/2} + 8\mathcal{S}(\mathcal{C}, n)e^{-n\varepsilon^2/512}. \tag{2.14}$$

A particularly interesting case is when the optimal rule belongs to $\mathcal{C}$, i.e. when $\eta \in \mathcal{C}$. This situation was investigated in [44, 5, 33] and other works. Obviously, in this case $\varphi_P \in \mathcal{C}$ and $\operatorname{err}(\varphi_P, P) = 0$ for any P. Moreover, the bound (2.13) can be improved (see e.g. [12], Theorem 12.7)

$$P(\operatorname{err}_n(\Gamma, P) > \varepsilon) \leq 2\mathcal{S}(\mathcal{C}, n)e^{-n\varepsilon/2}. \tag{2.15}$$

These inequalities provide estimates of probability of error for predictors minimising empirical risk, given the estimate of error of the best decision function available. Clearly, these bounds are nontrivial only in the case when the VC dimension of the class $\mathcal{C}$ is finite. Important cases of classes with finite VC dimension include classes of hyperplanes and classes of partitions with fixed number of cells.

However, finite VC dimension usually means that the class is rather small. Thus, it is natural to look for a balance between the size of available dataset and the capacity of $\mathcal{C}$. By extending the class $\mathcal{C}$ with the growth of n we can possibly obtain better generalisation. For asymptotic (strong universal consistency) results achieved on this way see [29, 45]. Some of these results are also generalised in the next section.

2.2.2 Extension to the conditional model

The following theorem provides analogues of the bounds (2.14) and (2.15) for the conditional model.

Theorem 2.3. *Let $\mathcal{C}$ be a class of decision functions and let Γ be a predictor which for each $n \in \mathbb{N}$ minimises $\overline{\mathrm{err}}_n$ over $\mathcal{C}$ on the observed examples $(z_1, \dots, z_n)$. Fix some $\delta \in (0, 1/2]$, let $p(n) := \frac{1}{n}\#\{i \le n : Y_i = 0\}$ and $C_n := \mathbf{P}(\delta \le p(n) \le 1 - \delta)$ for each $n \in \mathbb{N}$. Assume $n > 4/\varepsilon^2$ and let $\alpha_n := \frac{1}{1-1/\sqrt{n}}$. We have*

$$\Delta(P_0, P_1, n, \varepsilon) \le 16\mathcal{S}(\mathcal{C}, n)e^{-n\varepsilon^2/512}. \tag{2.16}$$

(which does not depend on the distributions P_0 and P_1) and

$$\begin{aligned}\mathbf{P}(\mathrm{err}_n(\Gamma, \mathbf{P}) > \varepsilon) \le {} & I_{2\,\mathrm{err}(\varphi_{P_{1/2}}, P_{1/2}) > \varepsilon/2} \\ & + 16\alpha_n C_n^{-1}\mathcal{S}(\mathcal{C}, n)e^{-n\delta^2\varepsilon^2/2048} + (1 - C_n).\end{aligned} \tag{2.17}$$

If in addition $\eta \in \mathcal{C}$ then

$$\Delta(n, \varepsilon) \le 4\mathcal{S}(\mathcal{C}, 2n)2^{-n\varepsilon/8} \tag{2.18}$$

and

$$\mathbf{P}(\mathrm{err}_n(\Gamma, \mathbf{P}) > \varepsilon) \le 4\alpha_n C_n^{-1}\mathcal{S}(\mathcal{C}, n)e^{-n\delta\varepsilon/16} + (1 - C_n). \tag{2.19}$$

Thus, if we have bounds on the VC dimension of some class of classifiers, we can obtain bounds on the performance of the empirical error minimising predictors for the conditional model.

The proof can be found in Section 2.6.

Next we show how asymptotic (strong consistency) results can be achieved in the conditional model.

Using Theorem 2.3, inequality (2.11) and Borel-Cantelli lemma, we obtain the following statement.

Corollary 2.4. *Let $\mathcal{C}^k$, $k \in \mathbb{N}$ be a sequence of classes of decision functions with finite VC dimension such that $\lim_{k\to 0} \inf_{\varphi\in\mathcal{C}^k} \operatorname{err}(\varphi, P) = 0$ for any distribution P on $\mathbf{X}$. If $k_n \to \infty$ and $\frac{V(\mathcal{C}^{k_n})\log n}{n} \to 0$ as $n \to \infty$ then*

$$\operatorname{err}(\Gamma, \mathbf{P}) \to 0 \ \mathbf{P}\text{–}a.s.$$

where Γ is a predictor which in each trial n minimises empirical risk over $\mathcal{C}^{k_n}$ and $\mathbf{P}$ is any distribution satisfying (2.1), (2.2) and $\sum_{n=1}^{\infty}(1-C_n) < \infty$.

In particular, if we use the bound (2.12) on the VC dimension on classes of neural networks then we obtain the following corollary (which generalises the corresponding result from [29]).

Corollary 2.5. *Let Γ be a classifier that minimises empirical error over the class $\mathcal{C}^{k_n}$, where $\mathcal{C}^{k_n}$ is the class of neural net classifiers with k_n nodes in the hidden layer and the threshold sigmoid, and $k_n \to \infty$ so that $k_n \log n/n \to 0$ as $n \to \infty$. Let $\mathbf{P}$ be any distribution on $\mathbf{X}^\infty$ satisfying (2.1) and (2.2) such that $\sum_{n=1}^{\infty}(1 - C_n) < \infty$. Then*

$$\lim_{n\to\infty} \operatorname{err}_n(\Gamma) = 0 \quad \mathbf{P}\text{–}a.s.$$

2.3 Application to classical nonparametric predictors

In this section we use two types of classical nonparametric predictors: partitioning and nearest neighbour classifiers, to show how weak consistency results can be generalised to the conditional model. In the first subsection we introduce the Nearest Neighbour and partitioning rules, and present in the next section our generalised consistency results.

2.3.1 Nearest Neighbour and partitioning estimators

Nearest neighbours rule, first introduced in [18, 19], has attracted attention of many researchers as a very simple yet powerful classification tool. Main results which are referred and used in the present section have been obtained in [8, 43, 14]. See also [12], Chapter 5 for an extensive overview.

Nearest Neighbour predictor assigns to a new object x the label of its nearest neighbours among $x_1, \dots, x_n$:

$$\Gamma_n(x_1, y_1, \dots, x_n, y_n, x) := y_j,$$

where $j := \operatorname{argmin}_{i=1,\dots,n} \|x - x_i\|$.

For i.i.d. distributions this predictor is weakly consistent, i.e.

$$E(\mathrm{err}_n(\Gamma, P^\infty)) \to 0$$

for any distribution P on $\mathbf{X}$. (This result follows from the results of [43] and [8].)

A partitioning predictor on each step n partitions the object space $\mathbf{X} = \mathbb{R}^d$, $d \in \mathbb{N}$ into disjoint cells $A_1^n, A_2^n, \dots$ and classifies in each cell according to the majority vote:

$$\Gamma(z_1, \dots, z_n, x) := \begin{cases} 0 & \text{if } \sum_{i=1}^n I_{y_i=1} I_{x_i \in A(x)} \le \sum_{i=1}^n I_{y_i=0} I_{x_i \in A(x)} \\ 1 & \text{otherwise,} \end{cases}$$

where $A(x)$ stands for the cell containing x. Introduce

$$\mathrm{diam}(A) := \sup_{x,y \in A} \|x - y\|$$

and

$$N(x) := \sum_{i=1}^{n} I_{x_i \in A(x)}.$$

It is a well known result (see, e.g. [12], Theorem 6.1) that a partitioning predictor is weakly consistent, provided certain regulatory conditions on the size of cells. More precisely, let Γ be a partitioning predictor such that $\operatorname{diam}(A(X)) \to 0$ in probability and $N(X) \to \infty$ in probability. Then for any distribution P on $\mathbf{X}$

$$E(\mathrm{err}_n(\Gamma, P^\infty)) \to 0.$$

A simple example of partitioning estimate is the cubic histogram rule. This rule partitions the object space $\mathbb{R}^d$ into sets of the type

$$\prod_{i=1}^{d} [k_i h_n, (k_i + 1)h_n)$$

where k_i's are integers. Such predictor is consistent (i.e. the probability of error tends to zero) if $h_n \to 0$ and $nh_n^d \to \infty$ as $n \to \infty$, see [21, 22].

We note that more general and strong (e.g. strong consistency) results exist for the described rules and their generalisations (see e.g. [14],[28]). However, we do not aim to generalise the state-of-the-art results in nonparametric classification, but rather to illustrate that weak consistency results can be extended to the conditional model.

2.3.2 Extension to the conditional model

We present the following generalisation of the results on nonparametric classification.

Theorem 2.6. *Let Γ be the nearest neighbour classifier. Let $\mathbf{P}$ be some* *distribution on* $\mathbf{X}^\infty$ *satisfying (2.1), (2.2) and (2.7). Then*

$$E(\mathrm{err}_n(\Gamma, \mathbf{P})) \to 0.$$

The proofs for this section can be found in Section 2.7.

Theorem 2.7. *Let* Γ *be a partitioning predictor such that* $\mathrm{diam}(A(X)) \to 0$ *in probability and* $N(X) \to \infty$ *in probability, for any distribution generating i.i.d. examples. Then*

$$E(\mathrm{err}_n(\Gamma, \mathbf{P})) \to 0.$$

for any distribution $\mathbf{P}$ *on* $\mathbf{X}^\infty$ *satisfying (2.1), (2.2) and (2.7).*

2.4 Discussion of the conditions of the model

In the section 2.1 we have introduced the "conditionally i.i.d." model for pattern recognition which generalises the commonly used i.i.d. model. Naturally, a question arises whether our conditions on the distributions and on predictors are necessary, or they can be yet more generalised in the same direction. In this section we discuss the conditions of the new model from this point of view.

The first question is: Can the same bounds on the probability of error in the conditional model be achieved without assumptions on tolerance to data? The following negative example shows that some bounds on tolerance to data are necessary.

Remark 2.8. *There exists a distribution* $\mathbf{P}$ *on* $\mathbf{X}^\infty$ *satisfying (2.1) and (2.2) such that* $\mathbf{P}(|p_n - 1/2| > 3/n) = 0$ *for any* n *(i.e.* $C_n = 1$ *for any* $\delta \in (0, 1/2)$ *and* $n > \frac{3}{(1/2-\delta)}$*) and a predictor* Γ *such that* $P_p^n(\mathrm{err}_n > 0) \leq 2^{1-n}$ *for any* $p \in [\delta, 1-\delta]$ *and* $\mathbf{P}(\mathrm{err}_n = 1) = 1$ *for* $n > 1$*.*

Proof. Let $\mathbf{X} = \mathbf{Y} = \{0, 1\}$. We define the distributions P_y as $P_y(X = y) = 1$, for each $y \in \mathbf{Y}$ (i.e. $\eta(x) = x$ for each x). The distribution $\mathbf{P}|_{\mathbf{Y}^\infty}$ is defined as a Markov distribution with transition probability matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, i.e. it always generates sequences of labels $\dots 01010101 \dots$.

We define the predictor Γ as follows

$$\Gamma_n := \begin{cases} 1 - x_n & \text{if } |\#\{i < n : y_i = 0\} - n/2| \leq 1, \\ x_n & \text{otherwise.} \end{cases}$$

So, in the case when the distribution $\mathbf{P}$ is used to generate the examples, Γ is always seeing either $n-1$ zeros and n ones, or n zeros and n ones which, consequently, will lead it to always predict the wrong label. It remains to note that this is almost improbable in the case of an i.i.d. distribution. □

Another point is the requirement on the frequencies of labels. In particular, the assumption (2.7) might appear redundant: if the rate of occurrence of some label tends to zero, can we just ignore this label without affecting the asymptotic? It appears that this is not the case, as the following example illustrates.

Remark 2.9. *There exists a distribution* $\mathbf{P}$ *on* $\mathbf{X}^\infty$ *which satisfies (2.1) and (2.2) but for which the nearest neighbour predictor is not consistent, i.e. the probability of error does not tend to zero.*

Proof. Let $\mathbf{X} = [0, 1]$, let $\eta(x) = 0$ if x is rational and $\eta(x) = 1$ otherwise. The distribution P_1 is uniform on the set of irrational numbers, while P_0 is any distribution such that $P(x) \neq 0$ for any rational x. (This construction is due to T. Cover.) The nearest neighbour predictor is consistent for any i.i.d. distribution which agrees with the definition, i.e. for any $p = P(Y = 1) \in [0, 1]$.

Next we construct the distribution $\mathbf{P}|_{\mathbf{Y}^\infty}$. Fix some ε, $0 < \varepsilon < 1$. Assume that according to $\mathbf{P}$ the first label is always 1, i.e. $\mathbf{P}(y_1 = 1) = 1$ (the object is an irrational number). Next k_1 labels are always 0 (rationals), then follows 1, then k_2 zeros, and so on. It is easy to check that there exists such sequence $k_1, k_2, \ldots$ that with probability at least ε we have

$$\max_{i<n:\ X_i \text{ is irrational}} P_1\{x : X_i \text{ is the nearest neighbour of } x\} \leq$$

where $m(n)$ is the total number of irrational objects up to the trial n. On each step n such that $n = t + \sum_{j=1}^{t} k_t$ for some $t \in \mathbb{N}$ (i.e. on each irrational object) we have

$$\begin{aligned} &E(\mathrm{err}_n(\Gamma, \mathbf{P})) \\ &\geq \varepsilon \left(1 - \sum_{j<n:\ X_j \text{ is irrational}} \mathbf{P}(X_j \text{ is the nearest neighbour of } X) \right) \geq \varepsilon^2 \end{aligned}$$

As irrational objects are generated infinitely often (that is, with intervals k_i), the probability of error does not tend to zero. □

Another question is whether the results can be generalised to the case of non-deterministically defined labels, which is often considered in literature. It should be noted that we consider the task of learning object-label dependence, ignoring the label-label dependence (and prohibiting any dependence apart from these). On one hand, it enables us to allow any sort of label-label dependence. On the other hand, the best bound on the probability of error we can obtain is is the maximum of the class-conditional probabilities of error (as nothing is known about the probability of the next label), and not the so-called Bayes error, which is the best achievable bound in the i.i.d. case.

Thus, if we want to consider stochastically defined labels, we should restrict our attention to class-conditional probabilities of error. On this way also several obstacles can be met. In particular, the function η, which in this case is defined as $\eta(x) := \mathbf{P}(Y_n = 1 | X_n = x)$ should not depend on n, which will require more restrictive definition of constants C_n and the condition (2.7). We leave this question for further investigation.

One more point which needs clarification is the choice of the constants $\varkappa_n$. We fixed these constants for the sake of simplicity of notation, however, they can be made variable. Specifically, for the asymptotic results the following condition is required:

$$\lim_{n\to\infty} \{n|p_n - p| \leq \varkappa_n\} = 0$$

almost surely for any $p \in (0,1)$ and any probability distribution P on $\mathbf{X}$ such that $P(y=1) = p$, where $p_n := \frac{1}{n}\#\{i \leq n : y_i = 0\}$.

2.5 Proofs for Section 2.1

Before proceeding with the proof of Theorem 2.1 we give some definitions and supplementary facts.

Define the conditional probabilities of error of Γ as follows

$$\mathrm{err}_n^0(\Gamma, \mathbf{P}, z_0, \dots, z_n) := \mathbf{P}(Y_{n+1} \neq \Gamma(z_1, \dots, z_n, X_{n+1}) | Y_{n+1} = 0),$$

$$\mathrm{err}_n^1(\Gamma, \mathbf{P}, z_0, \dots, z_n) := \mathbf{P}(Y_{n+1} \neq \Gamma(z_1, \dots, z_n, X_{n+1}) | Y_{n+1} = 1)$$

(with the same notational convention as used with the definition of $\mathrm{err}_n(\Gamma)$). In words, for each $y \in \mathbf{Y} = \{0,1\}$ we define err_n^y as the probability of all $x \in \mathbf{X}$, such that Γ makes an error on n'th trial, given that $Y_{n+1} = y$ and fixed $z_1, \dots, z_n$.

For any $\mathbf{y} := (y_1, y_2, \dots) \in \mathbf{Y}^\infty$, define $\mathbf{y}_n := (y_1, \dots, y_n)$ and $p_n(\mathbf{y}) := \frac{1}{n}\#\{i \leq n : y_i = 0\}$, for $n > 1$.

Clearly (from the assumption (2.1)) the random variables $X_1, \dots, X_n$ are mutually conditionally independent given $Y_1, \dots, Y_n$, and by (2.2) they are distributed according to P_{Y_i}, $1 \leq i \leq n$. Hence, the following statement is valid.

Lemma 2.10. *Fix some $n > 1$ and some $\mathbf{y} \in \mathbf{Y}^\infty$ such that*

$$\mathbf{P}((Y_1, \ldots Y_{n+1}) = \mathbf{y}_{n+1}) \neq 0.$$

Then

$$\mathbf{P}\big(\operatorname{err}_n^{y_{n+1}}(\Gamma) > \varepsilon \,\big|\, (Y_1, \ldots, Y_n) = \mathbf{y}_n\big) \\ = P_p^n\big(\operatorname{err}_n^{y_{n+1}}(\Gamma) > \varepsilon \,\big|\, (Y_1, \ldots, Y_n) = \mathbf{y}_n\big)$$

for any $p \in (0, 1)$.

Proof of Theorem 2.1. Fix some $n > 1$, some $y \in \mathbf{Y}$ and such $\mathbf{y}^1 \in \mathbf{Y}^\infty$ that $n\delta \leq p_n(\mathbf{y}^1) \leq n(1-\delta)$ and $\mathbf{P}((Y_1, \ldots, Y_n) = \mathbf{y}^1_n) \neq 0$. Let $p := p_n(\mathbf{y}^1)/n$. We will find bounds on $\mathbf{P}\big(\operatorname{err}_n(\Gamma) > \varepsilon \mid (Y_1, \ldots, Y_n) = \mathbf{y}^1_n\big)$, first in terms of Δ and then in terms of $\bar{\Delta}$.

Lemma 2.10 allows us to pass to the i.i.d. case:

$$\mathbf{P}\big(\operatorname{err}_n^y(\Gamma, X_1, y^1_1, \ldots, X_n, y^1_n, X_{n+1}) > \varepsilon\big) \\ = P_p^n\big(\operatorname{err}_n^y(\Gamma, X_1, y^1_1, \ldots, X_n, y^1_n, X_{n+1}) > \varepsilon\big)$$

for any y such that $\mathbf{P}(Y_1 = y^1_1, \ldots, Y_n = y^1_n, Y_{n+1} = y) \neq 0$ (recall that we use upper-case letters for random variables and lower-case for fixed variables, so that the probabilities in the above formula are labels-conditional).

Clearly, for $\delta \leq p \leq 1-\delta$ we have $\operatorname{err}_n(\Gamma, P_p) \leq \max_{y \in \mathbf{Y}}(\operatorname{err}_n^y(\Gamma, P_p))$, and if $\operatorname{err}_n(\Gamma, P_p) < \varepsilon$ then $\operatorname{err}_n^y(\Gamma, P_p) < \varepsilon/\delta$ for each $y \in \mathbf{Y}$.

Let m be such number that $m - \varkappa_m = n$. For any $\mathbf{y}^2 \in \mathbf{Y}^\infty$ such that $|mp_m(\mathbf{y}^2) - mp| \leq \varkappa_m/2$ there exist such mapping $\pi : \{1, \ldots, n\} \to \{1, \ldots, m\}$ that $y^2_{\pi(i)} = y^1_i$ for any $i \leq n$. Define random variables $X'_1 \ldots X'_m$ as follows: $X'_{\pi(i)} := X_i$ for $i \leq n$, while the rest $\varkappa_m$ of X'_i are some random variables independent from $X_1, \ldots, X_n$ and from each other, and distributed according to P_p (a "ghost sample"). We have

$$\begin{aligned}
&P_p^n\big(\operatorname{err}_n^y(X_1, y_1^1, \dots, X_n, y_n^1) > \varepsilon\big) \\
&\quad = P_p^m\Big(\operatorname{err}_n^y(X_1, y_1^1, \dots, X_n, y_n^1) - \operatorname{err}_n^y(X_1', y_1^2, \dots, X_m', y_m^2) \\
&\quad + \operatorname{err}_n^y(X_1', y_1^2, \dots, X_m', y_m^2) > \varepsilon\Big) \\
&\le P_p^m\Big(\big|\operatorname{err}_n^y(X_1', y_1^2, \dots, X_n', y_n^2) - \operatorname{err}_n^y(X_1, y_1^1, \dots, X_n, y_n^1)\big| > \varepsilon/2\Big) \\
&\quad + P_p^n\Big(\operatorname{err}_n^y(X_1', y_1^2, \dots, X_n', y_n^2) > \varepsilon/2\Big).
\end{aligned}$$

Observe that $\mathbf{y}^2$ was chosen arbitrary (among sequences for which $|mp_m(\mathbf{y}^2) - mp| \le \varkappa_m/2$) and

$$(X_1, y_1^1, \dots, X_n y_n^1)$$

can be obtained from

$$(X_1', y_1^2, \dots, X_m' y_m^2)$$

by removing at most $\varkappa_m$ elements and applying some permutation. Thus the first term is bounded by

$$\begin{aligned}
&P_p^m\big(\max_{j \le \varkappa_m;\ \pi:\{1,\dots,m\}\to\{1,\dots,m\}} |\operatorname{err}_m^y(\Gamma, Z_1, \dots, Z_m) - \\
&\qquad \operatorname{err}_{m-j}^y(\Gamma, Z_{\pi(1)}, \dots, Z_{\pi(m-j)})| > \varepsilon/2 \,\big|\, |mp(m) - mp| \le \varkappa_m/2\big) \\
&\qquad\qquad \le \frac{\Delta(P_p, m, \delta\varepsilon/2)}{P_p^n(|mp(m) - mp| \le \varkappa_m)} \le \frac{1}{1 - 1/\sqrt{m}} \Delta(P_p, m, \delta\varepsilon/2),
\end{aligned}$$

and the second term is bounded by $\frac{1}{1-1/\sqrt{m}} P_p^m(\operatorname{err}_m(\Gamma) > \delta\varepsilon/2)$. Hence

$$\begin{aligned}
&P_p^n\big(\operatorname{err}_n^y(X_1, y_1^1, \dots, X_n, y_n^1) > \varepsilon\big) \\
&\qquad \le \alpha_n\big(\Delta(P_p, m, \delta\varepsilon/2) + P_p^m(\operatorname{err}_m(\Gamma) > \delta\varepsilon/2)\big). \quad (2.20)
\end{aligned}$$

Next we establish a similar bound in terms of $\bar{\Delta}$. For any $\mathbf{y}_n^2 \in \mathbf{Y}^n$ such that $|np_n(\mathbf{y}^2) - np| \leq \varkappa_n/2$ there exist such permutations π_1, π_2 of the set $\{1, \ldots, n\}$ that $y^1_{\pi_1(i)} = y^2_{\pi_2(i)}$ for any $i \leq n - \delta\varkappa_n$. Denote $n - \delta\varkappa_n$ by n' and define random variables $X'_1 \ldots X'_n$ as follows: $X'_{\pi_2(i)} := X_{\pi_1(i)}$ for $i \leq n'$, while for $n' < i \leq n$ X'_i are some "ghost" random variables independent from $X_1, \ldots, X_n$ and from each other, and distributed according to P_p. We have

$$\begin{aligned} &P_p^n\big(\operatorname{err}^y_n(X_1, y^1_1, \ldots, X_n, y^1_n) > \varepsilon\big) \\ &\quad \leq P_p^{n+\varkappa_n}\Big(\big|\operatorname{err}^y_n(X'_1, y^2_1, \ldots, X'_n, y^2_n) - \operatorname{err}^y_n(X_1, y^1_1, \ldots, X_n, y^1_n)\big| > \varepsilon/2\Big) \\ &\qquad\qquad + P_p^n\Big(\operatorname{err}^y_n(X'_1, y^2_1, \ldots, X'_n, y^2_n) > \varepsilon/2\Big), \end{aligned}$$

Again, as $\mathbf{y}^2$ was chosen arbitrary (among sequences for which $|np_n(\mathbf{y}^2) - np| \leq \varkappa_n/2$) and $(X_1, y^1_1, \ldots, X_n y^1_n)$ differs from $(X'_1, y^2_1, \ldots, X'_n y^2_n)$ in at most $\varkappa_n$ elements, up to some permutation. Thus the first term is bounded by

$$\begin{aligned} &P_p^n\big(\sup_{j<\varkappa_n;\pi:\{1,\ldots,n\}\to\{1,\ldots,n\};z'_{n-j},\ldots,z'_n} |\operatorname{err}^y_n(Z_1, \ldots, Z_n) \\ &\qquad - \operatorname{err}^y_n(\zeta_1, \ldots, \zeta_n)| > \varepsilon/2 \,\big|\, |np(n) - np| \leq \varkappa_n/2\big) \\ &\qquad\qquad \leq \alpha_n \bar{\Delta}(P_p, n, \delta\varepsilon/2), \end{aligned}$$

and the second term is bounded by $\alpha_n P_p^n(\operatorname{err}_n(\Gamma) > \delta\varepsilon/2)$. Hence

$$\begin{aligned} &P_p^n\big(\operatorname{err}^y_n(X_1, y^1_1, \ldots, X_n, y^1_n) > \varepsilon\big) \\ &\qquad \leq \alpha_n\big(\bar{\Delta}(P_p, n, \delta\varepsilon/2) + P_p^n(\operatorname{err}_n(\Gamma) > \delta\varepsilon/2)\big). \end{aligned} \tag{2.21}$$

Finally, as $\mathbf{y}^1$ was chosen arbitrary among sequences $\mathbf{y} \in \mathbf{Y}^\infty$ such that $n\delta \leq p_n(\mathbf{y}^1) \leq n(1 - \delta)$ from (2.20) and (2.21) we obtain (2.5) and (2.6). □

2.6 Proofs for Section 2.2

Proof of Theorem 2.3. Fix some probability distribution P_p and some $n \in \mathbb{N}$. Let $\varphi^\times$ be any decision rule $\varphi \in \mathcal{C}$ picked by $\Gamma_{n-\varkappa_n}$ on which (along with the corresponding permutation) the maximum

$$\max_{j \le \varkappa_n;\ \pi:\{1,\dots,n\}\to\{1,\dots,n\}} |\operatorname{err}_n(\Gamma, z_1, \dots, z_n) - \operatorname{err}_{n-j}(\Gamma, z_{\pi(1)}, \dots, z_{\pi(n-j)})|$$

is reached. We need to estimate $P^n(|\operatorname{err}(\varphi^*) - \operatorname{err}(\varphi^\times)| > \varepsilon)$.

Clearly, $|\overline{\operatorname{err}}_n(\varphi^\times) - \overline{\operatorname{err}}_n(\varphi^*)| \le \varkappa_n$, as $\varkappa_n$ is the maximal number of errors which can be made on the difference of the two samples.

Moreover,

$$\begin{aligned}
&P^n\big(|\operatorname{err}(\varphi_n^*) - \operatorname{err}(\varphi^\times)| > \varepsilon\big)\\
&\quad\le P^n\big(|\operatorname{err}(\varphi_n^*) - \frac{1}{n}\overline{\operatorname{err}}_n(\varphi^*)| > \varepsilon/2\big)\\
&\quad+P^n\big(|\frac{1}{n}\overline{\operatorname{err}}_n(\varphi^\times) - \operatorname{err}(\varphi^\times)| > \varepsilon/2 - \varkappa_n/n\big)
\end{aligned}$$

Observe that

$$P^n(\sup_{\varphi\in\mathcal{C}} |\frac{1}{n}\overline{\operatorname{err}}_n(\varphi) - \operatorname{err}(\varphi)| > \varepsilon) \le 8\mathcal{S}(\mathcal{C}, n)e^{-n\varepsilon^2/32}, \tag{2.22}$$

see [12], Theorem 12.6. Thus,

$$\Delta(P_p, n, \varepsilon) \le 16\mathcal{S}(\mathcal{C}, n)e^{-n(\varepsilon/2-\varkappa_n/n)^2/32} \le 16\mathcal{S}(\mathcal{C}, n)e^{-n\varepsilon^2/512}$$

for $n > 4/\varepsilon^2$. So,

$$\begin{aligned}
\mathbf{P}(\operatorname{err}_n(\Gamma, \mathbf{P}) > \varepsilon) &\le I_{\sup_{p\in[\delta,1-\delta]} \operatorname{err}(\varphi_{P_p}, P_p) > \varepsilon/2}\\
&+16\alpha C_n^{-1}\mathcal{S}(\mathcal{C}, n)e^{-n\delta^2\varepsilon^2/2048} + (1 - C_n).
\end{aligned}$$

It remains to notice that

$$\operatorname{err}(\varphi_{P_p}, P_p) = \inf_{\varphi\in\mathcal{C}}(p \operatorname{err}^1(\varphi, P_p) + (1-p)\operatorname{err}^0(\varphi, P_p))$$
$$\le \inf_{\varphi\in\mathcal{C}}(\operatorname{err}^1(\varphi, P_{1/2}) + \operatorname{err}^0(\varphi, P_{1/2})) = 2\operatorname{err}(\varphi_{P_{1/2}}, P_{1/2})$$

for any $p \in [0, 1]$.

So far we have proven (2.16) and (2.17); (2.18) and (2.19) can be proven analogously, only for the case $\eta \in \mathcal{C}$ we have

$$P^n(\sup_{\varphi\in\mathcal{C}} |\frac{1}{n}\overline{\operatorname{err}}_n(\varphi) - \operatorname{err}(\varphi)| > \varepsilon) \le \mathcal{S}(\mathcal{C}, n)e^{-n\varepsilon}$$

instead of (2.22), and $\operatorname{err}(\varphi_{P_p}, P_p) = 0$. □

2.7 Proofs for Section 2.3

The first part of the proof is common for Theorems 2.6 and 2.7. Let us fix some distribution $\mathbf{P}$ satisfying conditions of the theorems. It is enough to show that

$$\sup_{p\in[\delta,1-\delta]} E(\operatorname{err}_n(\Gamma, P_p, Z_1, \dots, Z_n)) \to 0$$

and

$$\sup_{p\in[\delta,1-\delta]} E(\bar\Delta(P_p, n, Z_1, \dots, Z_n)) \to 0$$

for nearest neighbour and partitioning predictor, and apply Corollary 2.2.

Observe that both predictors are symmetric, i.e. do not depend on the order of $Z_1, \dots, Z_n$. Thus, for any $z_1, \dots, z_n$

$$\bar\Delta(P_p, n, z_1, \dots, z_n) = \sup_{j\le\varkappa_n;\ \pi:\{1,\dots,n\}\to\{1,\dots,n\}, z'_{n-j},\dots,z'_n}$$
$$|\operatorname{err}_n(\Gamma, P_p, z_1, \dots, z_n) - \operatorname{err}_n(\Gamma, P_p, z_{\pi(1)}, \dots, z_{\pi(n-j)}, z'_{n-j}, \dots, z'_n)|,$$

where the maximum is taken over all z'_i consistent with η, $n-j \le i \le n$.

Define also the class-conditional versions of $\bar{\Delta}$:

$$\bar{\Delta}^y(P_p, n, z_1, \dots, z_n) := \sup_{j \le \varkappa_n;\ \pi:\{1,\dots,n\}\to\{1,\dots,n\}, z'_{n-j},\dots,z'_n}$$
$$|\operatorname{err}_n^y(\Gamma, P_p, z_1, \dots, z_n) - \operatorname{err}_n^y(\Gamma, P_p, z_{\pi(1)}, \dots, z_{\pi(n-j)}, z'_{n-j}, \dots, z'_n)|.$$

Note that (omitting $z_1, \dots, z_n$ from the notation)

$$\operatorname{err}_n(\Gamma, P_p) \le \operatorname{err}_n^0(\Gamma, P_p) + \operatorname{err}_n^1(\Gamma, P_p)$$

and

$$\bar{\Delta}(P_p, n) \le \bar{\Delta}^0(P_p, n) + \bar{\Delta}^1(P_p, n).$$

Thus, it is enough to show that

$$\sup_{p\in[\delta,1-\delta]} E(\operatorname{err}_n^1(\Gamma, P_p)) \to 0 \tag{2.23}$$

and

$$\sup_{p\in[\delta,1-\delta]} E(\bar{\Delta}^1(P_p, n)) \to 0. \tag{2.24}$$

Observe that for each of the predictors in question the probability of error given that the true label is 1 will not decrease if an arbitrary (possibly large) portion of training examples labelled with ones is replaced with an arbitrary (but consistent with η) portion of the same size of examples labelled with zeros. Thus, for any n and any $p \in [\delta, 1-\delta]$ we can decrease the number of ones in our sample (by replacing the corresponding examples with examples from the other class) down to (say) $\delta/2$, not decreasing the probability of error on examples labelled with 1. So,

$$E(\operatorname{err}_n^1(\Gamma, P_p)) \le E(\operatorname{err}_n^1(\Gamma, P_{\delta/2}|p_n = \delta/2)) + P_p(p_n \le \delta/2), \tag{2.25}$$

where as usual $p_n := \frac{1}{n}\#\{i \le n : y_i = 1\}$. Obviously, the last term (quickly) tends to zero. Moreover, it is easy to see that

$$\begin{aligned}
&E(\mathrm{err}_n^1(\Gamma, P_{\delta/2})|p_n = n(\delta/2)) \\
&\qquad \le E\big(\,\mathrm{err}_n^1(\Gamma, P_{\delta/2})\big||n(\delta/2) - p_n| \le \varkappa_n/2\big) + E(\bar\Delta^1(P_{\delta/2}, n)) \\
&\qquad\qquad\qquad \le \frac{1}{1 - 1/\sqrt{n}} E(\mathrm{err}_n^1(\Gamma, P_{\delta/2})) + E(\bar\Delta^1(P_{\delta/2}, n)). \quad (2.26)
\end{aligned}$$

The first term tends to zero, as it is known from the results for i.i.d. processes; thus, to establish (2.23) we have to show that

$$E(\bar\Delta^1(P_p, n, Z_1, \dots, Z_n)) \to 0 \qquad (2.27)$$

for any $p \in (0,1)$.

We will also show that (2.27) is sufficient to prove (2.24). Indeed,

$$\begin{aligned}
&\bar\Delta^1(P_p, n, z_1, \dots, z_n) \le \mathrm{err}_n^1(\Gamma, P_p, z_1, \dots, z_n) + \\
&\qquad \sup_{j \le \varkappa_n;\ \pi:\{1,\dots,n\}\to\{1,\dots,n\}, z'_{n-j},\dots,z'_n} \mathrm{err}_n^1(\Gamma, P_p, z_{\pi(1)}, \dots, z_{\pi(n-j)}, z'_{n-j}, \dots, z'_n)
\end{aligned}$$

Denote the last summand by D. Again, we observe that D will not decrease if an arbitrary (possibly large) portion of training examples labelled with ones is replaced with an arbitrary (but consistent with η) portion of the same size of examples labelled with zeros. Introduce $\widetilde\Delta^1(P_p, n, z_1, \dots, z_n)$ as $\bar\Delta^1(P_p, n, z_1, \dots, z_n)$ with $\varkappa_n$ in the definition replaced by $\frac{2}{\delta}\varkappa_n$. Using the same argument as in (2.25) and (2.26) we have

$$E(D) \le \frac{1}{1 - 1/\sqrt{n}}\big(E(\widetilde\Delta^1(P_{\delta/2}, n)) + E(\mathrm{err}_n(\Gamma, P_{\delta/2})\big) + P_p(p_n \le \delta/2).$$

Thus, (2.24) holds true if (2.27) and

$$E(\widetilde\Delta^1(P_p, n, Z_1, \dots, Z_n)) \to 0. \qquad (2.28)$$

Finally, we will prove (2.27); it will be seen that the proof of (2.28) is analogous (i.e. replacing $\varkappa_n$ by $\frac{2}{\delta}\varkappa_n$ does not affect the proof). Note that

$$E(\bar{\Delta}(P_p, n, Z_1, \dots, Z_n)) \le P_p\Big(\sup_{j \le \varkappa_n;\ \pi:\{1,\dots,n\}\to\{1,\dots,n\}, z'_{n-j},\dots,z'_n}$$
$$\big| \operatorname{err}_n(\Gamma, P_p, Z_1, \dots, Z_n) \ne \operatorname{err}_n(\Gamma, P_p, Z_{\pi(1)}, \dots, Z_{\pi(n-j)}, z'_{n-j}, \dots, z') \big| \Big),$$

where the maximum is taken over all z'_i consistent with η, $n - j \le i \le n$. The last expression should be shown to tend to zero. This we will prove for each of the predictors separately.

Nearest Neighbour predictor. Fix some distribution P_p, $0 < p < 1$ and some $\varepsilon > 0$. Fix also some $n \in \mathbb{N}$ and define (leaving $x_1, \dots, x_n$ implicit)

$$B_n(x) := P_p^{n+1}\{t \in \mathbf{X} : t \text{ and } x \text{ have the same}$$
$$\text{nearest neighbour among } x_1, \dots, x_n\}$$

and $B_n := E(B_n(X))$ Note that $E(B_n) = 1/n$, where the expectation is taken over $X_1, \dots, X_n$. Define $\mathcal{B} := \{(x_1, \dots, x_n) \in \mathbf{X}^n : B_n \le 1/n\varepsilon\}$ and $\mathcal{A}(x_1, \dots, x_n) := \{x : B_n(x) \le 1/n\varepsilon^2\}$. Applying Markov's inequality twice, we obtain

$$\begin{aligned} E(\bar{\Delta}(P_p, n)) &\le E(\bar{\Delta}(P_p, n) | (X_1, \dots, X_n) \in \mathcal{B}) + \varepsilon \\ &\le E\Big(\sup_{j \le \varkappa_n;\ \pi:\{1,\dots,n\}\to\{1,\dots,n\}, z'_{n-j},\dots,z'_n} \end{aligned}$$
$$P_p\{x : \operatorname{err}_n(\Gamma, P_p, Z_1, \dots, Z_n) \ne \operatorname{err}_n(\Gamma, P_p, Z_{\pi(1)}, \dots, Z_{\pi(n-j)}, z'_{n-j}, \dots, z'_n)$$
$$\big| x \in \mathcal{A}(X_1, \dots, X_n)\} \big| (X_1, \dots, X_n) \in \mathcal{B}\Big) + 2\varepsilon. \tag{2.29}$$

Removing one point x_i from a sample $x_1, \dots, x_n$ we can only change the value of Γ in the area

$$\{x \in \mathbf{X} : x_i \text{ is the nearest neighbour of x}\} = B_n(x_i),$$

while adding one point x_0 to the sample we can change the value of Γ in the area

$$D_n(x_0) := \{x \in \mathbf{X} : x_0 \text{ is the nearest neighbour of x}\}.$$

It can be shown that the number of examples (among $x_1, \dots, x_n$) for which a point x_0 is the nearest neighbour is not greater than a constant γ which depends only the space $\mathbf{X}$ (see [12], Corollary 11.1). Thus,

$$D_n(x_0) \subset \cup_{i=j_1,\dots,j_\gamma} B_n(x_i)$$

for some $j_1, \dots, j_\gamma$, and so

$$\begin{aligned} E(\bar{\Delta}(P_p, n)) &\le 2\varepsilon + 2(\gamma+1)\varkappa_n E(\max_{x \in \mathcal{A}(X_1,\dots,X_n)} B_n(x) | (X_1, \dots, X_n) \in \mathcal{B}) \\ &\le 2\varkappa_n \frac{\gamma+1}{n\varepsilon^2} + 2\varepsilon, \end{aligned}$$

which, increasing n, can be made less than 3ε. □

Partitioning predictor. For any measurable sets $\mathcal{B} \subset \mathbf{X}^n$ and $\mathcal{A} \subset \mathbf{X}$ define

$$D(\mathcal{B}, \mathcal{A}) := E\Big(\sup_{j \le \varkappa_n;\ \pi:\{1,\dots,n\}\to\{1,\dots,n\}, z'_{n-j},\dots,z'_n}$$
$$P_p\big\{x : \mathrm{err}_n(\Gamma, P_p, Z_1, \dots, Z_n) \ne \mathrm{err}_n(\Gamma, P_p, Z_{\pi(1)}, \dots, Z_{\pi(n-j)}, z'_{n-j}, \dots, z'_n)$$
$$\big| x \in \mathcal{A}\big\} \big| (X_1, \dots, X_n) \in \mathcal{B}\Big) + 2\varepsilon.$$

and $D := D(\mathbf{X}^n, \mathbf{X})$.

Fix some distribution P_p, $0 < p < 1$ and some $\varepsilon > 0$. Introduce

$$\hat{\eta}(x, X_1, \dots, X_n) := \frac{1}{N(x)} \sum_{i=1}^n I_{Y_i=1} I_{X_i \in A(x)}$$

($X_1, \dots X_n$ will usually be omitted). From the consistency results for i.i.d. model (see, e.g. [12], Theorem 6.1) we know that $E^{n+1}|\hat{\eta}_n(X) - \eta(X)| \to 0$

(the upper index in E^{n+1} indicating the number of examples it is taken over).

Thus, $E|\hat{\eta}_n(X) - \eta(X)| \leq \varepsilon^4$ from some n on. Fix any such n and let $\mathcal{B} := \{(x_1, \ldots, x_n) : E^1|\hat{\eta}_n(X) - \eta(X)| \leq \varepsilon^2\}$. By Markov inequality we obtain $P_p(\mathcal{B}) \geq 1 - \varepsilon^2$. For any $(x_1, \ldots, x_n) \in \mathcal{B}$ let $\mathcal{A}(x_1, \ldots, x_n)$ be the union of all cells A_i^n for which $E^1(|\hat{\eta}_n(X) - \eta(X)||X \in A_i^n) \leq \varepsilon$. Clearly, with $x_1, \ldots, x_n$ fixed, $P_p(X \in \mathcal{A}(x_1, \ldots, x_n)) \geq 1 - \varepsilon$. Moreover, $D \leq D(\mathcal{B}, \mathcal{A}) + \varepsilon + \varepsilon^2$.

Fix $\mathcal{A} := (x_1, \ldots, x_n)$ for some $(x_1, \ldots, x_n) \in \mathcal{B}$. Since $\eta(x)$ is always either 0 or 1, to change a decision in any cell $A \subset \mathcal{A}$ we need to add or remove at least $(1-\varepsilon)N(A)$ examples, where $N(A) := N(x)$ for any $x \in A$. Let $N(n) := E(N(X))$ and $A(n) := E(P_p(A(X))$. Clearly, $\frac{N(n)}{nA(n)} = 1$ for any n, as $E\frac{N(X)}{n} = A(n)$.

As before, using Markov inequality and shrinking $\mathcal{A}$ if necessary we can have $P_p(\frac{\varepsilon^2 nA(X)}{N(n)} \leq \varepsilon|X \in \mathcal{A}) = 1$, $P_p(\frac{\varepsilon^2 nA(n)}{N(X)} \leq \varepsilon|X \in \mathcal{A}) = 1$, and $D \leq D(\mathcal{B}, \mathcal{A}) + 3\varepsilon + \varepsilon^2$. Thus, for all cells $A \subset \mathcal{A}$ we have $N(A) \geq \varepsilon nA(n)$, so that the probability of error can be changed in at most $2\frac{\varkappa_n}{(1-\varepsilon)\varepsilon nA(n)}$ cells; but the probability of each cell is not greater than $\frac{N(n)}{\varepsilon n}$. Hence $E(\bar{\Delta}(P_p, n)) \leq 2\frac{\varkappa_n}{n(1-\varepsilon)\varepsilon^2} + 3\varepsilon + \varepsilon^2$. □

Chapter 3

Online learning with weak teachers

3.1 Preliminaries

In this section we briefly introduce the online learning model and region predictors. We also provide some background on Transductive Confidence Machines, which is importand for understending the main results.

3.1.1 Notation

As before, we consider a sequence of object-label examples

$$(x_1, y_1), (x_2, y_2), \ldots;$$

In this chapter we consider a slightly more general situation, where labels can be defined probabilistically (rather then deterministically). However, only i.i.d. generated examples are considered.

That is, we just assume that examples are drawn according to some probability distribution P^∞ on $\mathbf{Z}^\infty$.

Also in this chapter we work with a generalised vesrion of predictors, so-called region predictors. Traditionally defined predictors which output one label are considered as a particular case of a region predictors: one-label prediction can be thought of as a region consisting of one element.

A *region predictor* is a measurable function

$$\Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n),$$

taking values in $\mathcal{P}(Y)$ where $n \in \mathbb{N}$, the $(x_i, y_i) \in \mathbf{Z}$, $i = 1, \ldots, n-1$ are examples and $x_n \in \mathbf{X}$ is an object.

For the cases when we are interested in prediction with confidence, the predictor is given an extra input $\gamma := (1 - \delta) \in (0, 1)$ which is called the *confidence level*; the complementary value δ is called the *significance level.* In this case, we assume that

$$\Gamma_{\gamma_1}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n) \subseteq \Gamma_{\gamma_2}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n)$$

whenever $\gamma_1 \leq \gamma_2$.

An important modification of the definition of a region predictor is where it is allowed to depend on additional inputs, random numbers $\tau_i \in [0, 1]$ (τ_i are assumed to be independently distributed according to the uniform distribution in $[0, 1]$ and to be independent of the examples); however, this case reduces to the case of deterministic region predictors by extending the object space $\mathbf{X}$ to $\mathbf{X} \times [0, 1]$, so that τ_i becomes an element of the extended object x_i. Therefore, we need not mention the random numbers τ_i explicitly.

A predictor is called *symmetric* if its predictions do not depend on the order of the examples learnt so far: if Γ is a (region) predictor then

$$\Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n) = \Gamma(x_{\pi(1)}, y_{\pi(1)}, \ldots, x_{\pi(n-1)}, y_{\pi(n-1)}, x_n)$$

for any permutation π of the set $\{1, \ldots, n-1\}$.

We often use the notation Γ_n instead of

$$\Gamma_{1-\delta}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n)$$

when the values of other parameters are clear.

The number of errors $\mathrm{Er}_n(\Gamma_{1-\delta})$ which Γ makes up to the trial $n \in \mathbb{N}$ is defined as

$$\#\{i = 1, \ldots, n : y_i \notin \Gamma_{1-\delta}(x_1, y_1, \ldots, x_{i-1}, y_{i-1}, x_i)\}.$$

The indicator of an individual error er_n at trial n is defined to be 1 if $y_n \notin \Gamma_n$ and 0 otherwise.

Similarly, the number of uncertain predictions $\mathrm{Un}_n(\Gamma)$ that $\Gamma_{1-\delta}$ makes up to the trial $n \in N$ is defined to be

$$\#\{i = 1, \ldots, n : |\Gamma_{1-\delta}(x_1, y_1, \ldots, x_{n-1}, y_{i-1}, x_i)| > 1\},$$

and the indicator un_n of uncertain prediction at individual trial n to be 1 if $|\Gamma_n| > 1$ and 0 otherwise.

A region predictor is called *well-calibrated* if, for any $\delta \in (0, 1)$,

$$\lim_{n \to \infty} \frac{\mathrm{Er}_n(\Gamma_{1-\delta})}{n} \to \delta \quad \text{a.s.}$$

under any probability distribution P^∞ generating the examples. Note that in the last definition a predictor is required to achieve exactly δ rate of errors for all distributions, including very clean ones. This means that sometimes it deliberately outputs wrong (e.g. empty) predictions. The merits of such definition versus the definition where the rate of errors should be less than or equal to δ are discussed in [54]; here we only note that all the results can be extended to the latter case, and consider the former case as more simple to analyse.

3.1.2 Transductive Confidence Machines

In this section we breifly introduce TCMs, following [51] in our exposition. For understending the rest of this chapter some explanation of what TCM is and how it works is important. However, formally, in what follows this section we only use the fact that TCMs are symmetric and well-calibrated.

A TCM is a way of constructing region predictor from a machine learning algorithm. A learning algorithm is implemented into a TCM in the form of individual strangeness measure, which is defined as follows.

A family of functions $\{A_n : n \in \mathbb{N}\}$ where $A_n : \mathbf{Z}^n \to \mathbb{R}^n$ is called *individual strangennes measure* if for any $n \in \mathbb{N}$, any $z_1, \ldots, z_n \in \mathbb{R}$, any $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and any permutation π

$$(\alpha_1, \ldots, \alpha_n) = A_n(z_1, \ldots, z_n) \Rightarrow (\alpha_{\pi(1)}, \ldots, \alpha_{\pi(n)}) = A_n(z_{\pi(1)}, \ldots, z_{\pi(n)}).$$

In other words, A_n it preserves the order.

A TCM associated with individual strangeness measure A_n is the following region predictor:

$$\Gamma_{1-\delta}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n)$$

is defined as a set of all labels $y \in \mathbf{Y}$ such that

$$\frac{1}{n}\#\{i = 1, \ldots, n : \alpha_i \geq \alpha_n\} > \delta,$$

where $(\alpha_1, \ldots, \alpha_n) := A_n(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, y)$.

Thus, a TCM "tries" each possible label with the current object, and includes it in the prediction if its strangeness is small, among strangeness of the examples of the training set.

As an example, a strangeness measure based on the Nearest Neighbour rule can be defined as follows:

$$\alpha_i := \frac{min_{j \neq i: y_j = y_i} d(x_i, x_j)}{min_{j \neq i: y_j \neq y_i} d(x_i, x_j)},$$

where d is the Eucledean distance. Thus, an object will appear strange if it is amidst the objects labelled differently. Other learning algorithms can be used to construct strangeness measures; examples include k-Nearest Neighbour rules, Support Vector Machines, Neural Networks, etc.

As it was already noted, a TCM is well-calibrated, i.e. achives any predefined level of erroneousness (rate of errors). Thus, to compare different TCMs (that is, different strangeness measures), the rate of uncertain predictions should be used as a measure of performance. Observer that there is an optimal region predictor among TCMs, i.e. the predictor which achives minimal asimptotic rate of uncertain predictions. Such a TCM was constructed in [52] based on the nearest neighbours strangeness measure.

3.2 Weak Teachers scenario

This section presents our generalisation of on-line learning scenario, allowing delayed and omitted labels. We find some sufficient conditions on the parameters of the new scenario under which asymptotic performance of a predictor is preserved.

3.2.1 Weak Deterministic Teachers

We suggest the following deterministic modified scenario for online prediction.

We call a sequence $\mathfrak{L} := \big((l_i, k_i) : i \in \mathbb{N}\big)$, where $(l_i, k_i) \in \mathbb{N} \times \mathbb{N}$, $i \in \mathbb{N}$ a *(deterministic) learning rule* if $k_i \leq n_i := \sum_{j=1}^{i} l_j$ for all $i \in N$, and $i \neq j$ implies $k_i \neq k_j$ for all $i, j \in \mathbb{N}$. The symbol N is used for the set $\{n_1, n_2, \dots\}$. Define the total amount of information available at the beginning of trial

n to a prediction algorithm taught according to the learning function $\mathfrak{L}$ as $s(n) := \#\{i \in N : i < n\}$.

In this definition, the numbers l_i specify the delays with which true labels are given, so that n_i are the numbers of trials on which the labels are disclosed, while k_i are the numbers of actual labels.

Suppose that $\Gamma_{1-\delta}$ is an online predictor and $\mathfrak{L}$ is a learning rule. Then we define the *$\mathfrak{L}$-taught version of* $\Gamma_{1-\delta}$ as follows:

$$\Gamma^{\mathfrak{L}}_{1-\delta}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n) := \Gamma_{1-\delta}(x_{k_1}, y_{k_1}, \ldots, x_{k_{s(n)}}, y_{k_{s(n)}}, x_n).$$

So at the end of each trial $n \in N$ the predictor $\Gamma^{\mathfrak{L}}_{1-\delta}$ "learns" the label y_n if $n \in N$ and "learns" nothing otherwise. The sequence $(l_i : i \in \mathbb{N})$ specifies the intervals in which labels are given, while the sequence $(k_i : i \in \mathbb{N})$ specifies the labels given on corresponding steps.

Consider several examples.

Ideal teacher. If $l_n = 1$ and $k_n = n$ for each $n \in \mathbb{N}$, then $\Gamma^{\mathfrak{L}}_{1-\delta}$ is equal to $\Gamma_{1-\delta}$.

Slow teacher with a fixed lag. If $l_n = d + 1$ for some $d \in \mathbb{N}$ and $k_n = n + d$ for $n \in N$, then $\Gamma^{\mathfrak{L}}_{1-\delta}$ is a predictor which learns true labels with the delay d.

Slow teacher. The previous example can be generalised as follows. Let $l_n = \mathrm{lag}(n)$, $n \in \mathbb{N}$ where $\mathrm{lag} : \mathbb{N} \to \mathbb{N} \cup \{0\}$ is an arbitrary function and let $k_n = n$ for all $n \in N$. Then $\Gamma^{\mathfrak{L}}_{1-\delta}$ models a predictor which learns the true label for each example x_n with the delay $\mathrm{lag}(n)$. This is what we call a *region predictor with slow teacher with the delay* $\mathrm{lag}()$.

Lazy teacher. Suppose that $N \neq \mathbb{N}$ and $l_n = 1$, for all $n \in N$; then $\Gamma^{\mathfrak{L}}_{1-\delta}$ is a *region predictor with lazy teachers*: it is given true labels immediately but not on every step.

Prior to stating the main theorem about $\mathfrak{L}$-taught predictors we need to give one more definition. If Γ is a (region) predictor, set

$$\overline{\mathrm{Er}}(\Gamma) := \left[\liminf_{n\to\infty} \frac{\mathrm{Er}_n(\Gamma)}{n}, \limsup_{n\to\infty} \frac{\mathrm{Er}_n(\Gamma)}{n}\right].$$

If Γ is a region predictor, set

$$\overline{\mathrm{U}}(\Gamma) := \left[\liminf_{n\to\infty} \frac{\mathrm{Un}_n(\Gamma)}{n}, \limsup_{n\to\infty} \frac{\mathrm{Un}_n(\Gamma)}{n}\right].$$

The intervals $\overline{\mathrm{Er}}(\Gamma)$ and $\overline{\mathrm{U}}(\Gamma)$ characterise the asymptotical error and uncertainty rates of Γ correspondingly; of course, these are random intervals, since they depend on the actual sequence of examples. It turns out, however, that in the case of symmetric predictors these intervals are close to being deterministic.

Lemma 3.1. *For each symmetric region predictor* Γ *and probability distribution* P *in* $\mathbf{Z}$ *there exist intervals* $[a_1, b_1] \subseteq \mathbb{R}$ *and* $[a_2, b_2] \subseteq \mathbb{R}$ *such that* $\overline{\mathrm{Er}}(\Gamma) = [a_1, b_1]$ *and* $\overline{\mathrm{U}}(\Gamma) = [a_2, b_2]$ P^∞*-almost surely.*

Proof. The statement of this lemma is an immediate consequence of the Hewitt-Savage zero-one law (see, e.g., [42]). □

We will use the notations $\overline{\mathrm{Er}}(\Gamma, P)$ and $\overline{\mathrm{U}}(\Gamma, P)$ for the intervals whose existence is asserted in the lemma.

Since TCMs are calibrated and so $\overline{\mathrm{Er}}(\Gamma_{1-\delta}, P)$ is just $\{\delta\}$ for any P for any TCM, we do not consider $\overline{\mathrm{Er}}(\Gamma, P)$ for TCMs.

We call $\overline{\mathrm{U}}(\Gamma, P)$ the *asymptotical uncertainty* of Γ with examples distributed according to P.

Theorem 3.2. *Let* $\Gamma_{1-\delta}$ *be a symmetric (region) predictor, let* $\mathfrak{L}$ *be a deterministic learning rule, and let* $\Gamma^{\mathfrak{L}}_{1-\delta}$ *be the* $\mathfrak{L}$*-taught version of* $\Gamma_{1-\delta}$*. The following statements hold for any probability distribution* P^∞ *generating the examples.*

- *If* $\Gamma_{1-\delta}$ *is a TCM and*
  $$\sum_{i=2}^{\infty} \frac{l_i^2}{n_i^2} < \infty \tag{3.1}$$
  then $\Gamma^{\mathfrak{L}}_{1-\delta}$ *is well calibrated.*

- *If for some* $l \in \mathbb{N}$, $l_i = l$ *from some* i *on, then* $\overline{\mathrm{U}}(\Gamma^{\mathfrak{L}}_{1-\delta}, P) = \overline{\mathrm{U}}(\Gamma_{1-\delta}, P)$ *and* $\overline{\mathrm{Er}}(\Gamma^{\mathfrak{L}}_{1-\delta}, P) = \overline{\mathrm{Er}}(\Gamma_{1-\delta}, P)$.

The proofs are provided in Section 3.4. A discussion of cases in which the first statement of the theorem is satisfied is given in Section 3.3.

3.2.2 Weak Randomised Teachers

We define a *randomised learning rule* as a sequence

$$\mathfrak{L} := \big((l_i, k_i) : i \in \mathbb{N}\big)$$

of random variables (l_i, k_i), $i \in \mathbb{N}$ distributed according to some probability distribution L on $(\mathbb{N} \times \mathbb{N})^\infty$, such that

i) random variables (l_i, k_i), $i \in \mathbb{N}$ are independent of examples, i.e. of z_i, $i \in \mathbb{N}$,

ii) random variables l_i, $i \in \mathbb{N}$ are independent of each other,

iii) for all $i, j \in \mathbb{N}$ if $i \neq j$ then $L(k_i = k_j) = 0$,

iv) $L(k_i > n_i) = 0$, for all $i \in \mathbb{N}$, where $n_i = \sum_{j=1}^{i} l_j$.

Obviously, such distributions exist; for instance, l_i, $i \in \mathbb{N}$ can be any sequence of i.i.d. random variables distributed on $\mathbb{N}$ (and independent of examples), while $k_i = n_i$ with probability 1.

The total amount of information available to a prediction algorithm taught according to the randomised learning rule is defined as the random variable $s(n) := \max\{i : n_i < n\}$.

Suppose that $\Gamma_{1-\delta}$ is an online predictor and $\mathfrak{L}$ is a randomised learning rule. We define the $\mathfrak{L}$ *-taught version of* $\Gamma_{1-\delta}$ as follows:

$$\Gamma^{\mathfrak{L}}_{1-\delta}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) := \Gamma_{1-\delta}(x_{k_1}, y_{k_1}, \dots, x_{k_{s(n)}}, y_{k_{s(n)}}, x_n).$$

As in the case of deterministic learning rules, the variables l_i, $i \in \mathbb{N}$ define the intervals with which new labels are given, while the variables k_i specify the labels which are given to a predictor on corresponding steps; for every $i \in \mathbb{N}$ on step n_i the label y_{k_i} is revealed.

Consider several examples.

Ideal teacher. Again, we start with a degenerate example. If $l_i = 1$ and $k_i = n_i$ with probability 1 for each $i \in \mathbb{N}$, then $\Gamma^{\mathfrak{L}}_{1-\delta}$ is equal to $\Gamma_{1-\delta}$.

Simple Bernoulli teacher. Suppose that $L(l_i = n) = p(1-p)^{n-1}$ for all $n, i \in \mathbb{N}$ and some $p \in (0, 1)$, and let k_i, $i \in \mathbb{N}$ be any random variables satisfying the definition of a randomised learning rule. Then on each step some past label is revealed to a predictor with probability p.

Bernoulli teacher. If in the previous example we allow p to vary, i.e. $L(l_i = n) = p^i_n \prod_{j=1}^{n-1}(1 - p^i_j)$ for some $p^i_j \in \mathbb{N}$, for all $n, i \in \mathbb{N}$, then on each step some past label is revealed to a predictor with some probability p^i_j.

(Simple) Poisson teacher. Let each l_i be distributed according to a Poisson distribution with parameter $\lambda > 0$, i.e. $L(l_i = n) = \frac{\lambda^n}{n!}e^{-\lambda}$ for all $i \in \mathbb{N}$. As before, we do not specify the distribution of random variables k_i, we just assume that they satisfy the definition. We call such learning rule a Simple Poisson teacher. Likewise the previous example, this case can be generalised to Poisson teacher by allowing different values of the parameter λ on different steps.

Analogously to the last example we can define Binomial teacher, Exponential teacher, etc.

Theorem 3.3. *Let $\Gamma_{1-\delta}$ be a symmetric (region) predictor, let $\mathfrak{L}$ be a randomised learning rule, and let $\Gamma^{\mathfrak{L}}_{1-\delta}$ be the $\mathfrak{L}$-taught version of $\Gamma_{1-\delta}$. The following statements hold for any probability distribution P^∞ generating the examples.*

- *If $\Gamma_{1-\delta}$ is a TCM and for each $i \in \mathbb{N}$ the second moment of l_i exists, and*
$$\sum_{i=2}^{\infty} \frac{\mathbb{E} l_i^2}{n_i^2} < \infty \ a.s. \tag{3.2}$$
then $\Gamma^{\mathfrak{L}}_{1-\delta}$ is well calibrated.

- *If for some $l > 0$, $\mathbb{E} l_i = l$ from some i on, the variances Δl_i of random variables l_i exist and*
$$\sum_{i=2}^{\infty} \frac{\Delta l_i}{n_i^2} < \infty \ a.s. \tag{3.3}$$
then $\overline{\mathrm{U}}(\Gamma^{\mathfrak{L}}_{1-\delta}, P) = \overline{\mathrm{U}}(\Gamma_{1-\delta}, P)$ and $\overline{\mathrm{Er}}(\Gamma^{\mathfrak{L}}_{1-\delta}, P) = \overline{\mathrm{Er}}(\Gamma_{1-\delta}, P)$.

It can be easily checked that Simple Bernoulli and Simple Poisson teachers both satisfy all conditions of Theorem 3.2, and thus the rates of errors and uncertain predictions are preserved in these cases.

3.3 Discussion of the conditions of the theorems

First we would like to give some explanations on the conditions of the first statement of the Theorem 3.2.

Remark 3.4. *$\Gamma^{\mathfrak{L}}_{1-\delta}$ is well calibrated when*

$$\frac{n_{k+1}}{n_k} = 1 + O\left(\frac{1}{\sqrt{k}\ln k}\right). \tag{3.4}$$

For example, it is well calibrated when n_k grows as $\exp(\sqrt{k}/\ln k)$; on the other hand, our result cannot guarantee that it is well calibrated if n_k grows as $\exp(\sqrt{k})$.

Proof. Indeed, condition (3.1) can be rewritten as

$$\sum_{i=2}^{\infty} \left(\ln \frac{n_{i+1}}{n_i} \right)^2 < \infty;$$

therefore, it is satisfied when $\ln(n_{k+1}/n_k) = O(1/(\sqrt{k} \ln k))$; this is equivalent to (3.4). □

Conditions of Theorem 3.3 concerning uncertain predictions (the second statement) are evidently weaker then those of Theorem 3.2; it seems that the strict condition of fixed l in Theorem 3.2 can be made weaker. However, the following simple example shows that it is not so.

Remark 3.5. *There exist a predictor Γ and a deterministic learning function $\mathfrak{L}$ for which*

$$\lim_{n\to\infty} \frac{1}{n} \operatorname{Er}_n(\Gamma^{\mathfrak{L}}) > \lim_{n\to\infty} \frac{1}{n} \operatorname{Er}_n(\Gamma)$$

although $\mathfrak{L}$ seems to be some deterministic version of Simple Bernoulli teacher with parameter $p = 1/3$.

Proof. Suppose that the object space consists of just one element: $X = \{x\}$, so objects play only formal role in this example. Let $Y = \{0, 1\}$, let $P(y_n = 0) = 0$ and $P(y_n = 1) = 1$ for all $n \in \mathbb{N}$. Thus, $z_n = 1$ for every $n \in \mathbb{N}$. However, Γ will sometimes make errors. For any $n \in \mathbb{N}$ we define $\Gamma(Z, x) = \{1\}$ for even n and $\Gamma(Z, x) = \{0\}$ for odd n, where $Z = (z_1, \ldots, z_{n-1})$. Observe that Γ always makes correct predictions on even steps and incorrect on odd: $\lim_{n\to\infty} \operatorname{Er}_n(\Gamma)/n = 1/2$.

Furthermore, we define $N = \{3k - 1, 3k : k \in \mathbb{N}\}$ and $\mathfrak{L}(n) = n$ for all $n \in N$. Clearly, $s(n) = [2/3n]$. It is easy to see that $\lim_{n\to\infty} \operatorname{Er}_n(\Gamma^{\mathfrak{L}})/n = 2/3 > 1/2$. □

3.4 Proofs

Proof of Theorem 3.2 Suppose that B is a region predictor (such as $\Gamma_{1-\delta}$ or $\Gamma^{\mathfrak{L}}_{1-\delta}$). We introduce the following "predictable" versions of $\mathrm{er}_n(B)$ and $\mathrm{un}_n(B)$:

$$\mathrm{err}_n(B) = P\Big\{(x,y) \in \mathbf{Z} : y \notin B_{1-\delta}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x)\Big\},$$

$$\mathrm{unc}_n(B) = P\Big\{(x,y) \in \mathbf{Z} : \big|B_{1-\delta}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x)\big| > 1\Big\},$$

$$\mathrm{Err}_n(B) = \sum_{i=1}^{n} \mathrm{err}_i(B),$$

$$\mathrm{Unc}_n(B) = \sum_{i=1}^{n} \mathrm{unc}_i(B)$$

Since $\mathrm{Er}_n(\Gamma^{\mathfrak{L}}_{1-\delta}) - \mathrm{Err}_n(\Gamma^{\mathfrak{L}}_{1-\delta})$ and $\mathrm{Un}_n(\Gamma^{\mathfrak{L}}_{1-\delta}) - \mathrm{Unc}_n(\Gamma^{\mathfrak{L}}_{1-\delta})$ are martingales, and

$$|\,\mathrm{er}_n(\Gamma^{\mathfrak{L}}_{1-\delta}) - \mathrm{err}_n(\Gamma^{\mathfrak{L}}_{1-\delta})| \leq 1,$$
$$|\,\mathrm{un}_n(\Gamma^{\mathfrak{L}}_{1-\delta}) - \mathrm{unc}_n(\Gamma^{\mathfrak{L}}_{1-\delta})| \leq 1,$$

the martingale strong law of large numbers (see, e.g., [42]) implies that

$$\lim_{n\to\infty} \frac{\mathrm{Er}_n(\Gamma^{\mathfrak{L}}_{1-\delta}) - \mathrm{Err}_n(\Gamma^{\mathfrak{L}}_{1-\delta})}{n} = 0 \text{ a.s.}$$

and

$$\lim_{n\to\infty} \frac{\mathrm{Un}_n(\Gamma^{\mathfrak{L}}_{1-\delta}) - \mathrm{Unc}_n(\Gamma^{\mathfrak{L}}_{1-\delta})}{n} = 0 \text{ a.s.};$$

this actually means that we can study $\mathrm{Err}_n(\Gamma^{\mathfrak{L}}_{1-\delta})$ and $\mathrm{Unc}_n(\Gamma^{\mathfrak{L}}_{1-\delta})$ instead of $\mathrm{Er}_n(\Gamma^{\mathfrak{L}}_{1-\delta})$ and $\mathrm{Un}_n(\Gamma^{\mathfrak{L}}_{1-\delta})$.

In this proof, where the arguments of functions $\Gamma_{1-\delta}$ and $\Gamma^{\mathfrak{L}}_{1-\delta}$ are not given explicitly, we assume that the predictor $\Gamma^{\mathfrak{L}}_{1-\delta}$ receives the sequence

$\bigl(z_i : i \in \mathbb{N}\bigr)$ while the predictor $\Gamma_{1-\delta}$ receives the sequence $\bigl(z_{k_i} : i \in \mathbb{N}\bigr)$. The latter sequence is distributed according to P^∞, since the choice of $\mathfrak{L}$, by definition, does not depend on the examples z_i, $i \in \mathbb{N}$.

We first prove the second statement of the theorem. For any $n \in \mathbb{N}$, by definition of $\Gamma^{\mathfrak{L}}_{1-\delta}$ and using the symmetry of $\Gamma_{1-\delta}$, we have

$$\begin{aligned}
&\mathrm{unc}_n(\Gamma^{\mathfrak{L}}_{1-\delta}) \\
&\quad = P\Bigl\{(x,y) \in \mathbf{Z} : |\Gamma^{\mathfrak{L}}_{1-\delta}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x)| > 1\Bigr\} \\
&\quad = P\Bigl\{(x,y) \in \mathbf{Z} : |\Gamma_{1-\delta}(x_{k_1}, y_{k_1}, \dots, x_{k_{s(n)}}, y_{k_{s(n)}}, x)| > 1\Bigr\} \\
&\qquad\qquad\qquad\qquad = \mathrm{unc}_{s(n)+1}(\Gamma_{1-\delta}).
\end{aligned}$$

Thus, since $s(n) = n/l + O(1)$, we have

$$\sum_{i=1}^{n} \mathrm{unc}_i(\Gamma^{\mathfrak{L}}_{1-\delta}) = l \sum_{i=1}^{\lfloor n/l \rfloor} \mathrm{unc}_i(\Gamma_{1-\delta}) + O(1),$$

and so

$$\mathrm{Unc}_n(\Gamma^{\mathfrak{L}}_{1-\delta}) = l\, \mathrm{Unc}_{\lfloor n/l \rfloor}(\Gamma_{1-\delta}) + o(n).$$

It follows that $\overline{\mathrm{U}}(\Gamma^{\mathfrak{L}}_{1-\delta}, P) = \overline{\mathrm{U}}(\Gamma_{1-\delta}, P)$.

The statement about $\overline{\mathrm{Er}}(\Gamma)$ is proven analogously.

Now we proceed with the first statement of the theorem. Clearly,

$$\mathrm{Err}_{n_k}(\Gamma^{\mathfrak{L}}_{1-\delta}) = l_1\, \mathrm{err}_1(\Gamma_{1-\delta}) + l_2\, \mathrm{err}_2(\Gamma_{1-\delta}) + \cdots + l_k\, \mathrm{err}_k(\Gamma_{1-\delta}) \qquad (3.5)$$

for any $k \in \mathbb{N}$. Denote

$$\overline{e}_1 := l_1\, \mathrm{err}_1(\Gamma_{1-\delta}),\ \overline{e}_2 := l_2\, \mathrm{err}_2(\Gamma_{1-\delta}),\ \dots$$

and

$$e_1 := l_1\, \mathrm{er}_1(\Gamma_{1-\delta}),\ e_2 := l_2\, \mathrm{er}_2(\Gamma_{1-\delta}),\ \dots .$$

It is easy to see that $\overline{e}_i - e_i$, $i \in \mathbb{N}$, is a martingale difference sequence with respect to $\Gamma_{1-\delta}$'s input sequence, z_{k_i}, $i \in \mathbb{N}$. Moreover,

$$\mathbb{E}\bigl((\overline{e}_i - e_i)^2 \mid z_{k_1}, \dots, z_{k_{i-1}}\bigr) \leq l_i^2,$$

for $i \in \mathbb{N}$. Thus,

$$\sum_{i=1}^{\infty} \frac{1}{n_i^2} \mathbb{E}\bigl((\overline{e}_i - e_i)^2 \mid z_{k_1}, \dots, z_{k_{i-1}}\bigr) \leq \sum_{i=1}^{\infty} \frac{l_i^2}{n_i^2} < \infty.$$

We can use (3.5) and the martingale strong law of large numbers to conclude that, as $k \to \infty$,

$$\frac{1}{n_k}\Bigl(\mathrm{Err}_{n_k}(\Gamma_{1-\delta}^{\mathfrak{L}}) - \sum_{i=1}^{k-2} e_i\Bigr) = \frac{1}{n_k}\sum_{i=1}^{k-2}(\overline{e}_i - e_i) \to 0 \text{ a.s.}$$

Analogously,

$$\frac{1}{n_k}\left(\sum_{i=1}^{k-2} e_i\right) - \delta = \frac{1}{n_k}\sum_{i=1}^{k-2}\bigl(e_i - \delta l_i\bigr) \to 0 \text{ a.s.}$$

And so

$$\frac{1}{n}\mathrm{Err}_n(\Gamma_{1-\delta}^{\mathfrak{L}}) \leq \frac{1}{n}\Bigl(\mathrm{Err}_{k_{s(n)}}(\Gamma_{1-\delta}^{\mathfrak{L}}) + l_{s(n)+1}\Bigr) \to \delta \text{ a.s.},$$

(the fact that $l_{s_{n+1}}/n \to 0$ follows from the first condition of the theorem) which implies the first statement of the theorem: $\frac{1}{n}\mathrm{Er}_n(\Gamma_{1-\delta}^{\mathfrak{L}}) \to \delta$ a.s.. □

Proof of Theorem 3.3 The proof of the first statement of the theorem is analogous to that of Theorem 3.2.

The proof of the second statement is also straightforward after that of Theorem 3.2, except for the ending: we have

$$\sum_{i=1}^{n} \mathrm{unc}_i(\Gamma_{1-\delta}^{\mathfrak{L}}) \leq l \sum_{i=1}^{\lfloor n/l \rfloor} \mathrm{unc}_i(\Gamma_{1-\delta}) + O(1) + \sum_{i=1}^{\lfloor n/l \rfloor} |l_i - \mathbb{E} l_i|. \tag{3.6}$$

We observe that $\sum_{i=1}^{\lfloor n/l \rfloor} |l_i - \mathbb{E}l_i|$ is a martingale, and use (3.3) and, again, martingale strong law of large numbers to derive

$$\frac{1}{n}\sum_{i=1}^{\lfloor n/l \rfloor} |l_i - \mathbb{E}l_i| \to 0 \text{ a.s.}$$

which, along with (3.6), implies $\overline{\mathrm{U}}(\Gamma_{1-\delta}^{\mathcal{L}}, P) = \overline{\mathrm{U}}(\Gamma_{1-\delta}, P)$. □

Chapter 4

Conclusion and future work

In the present work we have shown that methods developed in pattern recognition theory can be used in more general models than those they were designed for. That is, we have shown that certain assumptions traditionally imposed on the way examples are presented to learning algorithms can be relaxed without significant loss in performance.

The first assumption we have considered, the assumption that training and testing examples are distributed independently, was replaced with a weaker assumption of conditional independence. We have shown that for a wide range of predictors the results concerning their performance still hold true under the new assumptions. This concerns asymptotic as well as finite-step, distribution-free as well as distribution-specific results. Thus, it probably would not be an exaggeration to say that we have shown the independence assumption to be, to a considerable extend, redundant.

However there are still interesting problems to be solved on the way of relaxing the i.i.d. assumption. Thus, it appears important to find a way to generalise data-dependent estimates of performance (e.g. those known for SVMs, see [7]) to the new model. As mentioned before, it would also be interesting to obtain generalisations of our results to the case of non-deterministically defined labels. Some more strong consistency results are probably waiting to be obtained, as most of our results on non-parametric predictors concern weak consistency. The question of whether similar generalisations can be made to other learning tasks, such as regression estimation, is also may be worth an investigation.

The second assumption we have considered concerns the on-line learning scenario. We have introduced the possibility to omit or delay labels revealed to a predictor, and found conditions under which such "weakness" of a teaching protocol does not affect the performance of a predictor. These results were obtained in the framework of on-line region prediction, which is perhaps best suited for studying adaptive behaviour of learning algorithms. In this direction it appears interesting to obtain more results on the influence of weak teachers on the uncertainty of region predictors.

In sum, we can say that the traditionally used theoretical models for pattern recognition are too strict. However, there are negative results in pattern recognition theory that show that in a certain sense the theoretical models are too general. (For example, it is shown in [9] that for any pattern recognition task there exists a distribution under which the probability of error decreases arbitrarily slowly.)

Perhaps this controversy shows that the pattern recognition task is still awaiting its new theoretical models, and with them probably new practical methods.

Bibliography

[1] D. Aldous and U. Vazirani A Markovian extension of Valiant's learning model. In Proceedings of the 31st Symposium on Foundations of Computer Science, pp. 392–396, 1990.

[2] Y. Altun, I. Tsochantaridis, T. Hofmann. Hidden Markov Support Vector Machines. Proceedings of the Twentieth International Conference on Machine Learning, San Francisco: Morgan Kaufmann, 2003.

[3] P. Algoet, Universal Schemas for Learning the Best Nonlinear Predictor Given the Infinite Past and Side Information. IEEE Transactions on Information Theory, Vol. 45, No. 4, 1999.

[4] E. Baum and D. Haussler, What size net gives valid generalisation? Neural Computation, Vol. 1, pp. 151–160, 1989.

[5] A. Blumer, A. Ehrenfeucht, D. Haussler M and Warmuth Learnability and the Vapnik-Chervonenkis dimension. Journal of the ACM, 36, pp. 929–965, 1989.

[6] O. Bousquet, A. Elisseeff. Stability and Generalization. Journal of Machine Learning Research, 2, 499–526, 2002.

[7] N. Christianini, J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Methods. Cambridge University Press, Cambridge, 2000.

[8] T. Cover, P. Hart, Nearest neighbor pattern classification. IEEE Transactions on Information Theory, Vol. 13, pp. 21–27, 1967

[9] T. Cover. Rates if convergence for nearest neighbor procedures. In Proceedings of the Hawaii International Conference on Systems Sciences, pp. 413–415, Honolulu, 1968

[10] A.P. Dawid. Conditional Independence in Statistical Theory. Journal of the Royal Statistical Society, Series B (Methodological), Vol. 41, No 1, pp. 1–31, 1979

[11] A.P. Dawid, Conditional Independence, In Encyclopedia of Statistical Science (Update) Vol 3, Wiley, New York, 1999.

[12] L. Devroye, L. Györfi, G. Lugosi, A probabilistic theory of pattern recognition. New York: Springer, 1996.

[13] L. Devroye, On asymptotic probability of error in nonparametric discrimination. Annals of Statistics, Vol. 9, pp. 1320–1327.

[14] L. Devroye, L. Györfi, A. Krzyżak, G. Lugosi, On the strong universal consistency of nearest neighbor regression function estimates. Annals of Statistics, Vol. 22, pp. 1371–1385, 1994.

[15] L. Devroye, T. Wagner. Distribution-free inequalities for the deleted and holdout error estimates. IEEE Transactions on Information Theory, Vol. 25, No 2, pp. 202–207, 1979.

[16] L. Devroye, T. Wagner. Distribution-free performance bounds for potential function rules. IEEE Transactions on Information Theory, Vol. 25(5), pp. 601–604, 1979.

[17] R. Duda, P. Hart, D. Stork. Pattern Classification, Second edition, Wiley-Interscience, 2001.

[18] E. Fix, J. Hodges. Discriminatory analysis. Nonparametric discrimination: Consistency properties. Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX, 1951.

[19] E. Fix, J. Hodges. Discriminatory analysis: small sample performance Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX, 1952.

[20] D. Gamarnik, Extension of the PAC framework to finite and countable Markov chains IEEE Transactions on Information Theory, Vol. 49, No. 1, pp. 338–345, 2003.

[21] L. Gordon, R. Olshen, Asymptotically efficient solutions to the classification problem. Annals of Statistics, Vol. 6, pp. 515–533, 1978

[22] L. Gordon, R. Olshen, Consistent nonparametric regression from recursive partitioning schemes. Journal of Multivariate Analysis, Vol. 10 pp. 611–627, 1980.

[23] M. Kearns and D. Ron, Algorithmic stability and sanity-check bounds on leave-one-out cross-validation. Neural Computation, Vol. 11, No. 6, pp. 1427–1453, 1999.

[24] M. Kearns M. and U. Vazirani. An Introduction to Computational Learning Theory The MIT Press, Cambridge, Massachusetts, 1994.

[25] S. Kulkarni, S. Posner, S. Sandilya. Data-Dependent k_n-NN and Kernel Estimators Consistent for Arbitrary Processes. IEEE Transactions on Information Theory, Vol. 48, No. 10, pp.2785–2788, 2002.

[26] S. Kulkarni, S. Posner. Rates of Convergence of Nearest Neighbour Estimation Under Arbitrary Sampling. IEEE Transactions on Information Theory, Vol. 41, No. 10, 1995, pp.1028–1039.

[27] J. Lafferty, A. McCallum, F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282–289. San Francisco: Morgan Kaufmann, 2001.

[28] G. Lugosi, A. Nobel, Consistency of data-driven histogram methods for density estimation and classification. Annals of Statistics vol. 24, No.2, pp.687–706, 1996.

[29] G. Lugosi, K. Zeger, Nonparametric Estimation via empirical risk minimization. IEEE Transactions on Information Theory, Vol. 41, No. 3, pp. 677–687, 1995.

[30] A. McCallum, D, Freitag, F. Pereira. Maximum entropy Markov models for information extraction and segmentation. Proceedings of the Seventeenth International Conference on Machine Learning, pp. 591–598. San Francisco: Morgan Kaufmann, 2000.

[31] G. Morvai, S. Yakowitz, P. Algoet, Weakly Convergent Nonparametric Forecasting of Stationary Time Series. IEEE Transactions on Information Theory, Vol. 43, No. 2, 1997

[32] G. Morvai, S. Kulkarni, and A.B. Nobel, Regression estimation from an individual stable sequence, Statistics, Vol. 33, pp.99–118, 1999.

[33] B. Natarajan. Machine Learning: a Theoretical Approach. Morgan Kaufmann, San Mateo, CA, 1991.

[34] A.B. Nobel, Limits to classification and regression estimation from ergodic process, Annals of Statistics, Vol. 27, pp. 262–273, 1999.

[35] A.B. Nobel, G. Morvai, and S. Kulkarni, Density estimation from an individual numerical sequence, IEEE Transactions on Information Theory, vol. 44, pp.537–541, 1998.

[36] I. Nouretdinov, V. Vovk. Criterion of calibration for Transductive confidence machine with limited feedback. Proceedings of the 14th International Conference on Algorithmic Learning Theory (ed. by Ricard Gavaldá, Klaus P. Jantke and Eiji Takimoto). Lecture notes in Artificial Intelligence, vol. 2842. Springer, pp.268–282, 2003

[37] B. Novikoff, On convergence proofs on perceptrons, in Proceedings of the Symposium on the Mathematical Theory of Automata, Vol XII, Polytechnic Institute of Brooklin, pp. 615–622, 1962.

[38] Y. Pao, Adaptive pattern recognition and neural networks. Addison-Wesley , 1989

[39] L. Rabin, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, in Proceedings of the IEEE, Vol. 77 No. 2, 1989.

[40] W. Rogers and T. Wagner. A finite sample distribution-free performance bound for local discrimination rules. Annals of Statistics, Vol. 6, No. 3, pp.506–514, 1978.

[41] B. Ryabko, Prediction of random sequences and universal coding. Problems of Information Transmission, Vol. 24, pp. 87–96, 1988.

[42] A. Shiryaev, Probability, second edition. New York: Springer, 1996.

[43] C. Stone, Consistent nonparametric regression. Annals of Statistics, Vol. 14, pp. 1348–1360, 1977

[44] L. Valiant, A theory of the learnable. Communications of the ACM, Vol. 27, pp.1134–1142. 1984

[45] V. Vapnik, Statistical Learning Theory: New York etc.: John Wiley & Sons, Inc. 1998

[46] V. Vapnik, and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its applications, Vol. 16, pp. 264–280.

[47] V. Vapnik and A. Chervonenkis. Ordered risk minimisation I. Automation and Remote Control, Vol. 35 pp.1226–1235, 1974

[48] V. Vapnik and A. Chervonenkis Ordered risk minimisation II. Automation and Remote Control, Vol. 35, pp. 1403–1412, 1974

[49] V. Vapnik, and A. Chervonenkis. Theory of Pattern Recognition. Nauka, Moscow, 1974 (in Russian); German translation: Theorie der Zeichenerkennung, Akademie Verlag, Berlin 1979.

[50] M. Vidyasagar, A theory of learning and generalization with applications to neural networks and control systems. Springer, Berlin, 1997

[51] V. Vovk. On-line Confidence Machines are well-calibrated. Proceedings of the Forty Third Annual Symposium on Foundations of Computer Science:187–196, IEEE Computer Society, 2002.

[52] V. Vovk. Asymptotic optimality of Transductive Confidence Machine. Proceedings of the Thirteenth International Conference on Algorithmic Learning Theory, 2002.

[53] V. Vovk, A. Gammerman and C. Saunders. Machine-learning applications of algorithmic randomness. Proceedings of the Sixteenth International Conference on Machine Learning, pp.444–453. San Francisco, CA: Morgan Kaufmann, 1999.

[54] V. Vovk, A. Gammerman, G. Shafer. Algorithmic Learning in a Random World, Springer, 2004.