Online Learning of Conditionally I.I.D. Data

Daniil Ryabko DANIIL@CS.RHUL.AC.UK
Computer Learning Research Centre, Royal Holloway, University of London, Egham Hill, Egham, Surrey, TW20 0EX, UK

Abstract

In this work we consider the task of relaxing the i.i.d assumption in online pattern recognition (or classification), aiming to make existing learning algorithms applicable to a wider range of tasks. Online pattern recognition is predicting a sequence of labels based on objects given for each label and on examples (pairs of objects and labels) learned so far. Traditionally, this task is considered under the assumption that examples are independent and identically distributed. However, it turns out that many results of pattern recognition theory carry over under a much weaker assumption. Namely, under the assumption of conditional independence and identical distribution of objects only, while the only condition on the distribution of labels is that the rate of occurrence of each label should be above some positive threshold.

We find a broad class of learning algorithms for which estimations of the probability of a classification error achieved under the classical i.i.d. assumption can be generalised to the similar estimates for the case of conditionally i.i.d. distributed examples.

1. Introduction

Online pattern recognition (or classification) is, informally, the following task. There is a finite number of classes of some complex objects. A predictor is learning to label objects according to the class they belong to (i.e. to classify), based only on some examples (labelled objects). One of the typical practical examples is recognition of a hand-written text. In this case, an object is a hand-written letter and a label is the letter of an alphabet it denotes. Another example is recognising some illness in a patient. An object here is the set of symptoms of a patient, and the classes are those of normal and ill.

The formal model of the task used most widely is described, for example, in Vapnik (1998), and can be briefly introduced as follows (we will later refer to it as "the i.i.d. model"). 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 $\eta(x)$[1]. 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" $\eta(x)$.

Initially, the problem of pattern recognition had been considered in so-called *offline* (or *batch*) 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 first set and then is used to classify the objects from the second; the less errors it makes on the testing set the better. In another setting, so-called *online setting* of pattern recognition problem, 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 correct second label is given, and so on. This setting is more natural than the offline one in the practical applications dealing with constantly changing or slowly gained data (see e.g. Bottou and LeCun (2003) for a study in which cases online methods outperform offline methods).

There is a plenty of algorithms developed for solving the pattern recognition task (see Devroye, Györfi and Lugosi (1996) for the most widely used methods). However, the i.i.d assumption, which is central in the model, is too tight for many applications.

It turns out that it is also too tight for a wide range of methods developed under the assumptions of the

Appearing in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004. Copyright 2004 by the author.

[1] Often (e.g. in Vapnik (1998)) a more general situation is considered, the labels are drawn according to some probability distribution $P(y|x)$, i.e. each object can have more than one possible label.

model: they work nearly as well under weaker assumptions.

Consider the following situation. 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"). Does it mean that we can not use pattern recognition methods developed within the i.i.d. model for the text recognition task? No, we can shuffle the letters of the text and then use those methods. But will the results of recognition change significantly if we do not shuffle the letters? It is intuitively clear that the answer is negative; it is intuitively clear if we are having in mind nearly any popular pattern recognition method. Moreover, in online tasks we cannot shuffle examples, and so the question is not idle.

It turns out that the only needed assumption on the distribution of examples are the following two. First, that the dependence between *objects* is only that between their *labels*; in other words, the type of object-label dependence does not change in time. In our example, an image of a letter which in the beginning of the text denotes, say, "a", to the end of the text will not be interpreted as, say, "e". Second, each label should not cease in occurrence, i.e. 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 $\mathbf{P}(y)$, where $\mathbf{P}$ is a 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 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.

The main criterion in estimating how well a predictor works is the probabilities of its errors. In this work we provide a tool for obtaining estimations of probability of an error of a predictor in the conditional model from an estimation of the probability of an 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 change significantly the probability of an 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 (which are of the main concern in this work) as well as asymptotic results.

The general theorems about extending results concerning performance of a predictor to the conditional model are illustrated with the example of predictors minimising empirical error. We use some results of Vapnik-Chervonenkis theory to establish tolerance to data of such predictors and show what results about them the developed theory yields.

The idea of relaxing the i.i.d assumption in the i.i.d. model is not new. Thus, in Morvai, Yakowitz and Algoet (1997) the authors study the task of predicting a stationary and ergodic sequence objects, and also consider the generalisation of this task to the task of regression estimation (which itself is a more general variant of pattern recognition). Under the assumption that the joint distribution of objects and labels is stationary and ergodic, the authors manage to construct a weakly consistent predictor. This is a reasonable result in the framework and for the (more general) task at hand, but it is not the kind of results which is usually an achievable goal for pattern recognition methods, namely results on the performance of a predictor over a finite sample of data, and pointwise (almost surely) consistent predictors. Another approach is considered in Vovk (2002), where the authors construct a wide class of predictors for the case of exchangeable examples (i.e. the distribution generating examples is exchangeable).In both approaches the authors consider different types of assumptions on the *joint* distribution of objects and labels. Then they construct a predictor, as in Morvai, Yakowitz and Algoet (1997), or a class of predictors, as in Vovk (2002), to work well under the assumptions made. Our approach is different in that we find the conditions on the distribution of labels and (another condition) on conditional distribution of objects, under which a certain class of predictors which are *already known* to work well in the i.i.d. model, work as well.

2. Main Results

The traditional scenario for online pattern recognition is as follows.

Consider a sequence of *examples* $(x_1, y_1), (x_2, y_2), \ldots$; 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.

The traditional assumption about the distribution $\mathbf{P}$ generating objects is that $\mathbf{P} = P^\infty$ for some distribution P on $\mathbf{X}$, i.e. examples are i.i.d. This is what we call in this paper *the i.i.d. model.*

Here we replace this assumption with the following two conditions.

First, for any $n \in \mathbb{N}$

$$\mathbf{P}(x_n \in A \mid \mathcal{U}) = \mathbf{P}(x_n \in A \mid \mathcal{U}_n), \qquad (1)$$

where A is any measurable set (an event) in $\mathbf{X}$, $\mathcal{U}_n$ is σ-algebra generated by y_n and $\mathcal{U}$ is any σ-algebra which contains $\mathcal{U}_n$.

In more intuitive notation, for any $i_1, \ldots, i_k, j_1, \ldots, j_k \in \mathbb{N}$

$$\begin{aligned} &\mathbf{P}(x_n \in A \mid y_n, x_{i_1}, y_{j_1}, \ldots, x_{i_k}, y_{j_k}) \\ &\qquad\qquad = \mathbf{P}(x_n \in A \mid y_n). \end{aligned}$$

(This condition looks very much like 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 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 event A in $\mathbf{X}$

$$\mathbf{P}(x_{n_1} \in A \mid y_{n_1} = y) = \mathbf{P}(x_{n_2} \in A \mid y_{n_2} = y). \qquad (2)$$

(It is worth noting that (1) allows dependence in n, otherwise the present condition is not needed.)

For each $y \in \mathbf{Y}$ and any $n \in \mathbb{N}$ we will denote the distribution $\mathbf{P}(x_n \mid y_n = y)$ by $\mathbf{P}_y$ (it does not depend on n by (2)). As we want the function $\eta(x)$ which specifies the label for each object to be deterministic, we put an extra requirement on the distributions $\mathbf{P}_y$, $y \in \mathbf{Y}$, namely that there exist such sets $X_y \subset \mathbf{X}$, $y \in \mathbf{Y}$ such that $X_0 \cap X_1 = \varnothing$ and $\mathbf{P}_y(X_y) = 1$ for each $y \in \mathbf{Y}$.[2]

Under the conditions (1) and (2) we say that *objects are conditionally independent and identically distributed* (conditionally i.i.d).

Less formally, these conditions can be reformulated as follows. Assume that we have some sequence $(y_n)_{n \in \mathbb{N}}$ of labels and two probability distributions P_0 and P_1 on $\mathbf{X}$. Each example $x_n \in \mathbf{X}$ is drawn according to the distribution P_{y_n}; examples are drawn independently of each other.

A *predictor* is a measurable function $\Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n)$ taking values in $\mathbf{Y}$. Denote $\Gamma_n := \Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n)$.

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

$$\begin{aligned} &\mathrm{err}_n(\Gamma, \mathbf{P}, z_1, \ldots, z_{n-1}) \\ &\qquad := \mathbf{P}\bigl\{(x, y) \in \mathbf{Z} \bigm| y \neq \Gamma_n(z_1, \ldots, z_{n-1})\bigr\} \end{aligned}$$

(Here $\mathbf{P}$ in the list of arguments of err_n is understood as a distribution conditional on $z_1, \ldots, z_{n-1}$.)

We will often use a shorthand notation

$$\mathbf{P}(\mathrm{err}_n(\Gamma, z_1, \ldots, z_{n-1}) > \varepsilon)$$

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

$$\mathbf{P}\bigl\{z_1, \ldots, z_{n-1} \bigm| \mathrm{err}_n(\Gamma, \mathbf{P}, z_1, \ldots, z_{n-1}) > \varepsilon\bigr\}.$$

We call a predictor Γ *(finitely) universally consistent with the bounding function* $\bigtriangledown : \mathbb{N} \times \mathbb{R} \to [0, 1]$ if for any distribution P on $\mathbf{Z}$

$$P^\infty(\mathrm{err}_n(\Gamma) > \varepsilon) \leq \bigtriangledown(n, \varepsilon). \qquad (3)$$

We say that a predictor Γ is *tolerant to data with bounding function* $\Delta : \mathbb{N} \times \mathbb{R} \to [0, 1]$ if for any distribution P on $\mathbf{Z}$

$$\begin{aligned} P^\infty\Bigl(&\max_{j \leq \varkappa_n;\ \pi:\{1,\ldots,n\} \to \{1,\ldots,n\}} \bigl|\mathrm{err}_{n+1}(\Gamma, z_1, \ldots, z_n) - \\ &\mathrm{err}_{n-j+1}(\Gamma, z_{\pi(1)}, \ldots, z_{\pi(n-j)})\bigr| > \varepsilon\Bigr) \leq \Delta(n, \varepsilon), \quad (4) \end{aligned}$$

[2] Without this requirement, the conditions (1) and (2) would model a more general setting of the pattern recognition problem, in which each object has more than one possible label.

for any $n \in \mathbb{N}$, any $\varepsilon > 0$ and $\varkappa_n := \sqrt{n \log n}$ (see the end of the Section 5 for the discussion of the choice of the constants $\varkappa_n$). The probability in this definition is taken over $z_1, \dots, z_n$.

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

Theorem 1. *Suppose that a distribution* $\mathbf{P}$ *is such that the objects are conditionally i.i.d, i.e.* $\mathbf{P}$ *satisfies (1) and (2). Fix some* $\delta \in (0, 1/2]$, *denote* $p(n) := \frac{1}{n}\#\{i \le n : y_i = 0\}$ *and* $C_n := \mathbf{P}(\delta \le |p(n)| \le 1 - \delta)$ *for each* $n \in \mathbb{N}$. *For any predictor* Γ *if* Γ *is finitely universally consistent with some bounding function* $\bigtriangledown(n, \varepsilon)$ *and universally tolerant to data with some bounding function* $\Delta(n, \varepsilon)$, *then*

$$\mathbf{P}(\mathrm{err}_n(\Gamma) > \varepsilon) \le 2C_n^{-1}\Big(\bigtriangledown(n, \delta\varepsilon/3) \\ +2\Delta(n + \varkappa_n/2, \delta\varepsilon/6)\Big) + (1 - C_n^{-1}).$$

for any $\varepsilon > 0$ *and any* $n > e^{4\delta^{-2}}$.

The proof of this and the following theorem can be found in Appendix A.

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.

A predictor developed to work in the offline setting should be, loosely speaking, tolerant to permutations of the training sample. The theorem shows under which conditions in the online model this property of a predictor can be utilised.

Theorem 1 provides a tool for obtaining *distribution–free* bounds on probability of an error in the conditional model given the bounds in the i.i.d. model. However, often for certain classes of distributions there exist much better bounds on the probability of an error than for the universal (distribution-free) case. Next we show how *distribution–specific* results achieved in the i.i.d. model can be extended to the conditional model.

Let $\mathbf{P}$ be some distribution on $\mathbf{X}^\infty$ satisfying (1) and (2). We say that a distribution P on $\mathbf{X}$ *agrees* with $\mathbf{P}$ if the conditional distribution $P(x|y)$ is equal to $\mathbf{P}_y$ and $P(y) \neq 0$ for each $y \in \mathbf{Y}$. Clearly, this defines the distribution P up to the parameter $p = P(y = 1) \in (0, 1)$. For a distribution $\mathbf{P}$ on $\mathbf{X}^\infty$ we denote the family of distributions which agree with $\mathbf{P}$ by $(P_p)_{p \in (0,1)}$ where $P_p(y = 1) = p$ for each p in $(0, 1)$.

For a distribution $\mathbf{P}$ on $\mathbf{X}^\infty$ which satisfies (1) and (2) we call a predictor Γ *(finitely) consistent for the distribution* $\mathbf{P}$ *with the bounding function* $\bigtriangledown : \mathbb{N} \times \mathbb{R} \to [0, 1]$ if (3) holds for any distribution P on $\mathbf{X}$ which agrees with $\mathbf{P}$. Furthermore, we say that a predictor Γ is *tolerant to data for a distribution* $\mathbf{P}$ *with bounding function* $\Delta : \mathbb{N} \times \mathbb{R} \to [0, 1]$ if (4) holds for any distribution P on $\mathbf{X}$ which agrees with $\mathbf{P}$.

Theorem 2. *Suppose that a distribution* $\mathbf{P}$ *on* $\mathbf{X}^\infty$ *satisfies (1) and (2). Fix some* $\delta \in (0, 1/2]$, *denote* $p(n) := \frac{1}{n}\#\{i \le n : y_i = 0\}$ *and* $C_n := \mathbf{P}(\delta \le |p(n)| \le 1 - \delta)$ *for any* $n \in \mathbb{N}$. *For any predictor* Γ *if* Γ *is finitely consistent for* $\mathbf{P}$ *with some bounding function* $\bigtriangledown(n, \varepsilon)$ *and tolerant to data for* $\mathbf{P}$ *with some bounding function* $\Delta(n, \varepsilon)$, *then*

$$\mathbf{P}(\mathrm{err}_n(\Gamma) > \varepsilon) \le 2C_n^{-1}\Big(\bigtriangledown(n, \delta\varepsilon/3) \\ +2\Delta(n + \varkappa_n/2, \delta\varepsilon/6)\Big) + (1 - C_n^{-1}).$$

for any $\varepsilon > 0$, *any* $n > e^{4\delta^{-2}}$.

Let us call a class of distributions $\mathcal{P}$ on $\mathbf{X}$ *conditionally closed* if with any distribution $P \in \mathcal{P}$ the class $\mathcal{P}$ also includes any distribution P' such that $P(A|y = i) = P'(A|y = i)$ for any $A \subset \mathbf{X}$ and each $i \in \mathbf{Y}$ (i.e. P'^∞ agrees with P^∞).

As important examples of conditionally closed classes of distributions we mention the class of distributions which have densities, which have smooth densities, distributions which generate examples separable by a hyperplane.

Theorem 2 means that if we have some bounds on the error probabilities of a predictor Γ for some conditionally closed class of distributions $\mathcal{P}$ then we can obtain bounds on the error probabilities of Γ for any distribution $\mathbf{P}$ on $\mathbf{X}$ such that any (some) distribution on $\mathbf{X}$ which agrees with $\mathbf{P}$ is in $\mathcal{P}$ (having bounds on tolerance of Γ to data for the distributions from $\mathcal{P}$). In fact, Theorem 1 is an immediate consequence of Theorem 2.

3. Application to PAC Learning Theory

Here we show how the developed concepts relate to the PAC (Probably Approximately Correct) theory (see, e.g. Vidyasagar (1997); Kearns and Vazirani (1994)); here we mainly follow Vidyasagar (1997) in definitions).

For the purpose of this section we fix some conditionally closed class $\mathcal{P}$ of distributions on $\mathbf{Z}$.

A predictor Γ is called *PAC for the class of distributions* $\mathcal{P}$ if

$$\lim_{n\to\infty}\sup_{P\in\mathcal{P}} P^\infty(\mathrm{err}_n(\Gamma) > \varepsilon) = 0$$

for each $\varepsilon > 0$.

Denote by $\overline{\mathcal{P}}$ the set of distributions on $\mathbf{Z}^\infty$ which satisfy (1) and (2), such that for each $\mathbf{P} \in \overline{\mathcal{P}}$ there exist $P \in \mathcal{P}$ such that P agrees with $\mathbf{P}$. We call a predictor Γ *PAC in conditional model for the class of distributions* $\overline{\mathcal{P}}$ if

$$\lim_{n\to\infty}\sup_{\mathbf{P}\in\overline{\mathcal{P}}} \mathbf{P}(\mathrm{err}_n(\Gamma) > \varepsilon) = 0$$

for each $\varepsilon > 0$. For each $\delta \in (0, 1/2]$, denote $p(n) := \frac{1}{n}\#\{i \le n : y_i = 0\}$ and $\mathbf{C_n}(\delta) := \sup_{\mathbf{P}\in\overline{\mathcal{P}}} \mathbf{P}(\delta \le |\mathbf{p(n)}| \le \mathbf{1} - \delta)$ for each $n \in \mathbb{N}$.

Theorem 2 implies the following statement.

Corollary 1. *Suppose that* $\lim_{n\to\infty} \mathbf{C_n}(\delta) \to \mathbf{1}$ *for some* $\delta \in (0, 1/2]$. *Suppose further, that a predictor* Γ *is tolerant data for each* $\mathbf{P} \in \overline{\mathcal{P}}$ *with some bounding function* $\Delta(n, \varepsilon)$, *such that* $\lim_{n\to\infty} \Delta(n, \varepsilon) = 0$ *for each* $\varepsilon > 0$. *Then if* Γ *is PAC for* $\mathcal{P}$ *then it is PAC in conditional model for* $\overline{\mathcal{P}}$.

4. Structural Risk Minimisation and Applications to Popular Predictors

In this section we use some results of Vapnik-Chervonenkis theory to establish tolerance to data of certain popular classes of predictors and show how the asymptotic results (strong universal consistency) can be obtained within the conditional model.

The concepts of Vapnik-Chervonenkis theory used here were developed in Vapnik and Chervonenkis (1974a; 1974b; 1974c). See also Vapnik (1998) and Devroye, Györfi and Lugosi (1996) for detailed overviews. Here we mainly follow Devroye, Györfi and Lugosi (1996) in notations.

Let $\mathbf{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$ and let $\mathcal{C}$ be a class functions of the form $\varphi : \mathbf{X} \to \mathbf{Y} = \{0, 1\}$, called *decision functions*. For a probability distribution P on $\mathbf{X}$ we denote $\mathrm{err}(P, \varphi) := P(\varphi(x_i) \ne y_i)$. The symbol $\mathcal{S}(\mathcal{C}, n)$ denotes the n-th shatter coefficient of the class $\mathcal{C}$. For a sample of examples $(z_1, \dots, z_n)$ and a decision function $\varphi \in \mathcal{C}$ the empirical error functional $\overline{\mathrm{err}}_n(\varphi)$ is defined as $\overline{\mathrm{err}}_n(\varphi) := \sum_{i=1}^n I_{\varphi(x_i)\ne y_i}$, where, as usual, $z_i = (x_i, y_i))$.

Theorem 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* $(z_1, \dots, z_n)$. *Then* Γ *is universally tolerant to data with the bounding function*

$$\Delta(n, \varepsilon) = 16\mathcal{S}(\mathcal{C}, n)e^{-n(\varepsilon - 4\varkappa_n/n)^2/128}. \qquad (5)$$

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

For example, for bounds on the VC dimension of classes of neural networks see e.g. Baum and Haussler (1989) (also in Devroye, Györfi and Lugosi (1996), Theorem 30.6).

Next we show that the asymptotic performance of an empirical risk minimising predictor in the conditional model can also be estimated with the help of the theorems of the previous section.

Lemma 1. *Let* $\mathbf{P}$ *be some distribution on* $\mathbf{X}^\infty$ *satisfying (1) and (2). Assume that a sequence of classes* $\mathcal{C}^{(k)}$ *of decision rules of the form* $\mathbf{X} \to \mathbf{Y}$ *is such that* $\lim_{k\to\infty}\inf_{\varphi\in\mathcal{C}^{(k)}} \mathrm{err}(P, \varphi) = 0$ *for any distribution* P *which agrees with* $\mathbf{P}$. *Then*

$$\lim_{k\to\infty}\sup_{p\in[0,1]}\inf_{\varphi\in\mathcal{C}^{(k)}} \mathrm{err}(P_p, \varphi) = 0.$$

Corollary 2. *Let* Γ *be a classifier that minimises the empirical error over the class* $\mathcal{C}^{(k)}$, *where* $\mathcal{C}^{(k)}$ *is the class of neural net classifiers with* k *nodes in the hidden layer and the threshold sigmoid, and* $k \to \infty$ *so that* $k \log n/n \to 0$ *as* $n \to \infty$. *Let* $\mathbf{P}$ *be any distribution on* $\mathbf{X}^\infty$ *satisfying (1) and (2) such that* $\sum_{n=1}^\infty (1 - C_n^{-1}) < \infty$. *Then* Γ *is strongly consistent for* $\mathbf{P}$, *i.e.*

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

5. Discussion

In the section 2 we have introduced “conditionally i.i.d.” model for pattern recognition which generalises the commonly used i.i.d. model. A general tool is presented which makes it possible to extend the results achieved in the i.i.d. model to the conditional one.

The first question which arises is how much more general the conditional model is, and how useful is the generalisation. In response to the first part, observe that in the i.i.d. model, labels should be i.i.d, while in the conditional model labels can be distributed arbitrary, with the only restriction that the rate of occurrence of each label does not tend zero. To compare, in the i.i.d model the rate of occurrence of each label quickly tends to a certain limit. The assumption that objects are i.i.d conditionally on labels seems to capture the idea that this is only the object-label dependence that a predictor is required to learn, which itself does not put any restrictions on the dependence between the examples.

Another question is in what cases the bounds on the probability of an error provided by the (general) theorems 1 and 2 are of the same quality as those in the conditional model. Which means, are the bounds on tolerance to data of the same (or lower) order then the bounds on the probability of an error in the i.i.d. model? To show that this is often the case, we consider empirical error minimising predictors in section 4.

In the theory of structural error minimisation the probability of an error is split into two parts: estimation error and approximation error (see, e.g. Devroye, Györfi and Lugosi (1996)). The *estimation error* is the difference between the probability of an error of the considered predictor and that of the minimum of probability of an error among all decision rules in the class. The bounds on this probability of an error are universal and, in general, good, subject to the VC dimension of the class of decision rules. The second part, the *approximation error* is the minimum of probability of an error among all decision rules in the class. This variable is, in general, greater than the estimation error and can tend to zero arbitrarily slow. We show that for predictors minimising empirical risk the constants bounding the tolerance to data are of the same order that the estimation error, i.e. small, subject to the VC dimension of the class.

Still another question remains, can the same bounds on the probability of an error in the conditional model be achieved without assumptions on tolerance to data? The following negative example shows that the bounds on tolerance to data are necessary.

Proposition 1. *There exists a distribution* $\mathbf{P}$ *on* $\mathbf{X}^\infty$ *satisfying (1) and (2) such that* $\mathbf{P}(|p_n - 1/2| > 3/n) = 0$ *for any* n *(i.e.* $C_n = 1$ *for any* $\delta > 0$ *and* $n > 3$*) and a predictor* Γ *such that* $P^\infty(\mathrm{err}_n > 0) \le 2^{1-n}$ *for any distribution* P *which agrees with* $\mathbf{P}$ *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 $y \in \mathbf{Y}$. The distribution $\mathbf{P}$ is defined as follows: $\mathbf{P}_y = P_y$ for each $y \in \mathbf{Y}$ and $\mathbf{P}|_{\mathbf{Y}^\infty}$ is the Markov distribution with transition probability matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, i.e. it always generates sequences of labels $\ldots 01010101 \ldots$.

We define the predictor Γ as follows

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

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. □

One more point which needs clarification is the choice of the constants $\varkappa_n$. We fixed these constants for the sake of simplicity of notations, however, they can be made variable, as long as $\varkappa_n$ obeys the following condition.

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

almost surely for any $p \in (0, 1)$ and any probability distribution P on $\mathbf{X}$ such that $P(y = 1) = p$, where $p_n := \frac{1}{n}\#\{i \le n : y_i = 0\}$. Increasing $\varkappa_n$ increases the Δ function and decreases C_n^{-1}

Acknowledgements

The author would like to thank the reviewers for useful comments and constructive criticism. I also thank Alex Gammerman and Vladimir Vovk for fruitful discussions. I am grateful to Royal Holloway, University of London for funding my Ph.D. study. This research was partially supported by EPSRC (grant GR/R46670/01), BBSRC (grant 111/BIO14428) and Royal Society (grant ref: 15995).

Appendix A: proofs for Section 2

Theorem 1 is an immediate consequence of Theorem 2, so we proceed with the proof of the latter.

Proof of Theorem 2. We define the conditional probabilities of errors of Γ as follows

$$\begin{aligned} \mathrm{err}_n^0(\Gamma) := \mathbf{P}(&y_n \ne \Gamma_n \mid \\ &y_n = 0; x_1, y_1, \ldots, x_{n-1}, y_{n-1}), \\ \mathrm{err}_n^1(\Gamma) := \mathbf{P}(&y_n \ne \Gamma_n \mid \\ &y_n = 1; x_1, y_1, \ldots, x_{n-1}, y_{n-1}), \end{aligned}$$

(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 $y_n = y$ and given (random variables) $x_1, y_1, \ldots, x_{n-1}, y_{n-1}$. We will also use more explicit notations for $\mathrm{err}_n^y(\Gamma)$ specifying the distribution or the input sequence of labels, when the context requires. Obviously, $\mathrm{err}_n(\Gamma) \le \max_{y\in\mathbf{Y}} \mathrm{err}_n^y(\Gamma)$.

For any $\mathbf{y} := (y_1, y_2, \dots) \in \mathbf{Y}^\infty$, denote $\mathbf{y}_n := (y_1, \dots, y_n)$ and $p_n(\mathbf{y}) := \#\{i \le n : y_i = 0\}$.

Fix some $n > 1$, some $y \in \mathbf{Y}$ and such $\mathbf{y}^1 \in \mathbf{Y}^\infty$ that $n\delta \le p_n(\mathbf{y}^1) \le n(1-\delta)$. We shell find bounds on $\mathbf{P}\big(\mathrm{err}_n > \varepsilon \mid \mathbf{y}_n = \mathbf{y}_n^1\big)$. The following fact will allow us to pass from i.i.d. distributions to conditionally i.i.d. Observe that

$$\begin{aligned}&\mathbf{P}\big(\mathrm{err}_n^y(\Gamma) > \varepsilon \mid \mathbf{y}_n = \mathbf{y}_n^1\big)\\ &\qquad = P_p^\infty\big(\mathrm{err}_n(\Gamma) > \varepsilon \mid \mathbf{y}_n = \mathbf{y}_n^1\big)\end{aligned}$$

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

It is also clear that if $\mathrm{err}_n(\Gamma) < \varepsilon$ then $\mathrm{err}_n^y(\Gamma) < \varepsilon/\delta$ for each $y \in \mathbf{Y}$, if underlying probability distribution is P_p, $\delta \le p \le 1-\delta$. Denote $p = p_n(\mathbf{y}^1)/n$. For any $\mathbf{y}^2 \in \mathbf{Y}^\infty$ such that $|p_n(\mathbf{y}^2) - np| \le \delta\varkappa_n/2$ there exist such permutations π_1, π_2 of the set $\{1, \dots, n\}$ that $y^1_{\pi_1(i)} = y^2_{\pi_2(i)}$ for any $i \le n - \delta\varkappa_n$. Hence (denoting $n' := n - \delta\varkappa_n$) we have,

$$\begin{aligned}
&P_p^\infty\big(\mathrm{err}_n^y(x_1, y_1^1, \dots, x_n, y_n^1) > \varepsilon\big)\\
&\qquad = P_p^\infty\Big(\mathrm{err}_n^y(x_1, y_1^1, \dots, x_n, y_n^1)\\
&\qquad - \mathrm{err}_{n'}^y(x_{\pi_1(1)}, y^1_{\pi_1(1)}, \dots, x_{\pi_1(n')}, y^1_{\pi_1(n')})\\
&\qquad + \mathrm{err}_{n'}^y(x_{\pi_1(1)}, y^2_{\pi_2(1)}, \dots, x_{\pi_1(n')}, y^2_{\pi_2(n')})\\
&\qquad - \mathrm{err}_n^y(x'_1, y_1^2, \dots, x'_n, y_n^2)\\
&\qquad + \mathrm{err}_n^y(x'_1, y_1^2, \dots, x'_n, y_n^2) > \varepsilon\Big)\\
&\le P_p^\infty\Big(\big|\,\mathrm{err}_n^y(x_1, y_1^1, \dots, x_n, y_n^1)\\
&- \mathrm{err}_{n'}^y(x_{\pi_1(1)}, y^1_{\pi_1(1)}, \dots, x_{\pi_1(n')}, y^1_{\pi_1(n')})\big| > \varepsilon/3\Big)\\
&+P_p^\infty\Big(\big|\,\mathrm{err}_{n'}^y(x'_{\pi_2(1)}, y^2_{\pi_2(1)}, \dots, x'_{\pi_2(n')}, y^2_{\pi_2(n')})\\
&\qquad - \mathrm{err}_n^y(x'_1, y_1^2, \dots, x'_n, y_n^2)\big| > \varepsilon/3\Big)\\
&+P_p^\infty\Big(\mathrm{err}_n^y(x'_1, y_1^2, \dots, x'_n, y_n^2) > \varepsilon/3\Big),
\end{aligned}$$

where $x'_{\pi_2(i)} = x_{\pi_1(i)}$ for $i \le n'$ and so the probability in the first line is taken over the space $\mathbf{X}^n$, in the second line over $\mathbf{X}^{n+\delta\varkappa_n}$ and everywhere rest over $\mathbf{X}^n$.

To bound the first two terms, we observe that for any $\mathbf{y}^c \in \mathbf{Y}^\infty$ such that $|p_m(\mathbf{y}_n^c) - pm| \le \varkappa_n/4$, where $m = n + \varkappa_n/2$, there exist permutations ψ_1, ψ_2 of the set $\{1, \dots, n+\varkappa_n/2\}$ such that $y^c_{\psi_j(i)} = y_i^j$ for $j = 1,2$ and $i \le n$. Hence, for $j = 1,2$ and any permutation π of the set $\{1, \dots, n\}$ we have

$$\begin{aligned}
&P_p^\infty\Big(\big|\,\mathrm{err}_n^y(x_1, y_1^j, \dots, x_n, y_n^j)\\
&- \mathrm{err}_{n'}^y(x_{\pi(1)}, y^j_{\pi(1)}, \dots, x_{\pi(n')}, y^j_{\pi(n')})\big| > \varepsilon/3\Big)
\end{aligned}$$

$$\begin{aligned}
&\le P_p^\infty\Big(\big|\,\mathrm{err}_m^y(x_1, y_1^c, \dots, x_m, y_m^c)\\
&- \mathrm{err}_n^y(x_{\psi_j(1)}, y^c_{\psi_j(1)}, \dots, x_{\psi_j(n)}, y^c_{\psi_j(n)}\big| > \varepsilon/6\Big)\\
&+P_p^\infty\Big(\big|\,\mathrm{err}_m^y(x_1, y_1^c, \dots, x_m, y_m^c)\\
&\qquad - \mathrm{err}_{n'}^y(x_{\pi(\psi_j(1))}, y^c_{\pi(\psi_j(1))}, \dots,\\
&\qquad\qquad x_{\pi_j(\psi_j(n')}, y^c_{\pi_j(\psi_j(n'))})\big| > \varepsilon/6\Big)
\end{aligned}$$

$$\begin{aligned}
&\le 2P_p^\infty\Big(\max_{i\le\varkappa_n, \pi:\{1,\dots,m\}\to\{1,\dots,m\}} \big|\,\mathrm{err}_m^y(z_1, \dots, z_m)\\
&\qquad - \mathrm{err}_{m-i}^y(z_{\pi(1)}, \dots, z_{\pi(m-i)})\big| > \varepsilon/6 \;\Big|\\
&\qquad\qquad |p_m(\mathbf{y}) - p(m)| \le \varkappa_n/2\Big)\\
&\qquad \le \frac{2}{P_p^\infty(|p_m(\mathbf{y}) - pm| \le \varkappa_n/2)}\,\Delta(m, \delta\varepsilon/6)\\
&\qquad\qquad \le 4\Delta(m, \delta\varepsilon/6)
\end{aligned}$$

if $n > 8$.

Thus,

$$\begin{aligned}
&P_p^\infty\big(\mathrm{err}_n^y(x_1, y_1^1, \dots, x_n, y_n^1) > \varepsilon\big)\\
&\qquad \le P_p^\infty\big(\mathrm{err}_n^y(x_1, y_1^2, \dots, x_n, y_n^2) > \varepsilon/3\big)\\
&\qquad\qquad +4\Delta(m, \delta\varepsilon/6),
\end{aligned}$$

and, hence $\mathbf{y}^2$ was chosen arbitrary among sequences $\mathbf{y} \in \mathbf{Y}^\infty$ for which $|p_n(\mathbf{y}) - np| \le \delta\varkappa_n/2$, we conclude

$$\begin{aligned}
&P_p^\infty\big(\mathrm{err}_n^y(x_1, y_1^1, \dots, x_n, y_n^1) > \varepsilon\big)\\
&\le P_p^\infty\big(\mathrm{err}_n^y > \varepsilon/3 \,\big|\, |p_n(\mathbf{y}) - np| \le \delta\varkappa_n/2\big)\\
&\qquad +4\Delta(m, \delta\varepsilon/6) \le\\
&\quad 2P_p^\infty\big(\mathrm{err}_n^y > \varepsilon/3\big) + 4\Delta(m, \delta\varepsilon/6)\\
&\quad \le 2\triangledown(n, \delta\varepsilon/3) + 4\Delta(n + \varkappa_n/2, \delta\varepsilon/6)
\end{aligned}$$

(here we used that $n > e^{16\delta^{-2}}$). Finally, as $\mathbf{y}^1$ was chosen arbitrary among sequences $\mathbf{y} \in \mathbf{Y}^\infty$ such that $n\delta \le p_n(\mathbf{y}^1) \le n(1-\delta)$ we have

$$\begin{aligned}
&\mathbf{P}(\mathrm{err}_n > \varepsilon) \le \mathbf{P}(\max_{y\in\mathbf{Y}} \mathrm{err}_n^y > \varepsilon) \le 2C_n^{-1}\Big(\triangledown(n, \delta\varepsilon/3)\\
&\qquad +2\Delta(n + \varkappa_n/2, \delta\varepsilon/6)\Big) + (1 - C_n^{-1}).
\end{aligned}$$

which concludes the proof. ∎

Appendix B: proofs for Section 4

Proof of Theorem 3. Fix some probability distribution P on $\mathbf{X}$ and some $n \in \mathbb{N}$. Denote $\varphi_n^* := \arg\min_{\varphi\in\mathcal{C}} \overline{\mathrm{err}}_n(\varphi)$ (so that $\Gamma_n = \varphi_n^*$). We also denote by $\varphi^\times$ any such decision rule $\varphi \in \mathcal{C}$ that

$$\overline{\mathrm{err}}_n(\varphi) = \max_{j\le\varkappa_n;\ \pi:\{1,\dots,n\}\to\{1,\dots,n\}} \min_{\varphi\in\mathcal{C}} \overline{\mathrm{err}}_{n-j}\big(\varphi, z_{\pi(1)},\dots,z_{\pi(n-j)}\big)$$

We need to show that $P^n(|\operatorname{err}(\varphi^*)-\operatorname{err}(\varphi^\times)|>\varepsilon) \le \Delta(n,\varepsilon)$.

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

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

Now the statement of the theorem follows from the fact that

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

see Devroye, Györfi and Lugosi (1996), Theorem 12.6. ∎

Proof of Corollary 2. Applying Theorem 2 and using its notations we have

$$\nabla(n,\varepsilon) \le P_p(\mathrm{err}_n - \inf_{\varphi\in\mathcal{C}^{(k)}}\operatorname{err}(P_p,\varphi)>\varepsilon/2) + I_{\inf_{\varphi\in\mathcal{C}^{(k)}}\operatorname{err}(P_p,\varphi)>\varepsilon/2}.$$

By Theorem 12.6, Devroye, Györfi and Lugosi (1996) the first term is bounded by

$$16\mathcal{S}(\mathcal{C}^{(k)},n)e^{-n\varepsilon^2/128}$$

and is summable if we use the bound

$$\mathcal{S}(\mathcal{C}(k),n) \le (n\varepsilon)^{kd+2k+1}, \qquad (6)$$

see Theorem 30.6, Devroye, Györfi and Lugosi (1996). The second term is bounded by

$$I_{\sup_{p\in[0,1]}\inf_{\varphi\in\mathcal{C}^{(k)}}\operatorname{err}(P_p,\varphi)>\varepsilon/2}$$

and so is summable by Corollary 30.1, Devroye, Györfi and Lugosi (1996), which says that

$$\lim_{k\to\infty}\inf_{\varphi\in\mathcal{C}^{(k)}}\operatorname{err}(P,\varphi)=0$$

for any distribution P on $\mathbf{X}$ and from Lemma 1.

For the function $\Delta(n,\varepsilon)$ we have the bound provided by Theorem 3, which is also summable if we use the bound (6). ∎

References

Baum, E. and Haussler, D. (1989). *What size net gives valid generalisation?* Neural Computation, 1:151-160.

Bottou L., LeCun Y. (2003). *Large Scale Online Learning.* Advances in Neural Information Processing Systems 16 (*proceedings of NIPS 2003*)

Devroye L., Györfi G., Lugosi G (1996). *A probabilistic theory of pattern recognition.* New York: Springer.

Kearns M.J. and Vazirani U.V. (1994) *An Introduction to Computational Learning Theory* The MIT Press, Cambridge, Massachusetts.

Morvai G., Yakowitz S. J., Algoet P. (1997). *Weakly Convergent Nonparametric Forecasting of Stationary Time Series* IEEE Transactions on Information Theory, Vol. 43, No. 2, pp. 483–498.

Vapnik V. N. (1998), *Statistical Learning Theory* New York etc.: John Wiley & Sons, Inc.

Vapnik, V. and Chervonenkis, A. (1974) *Ordered risk minimisation I.* Automation and Remote Control, 35: 1226-1235.

Vapnik, V. and Chervonenkis, A. (1974). *Ordered risk minimisation II.* Automation and Remote Control, 35: 1403-1412.

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

Vidyasagar M. (1997) *A Theory of Learning and Generalization* New York: Springer.

Vovk V.(2002). On-line Confidence Machines are well-calibrated *Proceedings of the Forty Third Annual Symposium on Foundations of Computer Science, pp. 187–196* IEEE Computer Society.