



Gaining degrees of freedom in subsymbolic learning

B. Apolloni^{a,*}, D. Malchiodi^b

^a *Department of Computer Science, University of Milan, Via Comelico 39-41, 20135 Milan, Italy*

^b *Department of Mathematics, University of Milan, Via C. Saldini 50, 20133 Milan, Italy*

Received April 1998; revised May 1999

Communicated by M. Nivat

Abstract

We provide some theoretical results on sample complexity of PAC learning when the hypotheses are given by subsymbolical devices such as neural networks. In this framework we give new foundations to the notion of degrees of freedom of a statistic and relate it to the complexity of a concept class. Thus, for a given concept class and a given sample size, we discuss the efficiency of subsymbolical learning algorithms in terms of degrees of freedom of the computed statistic. In this setting we appraise the sample complexity overhead coming from relying on approximate hypotheses and display an increase in the degrees of freedom yield by embedding available formal knowledge into the algorithm. For known sample distribution, these quantities are related to the learning approximation goal and a special *production prize* is shown. Finally, we prove that testing the approximation capability of a neural network generally demands smaller sample size than training it. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Computational learning; Sentry functions; Nested concept classes; Approximate learning; Neural networks

1. Introduction

Drawing a sample of 10 000 items is generally wasteful if we want to estimate the parameter p of a Bernoulli distribution law from the observation of the values assumed by the random variable. In fact, already with sample size 2000 the probability of drawing a sample whose mean value is more than 0.05 far from p is less than 0.05. On the contrary, the same size is generally too small if we want to estimate the location and shape of a convex polygon of 20 sides separating two random variables that differ by a given property on the real plane. In this case, with the same sample size no any

* Corresponding author. Fax: +39-2-5500-6276
E-mail address: apolloni@dsi.unimi.it (B. Apolloni).

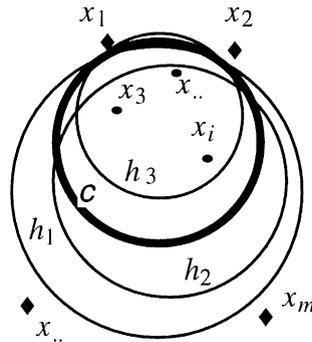


Fig. 1. Hypotheses h_i on circle c that are consistent with sampled points $\{x_j\}$.

estimator can guarantee sampling probability >0 to compute a polygon separating the two properties with a mistake probability less than 0.05 for each distribution law of the random variables.

The difference between the two inference instances can be suitably characterised through the notion of degrees of freedom. This parameter is generally used in a pragmatic way to instantiate some distribution law such as Chi-square, Student and so on [14].

In this paper we shed a new light on the notion of degrees of freedom in the frame of PAC learning theory and employ the management of this parameter as a meaningful tool for penetrating some theoretical aspects of subsymbolical learning [24].

In a previous paper [5] we characterised within a sample set drawn to learn a boolean function special elements constituting the pivots of the related statistic. In greater detail, let us consider the following paradigmatic situation. Some one told me that a polluting load has been injected in the groundwater at some point of a given region. Due to the radial symmetry of the site geomorphology, we can assume at a given time that the polluted region is a circle c of unknown centre and diameter as in Fig. 1. Concerned about the health of the inhabitants of this region, I order a set of samplings of the groundwater in some points. These points are selected within the region according to the same distribution law that describes the probability of meeting an inhabitant in any place of the region. Some water samples have shown to be affected and others immune from the pollutants. On the basis of these answers I will draw a circle, like any h_i in Fig. 1, that constitutes my hypothesis about the polluted region. In the PAC learning terminology computing this statistics realises the learning of the concept c within the concept class \mathbf{C} of circles through a hypothesis h within the hypothesis class \mathbf{H} coinciding with \mathbf{C} .

In greater detail, under the unique consistency constraint that no sampled unpolluted point falls within my hypothesis and no polluted point outside it, I am free to draw any circle within a large family where each element of the complementary family of forbidden circles contains at least one faulty point. Given our hypothesis h , the faulty region is the symmetric difference $c \div h$ between c and h . We are not interested in

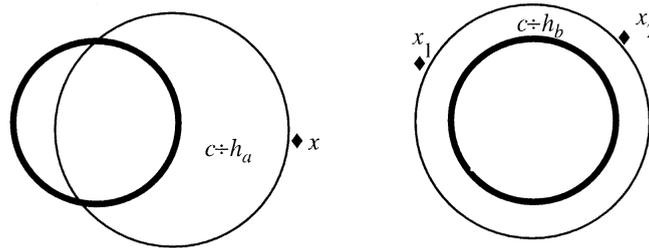


Fig. 2. Contrasting symmetric differences between circles.

comparing it with any faulty region coming from a different arbitrary selection of h , but only with the regions which contain our selection. In fact, our question is how does the sample set intervene on the algorithm computing h to bind its symmetric difference w.r.t. c ?

To disregard the arbitrary part of the algorithm, let us consider the maximality set constituted by consistent worst-case hypotheses characterised by the following statement: for each hypothesis h in the set there is no other consistent circle h' whose symmetric difference with c includes $c \div h$. As evident in Fig. 2 for the maximality set of circles vs. circles, only one sampled point is sufficient for this class of hypotheses to *contrast* the enlargement of $c \div h_a$ and 2 points for $c \div h_b$.

The number of necessary points changes with the class of the unknown boolean functions we want to discover and with the class of hypotheses we can compute. For instance, this number is at most 2 if the symmetric differences are segments, at most $\lfloor \frac{4}{3}k \rfloor$ if they are k -edges convex polygons, at most 2 if they are circles (see points x_1 and x_2 in Fig. 1), infinite if they are sets of unlimited number of sets, and so on.

Two distinguishing features are evident:

1. Because of the consistency constraint all sample points will always be outside the symmetric difference $c \div h$.
2. Whatever the worst-case hypothesis we can compute, some points of the sample must be employed to contrast the enlargement of its symmetric difference w.r.t. c . The other points can instead stay everywhere.

The relevance of the contrasting points comes from the following:

Consider a set \mathbf{B} of growing domains fully ordered by the inclusion relation, and assume that $c \div h \cup \{\text{contrasting points}\}$ belongs to this set and the previous element in the order is included in $c \div h$. Then, for each domain in the set, the contrasting points are the witnesses of the inclusion or not of $c \div h$ in this domain, and consequently of an analogous inequality relation on the probabilities of the two domains.

This is the starting point for our results. We will identify with $m-v$ the degrees of freedom of an m sized sample with respect to hypothesis h , where v is the number of contrasting points. We focus on these points within the sample and consider the event A that a domain of probability measure α in the set \mathbf{B} includes $c \div h$. We show that for any distribution on the sample space the probability of drawing a sample – and

then a hypothesis consistent with it – such that A occurs is minored by the incomplete Beta function [23] $I_x(\mu, m - \mu + 1)$, where μ is a tight upperbound on the number of contrasting points. The statistic feature of the contrasting points comes from the fact that I_x is a decreasing function on μ , and μ is, obviously, a non-decreasing function on v .

In this perspective contrasting points might be viewed as an extension of multivariate rank order statistics [25, 26]. Instead, they play a role quite different from the support vectors recently introduced by Vapnik [29] as a minimal set of points just delimiting the variation range of the hypotheses consistent with them.

Our paper deepens the role of contrasting points in learning algorithms for sub-symbolic devices such as neural network. In particular, the paper will concern general inconveniences affecting favourable circumstances that enhance the degrees of freedom of a sample, realising that:

- A sample complexity overhead intervenes when we rely on approximate classes of hypotheses [19]. On the contrary, v might be decreased by points contrasting irrelevant increments of the probability measure of $c \div h$, when we know the sample space distribution law.
- An increase in the degrees of freedom comes from embedding into the learning algorithm formal knowledge available about the goal concept [1].
- Specialising μ as a function of an upperbound ε on $p(c \div h)$, we have that for some probability distribution $\mu(1/\varepsilon)$ is a decreasing function, thus realising a sort of *production prize*: the better you learn the more you are rewarded by a bonus containing an additional amount of degrees of freedom,
- A further increment in the degrees of freedom occurs when passing from training to testing a learnt hypothesis – an usual procedure when working with subsymbolical devices such as neural networks [24]. Thus, testing the approximation capability of a neural network generally demands smaller sample size than training it.

This paper is organised as follows. In Section 2 we formalise the notion of detail of a concept class, we revisit in our framework the notion of degrees of freedom of a sample in respect to a given statistic, and state some elementary properties of these parameters. In Section 3 we give our learnability results on boolean functions, both in general and in relation to peculiarities of subsymbolic learning devices. Section 4 focuses on the inherent differences between training and testing procedures in regard to the sample size demand. Theoretical statements on boolean functions are extended to real functions for some learning strategies. Conclusions and outlooks of future work are reported in Section 5.

2. Detail of classes of boolean functions

Facing an approximating hypothesis h from \mathbf{H} for a concept c belonging to \mathbf{C} , we distinguish within a sample of observations of c the contrasting points. They act as sentinels that forbid the expansion of the symmetric difference $c \div h$ toward some other

$c \div h'$. The remaining sampled points constitute the rear-guard which, if numerous and fairly scattered on \mathbf{X} , give confidence that each sentinel has been considered. Linking sentinels, probability and sample size allows us to bind the sample complexity of learning. Let us start with the problem of sentinelling concepts within any class \mathbf{C} . Then we focus on classes $\{c \div h\}$ of symmetric differences between a goal concept from \mathbf{C} and candidate approximating concepts within \mathbf{H} .

Definition 1. Given a set \mathbf{X} , a class of concepts \mathbf{C} is a set of boolean functions c on \mathbf{X} . These functions are also called concepts c . By abuse of notation, we shall not make distinction between c and its support, i.e. the set of points x such that $c(x_i) = 1$. Therefore we also view \mathbf{C} as a set of subsets of \mathbf{X} .

Definition 2. Given a concept class \mathbf{C} on \mathbf{X} , an *outer sentry* function on \mathbf{C} is a total function $\mathbf{S}: \mathbf{C} \cup \{\emptyset, \mathbf{X}\} \mapsto 2^{\mathbf{X}}$ satisfying the following conditions:¹

- (1) The elements of $\mathbf{S}(c)$ are outside c , i.e. $c \cap \mathbf{S}(c) = \emptyset$.
- (2) Let us denote $c^+ = c \cup \mathbf{S}(c)$ and $up(c) = \{(c' \in \mathbf{C} \mid c' \not\subseteq c \text{ and } c^+ \subseteq c'^+)\}$,

if $c_2 \in up(c_1)$ then $c_2 \cap \mathbf{S}(c_1) \neq \emptyset$.

- (3) No $\mathbf{S}' \neq \mathbf{S}$ exists satisfying (1) and (2) and having the property that

$\mathbf{S}'(c) \subseteq \mathbf{S}(c)$ for every c .

- (4) Whenever c_1 and c_2 are such that $c_1 \subset c_2 \cup \mathbf{S}(c_2)$ and $c_2 \cap \mathbf{S}(c_1) = \emptyset$, then the restriction of \mathbf{S} to $c_1 \cup up(c_1) - \{c_2\}$ is a sentry function on this set.

Terminology $\mathbf{S}(c)$ is the *outer frontier* upon \mathbf{S} of c , their elements are called sentry points. Concept c_2 is *sentinelled* by $\mathbf{S}(c_1)$ iff $c_2 \cap \mathbf{S}(c_1) \neq \emptyset$.

Remark 1. A frontier does not fully identify the warded concept. A given concept class might admit more than one (possibly infinite) outer sentry functions obeying the conciseness constraint (3). Condition (4) prevents us from building sentry functions which are *unnatural*, where some sentry points of c_1 have the sole role of artificially increasing the elements of c_1^+ in order to prevent it from being included in another concept c_2 . The mentioned condition states that this role can be considered only as a side effect of points which are primarily involved in sentinelling some formula of $up(c_1)$. Note that the condition is stated in a rather redundant form, since $c_2 \notin up(c_1)$ by definition. The very essential statement is: \mathbf{S} is a sentry function to $up(c)$ for every c .

Example 1. A possible frontier of an item c of the concept class \mathbf{C} of circle c consists of points x_1, x_2 shown in Fig. 1.

¹ The reader should recognize this definition as an improved formulation of definition 3 in [5], cleaned of a noisy misprint in condition (2).

Example 2. Let us consider the class \mathbf{B}_2 of boolean forms on $\{0, 1\}^2$:

$$\mathbf{C} = \{0, 1, x_1, x_2, x_1x_2\}.$$

The related supports are:

	00 01 10 11
$c_1 = 0$	– – – –
$c_2 = x_1x_2$	– – – +
$c_3 = x_1$	– – + +
$c_4 = x_2$	– + – +
$c_5 = 1$	+ + + +

A possible outer sentry function for \mathbf{C} is:

$$\mathbf{S}(c_1) = \{11\}, \quad \mathbf{S}(c_2) = \{01, 10\}, \quad \mathbf{S}(c_3) = \{01\}, \quad \mathbf{S}(c_4) = \{10\}, \quad \mathbf{S}(c_5) = \emptyset.$$

Note that $\mathbf{S}(c_2) = \{00, 10\}$ should violate statement (4), since remotion of c_4 would make point 00 useless.

Definition 3. We call *outer detail* \mathbf{D}_C of a concept class \mathbf{C} the supremum of the cardinalities of the frontiers of its concepts with respect to all possible sentry functions. In symbols, $\mathbf{D}_C = \sup_{\mathbf{S}, c} \#\mathbf{S}(c)$

Fact 1. The class \mathbf{C} of convex polygons of k edges (k -gons) has detail $\mathbf{D}_C = \lfloor \frac{4}{3}k \rfloor$.

Proof. Let us consider any convex polygon B and assume w.l.o.g. that edges do not belong to B . For any edge (a, b) and any external point x and any straight line \mathfrak{l} say that x and \mathfrak{l} are on the same side w.r.t. (a, b) if there exists a tangent \mathfrak{t} to B passing through either a and/or b which leaves x and \mathfrak{l} on a same half-plane and (a, b) on the opposite one (see Fig. 3). Then for any such (a, b) , x and \mathfrak{l} , the half-plane α delimited by \mathfrak{l} and containing B is not sentinelled by x if and only if \mathfrak{l} passes through points $y' \in (a, x)$ and $y'' \in (b, x)$. Now, because of convexity of B , for any edge (a, b) we can always pick two such points y', y'' – and therefore an α not sentinelled by x – if and only if either x does not belong to (a, b) or x coincides with a or b . In the last case the opposite vertex can sentinel α .

Thus, to avoid that a polygon B' includes B eluding $\mathbf{S}(B)$ we can put one sentinel on each edge or, alternatively a sentinel on each vertex. Both frontiers are minimal, since possible B edges belong to the above considered straight lines. Moreover, we conclude by inspection that a mix of these results is still minimal and effective if at least a couple of consecutive edges bringing a sentinel inside alternate with edges bringing sentinels on the vertices (Fig. 4). \square

In a similar way, we can define the complementary notion of inner sentries as follows:

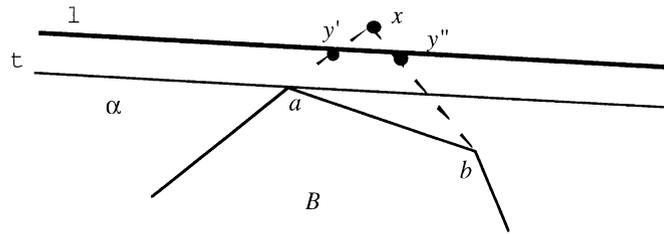


Fig. 3. Sentinelling by x the half-plane delimited by 1.

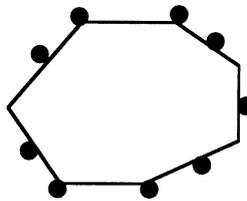


Fig. 4. Sentinelling a concept within the class of convex 7-gons.

Definition 2' Given a concept class \mathbf{C} on \mathbf{X} , an *inner sentry* function on \mathbf{C} is a total function $\mathbf{s}: \mathbf{C} \cup \{\emptyset, \mathbf{X}\} \mapsto 2^{\mathbf{X}}$ satisfying the following conditions:

- (1') The elements of $\mathbf{s}(c)$ are inside c , i.e. $c \cap \mathbf{s}(c) = \mathbf{s}(c)$.
- (2') Let us denote $c^- = c - \mathbf{s}(c)$ and $dw(c) = \{c' \in \mathbf{C} \mid c \not\subseteq c' \text{ and } c'^- \subseteq c^-\}$, if $c_2 \in dw(c_1)$ then $\overline{c_2} \cap \mathbf{s}(c_1) \neq \emptyset$.
- (3') No $\mathbf{s}' \neq \mathbf{s}$ exists satisfying (1) and (2) and having the property that $\mathbf{s}'(c) \subseteq \mathbf{s}(c)$ for each c .
- (4') Whenever c_1 and c_2 are such that $c_2 - \mathbf{s}(c_2) \subset c_1$ and $\overline{c_2} \cap \mathbf{s}(c_1) = \emptyset$, then the restriction of \mathbf{s} to $c_1 \cup dw(c_1) - \{c_2\}$ is a sentry function on this set.

Terminology c_2 is *sentinelled* by $\mathbf{s}(c_1)$ iff $\overline{c_2} \cap \mathbf{s}(c_1) \neq \emptyset$.

Definition 3' The *inner detail* $d_{\mathbf{C}}$ of a concept class \mathbf{C} is defined by

$$d_{\mathbf{C}} = \sup_{s,c} \#\mathbf{s}(c)$$

Fact 1' The class \mathbf{C} of convex k -gons has detail $d_{\mathbf{C}} = \lfloor \frac{4}{3}k \rfloor$.

Definition 4 (Vapnik [28]). Given a concept class \mathbf{C} and a finite set $Q \subseteq \mathbf{X}$, let $\Pi_{\mathbf{C}}(Q)$ denote the set of all subsets of Q that can be obtained by intersecting Q with a concept in \mathbf{C} , i.e. $\Pi_{\mathbf{C}}(Q) = \{Q \cap c \mid c \in \mathbf{C}\}$. The *Vapnik–Chervonenkis dimension* of \mathbf{C} (shortly, $d_{VC}(\mathbf{C})$) is the last integer d such that $\max_{Q \mid \#Q=d} \#\Pi_{\mathbf{C}}(Q) = 2^d$; if no such d exists, then $d_{VC}(\mathbf{C})$ is assumed to be infinite. If $\#\Pi_{\mathbf{C}}(Q) = 2^{\#Q}$, then we say that Q is *shattered* by \mathbf{C} [11].

Theorem 1. For any concept class \mathbf{C} , sentry functions \mathbf{S} and \mathbf{s} and concept $c \in \mathbf{C}$, the sets $\mathbf{S}(c)$ and $\mathbf{s}(c)$ are both shattered by $\mathbf{C} \cup \{\emptyset, \mathbf{X}\}$. Thus $\mathbf{D}_{\mathbf{C}} \leq d_{\text{VC}}(\mathbf{C}) + 1$, and $\mathbf{d}_{\mathbf{C}} \leq d_{\text{VC}}(\mathbf{C}) + 1$.

Proof. The statement on \mathbf{S} is proved in [5]. For its trivial extension to \mathbf{s} see [4]. \square

Fact 2. Further considerations [5] allow us to bind the detail also from below. Precisely

(i) $(d_{\text{VC}}(\mathbf{C}) - 1)/176 < \mathbf{d}_{\mathbf{C}}$, (ii) $(d_{\text{VC}}(\mathbf{C}) - 1)/176 < \mathbf{D}_{\mathbf{C}}$.

Remark 2. The class \mathbf{C} of the convex k -gons has Vapnik–Chervonenkis dimension $d_{\text{VC}}(\mathbf{C}) = 2k + 1$ [12, 31].

Details are a complexity measure of concept classes dual to the VC dimension. The former use points to separate sets of concepts, the latter concepts for partitioning sets of points. In this sense details should result a more constructive measure to which a partial algebra can be defined. More precisely, it is easy to prove by simple counting arguments that:

Fact 3. Given a class of concepts \mathbf{C} , for any enlargement $\mathbf{C}' \supseteq \mathbf{C}$:

- (1) For any pair of outer and inner sentry function \mathbf{S}, \mathbf{s} on \mathbf{C} , a pair of sentry function \mathbf{S}', \mathbf{s}' exists on \mathbf{C}' such that for each $c \in \mathbf{C}$ $\mathbf{S}(c) \subseteq \mathbf{S}'(c)$ and $\mathbf{s}(c) \subseteq \mathbf{s}'(c)$.
- (2) $\mathbf{D}_{\mathbf{C}'} > \mathbf{D}_{\mathbf{C}}$ and $\mathbf{d}_{\mathbf{C}'} > \mathbf{d}_{\mathbf{C}}$.

Proof. Point (2) is a straight consequence of point (1). The first statement is derived from the fact that $up(c)$ inside \mathbf{C} is included in the same set built inside \mathbf{C}' . \square

Definition 5. Let us denote by a $\rho\langle \mathbf{b} \rangle$ any relation ρ singularly satisfied by a and each element b_i of the set \mathbf{b} . For given concept classes \mathbf{C} and \mathbf{H} on \mathbf{X} , \mathbf{H} is dense w.r.t. \mathbf{C} if for each $h \in \mathbf{H}$ and for each $c \in \mathbf{C}$, a h' and h'' exist in \mathbf{H} such that:

- (i) $(c \cup h) \subseteq h'$ and $h'' \subseteq (c \cap h)$,
- (ii) for each pair h_1, h_2 such that $\langle (c \cup h_1), (c \cup h_2) \rangle \in up(c \cup h)$ and $(h_1 - h - c) \cup (h_2 - h - c) \not\subseteq \langle h_1, h_2 \rangle$, $(h_1 - h - c) \cup (h_2 - h - c) \not\subseteq \langle h'_1, h'_2 \rangle$.
- (iii) for each pair h_1, h_2 such that $\langle (c \cap h_1), (c \cap h_2) \rangle \in dw(c \cap h)$ and $(h - h_1 - \bar{c}) \cup (h - h_2 - \bar{c}) \not\subseteq \langle h - h_1, h - h_2 \rangle$, $(h - h_1 - \bar{c}) \cup (h - h_2 - \bar{c}) \not\subseteq \langle h - h''_1, h - h''_2 \rangle$.

A point x belonging to $(h_1 - h - c) - h'_2$ as in point (ii) or satisfying the analogous condition concerning point (iii) is said to be in the exclusive frontier of h w.r.t. h_1 against h_2 .

Example 3. The class \mathbf{H} of convex k -gons in a plane is dense on any convex figure – see Fig. 5a for density property exploitation of 4-gons w.r.t. ellipses in regard of the part concerning union of concepts. The same is not always true if the convexity constraint is removed – see Fig. 5b for \mathbf{H} constituted by difference of convex exagons with respect to ellipses).

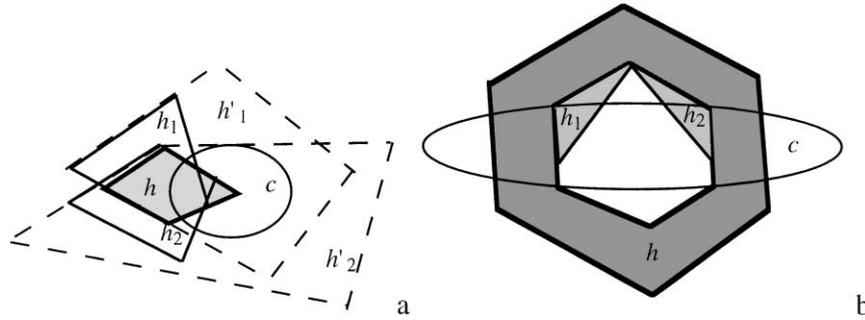


Fig. 5. Denseness properties for convex (a) and concave (b) figures. (a) for any pair h_1, h_2 such that their unions with c includes $(c \cup h), h'_1, h'_2$ are available that allow sentinelling separately both $(c \cup h_1), (c \cup h_2)$ within $S(c \cup h)$ and h'_1, h'_2 within $S(h)$. (b) the same does not happen for h concave, where h_1 is obtained by adding the left triangle and h_2 the right triangle to h .

The class \mathbf{H} of circles in a plane is dense on any convex figure.

Density failures may occur on convex polygons as well, when they are framed in a limited plane region.

Fact 4. For given concept classes \mathbf{C} and \mathbf{H} dense w.r.t. \mathbf{C} ,

- let us denote by $\mathbf{S}_\mathbf{H}$ an outer sentry function for \mathbf{H} . For a given $c \in \mathbf{C}$, let \mathbf{D}_c be the class of concepts $d = c \cup h$ for every $h \in \mathbf{H}$ and $\mathbf{D}_\mathbf{D} = \sup_{\mathbf{S}, c} \mathbf{D}_{\mathbf{D}_c}$, then:
 - (i) for every outer sentry function \mathbf{S} on \mathbf{D}_c there exists a $\mathbf{S}_\mathbf{H}$ such that $\mathbf{S}(d) \subseteq \mathbf{S}_\mathbf{H}(h)$ for each h, d s.t. $d = c \cup h$,
 - (ii) $\mathbf{D}_\mathbf{D} \leq \mathbf{D}_\mathbf{H}$.
- let us denote by $\mathbf{s}_\mathbf{H}$ an inner sentry function for \mathbf{H} and by \mathbf{Q}_c the class of concepts $q = c \cap h$ for a given c and every $h \in \mathbf{H}$. Let $\mathbf{d}_\mathbf{Q} = \sup_{\mathbf{S}, c} \mathbf{d}_{\mathbf{Q}_c}$, then:
 - (iii) for every sentry function \mathbf{s} on \mathbf{Q}_c there exists a $\mathbf{s}_\mathbf{H}$ such that $\mathbf{s}(q) \subseteq \mathbf{s}_\mathbf{H}(h)$ for each h, q s.t. $q = c \cap h$.
 - (iv) $\mathbf{d}_\mathbf{Q} \leq \mathbf{d}_\mathbf{H}$.

Proof. Point (i) descends from the fact that $c \cup \tilde{h} \not\subseteq c \cup h$ means that $\tilde{h} - h \neq \emptyset$. Thus, if $(c \cup h)^+ \subseteq (c \cup \tilde{h})^+$ in addition, in order to sentinel both $c \cup \tilde{h}$ and $h' \supseteq (c \cup \tilde{h})$ a point $x \in \tilde{h} - h$ can be used belonging to the exclusive frontier of h w.r.t. \tilde{h} against every hypothesis in $up(c \cup h)$ sentinelled by other points of $\mathbf{S}(c \cup h)$. With this choice, each sentinelling point of $\mathbf{S}(d)$ is also a sentinelling point of a suitable $\mathbf{S}_\mathbf{H}(h)$, owing to property (ii) of Definition 5.² This implies $\mathbf{S}(d) \subseteq \mathbf{S}_\mathbf{H}(h)$, where $\mathbf{S}_\mathbf{H}(c) - \mathbf{S}(d)$ eventually includes further sentinels to concepts \hat{h} such that $h^+ \subseteq \hat{h}^+$ but $(h \cup c)^+ \not\subseteq (\hat{h}^+ \cup c)$.

² Indeed, assume that $x_1 \in (h_1 - h - c) - h'_2$ and $x_2 \in (h_2 - h - c) - h'_1$ belong to $\mathbf{S}(h \cup c)$ but does not belong to $\mathbf{S}_\mathbf{H}(h)$. In force of Definition 2, this requires that $x_2 \in \mathbf{S}(h_1 \cup c)$, $x_1 \in \mathbf{S}(h_2 \cup c)$ but $(x_2 \in \mathbf{S}(h'_1)) \wedge (x_1 \in \mathbf{S}(h'_2))$ cannot hold. This unavoidably happens for instance on h'_1 only if either x_1 also belongs to $\mathbf{S}(h'_1)$, or an $x_3 \in \mathbf{S}(h'_1)$ lies in a forbidden region to $\mathbf{S}(h_1 \cup c)$. These events would make x_2 useless, but are both impossible, since $x_2 \in h'_1$ and $h'_1 \supseteq c$.

Point (ii) comes from applying (i) to the worst sentry \mathbf{S}_{D_c} of the worst D_c on the worst $c \cup h$.

Points (iii) and (iv) are shown in the same way. \square

Notation We denote by $\mathbf{H} \div c$ the set $\{(h \div c \mid h \in \mathbf{H})\}$, remember that $d_{VC}(\mathbf{C} \div c) = d_{VC}(\mathbf{C})$ [11], and define $\mathbf{H} \div \mathbf{C} = \bigcup_{C \in \mathbf{C}} \mathbf{H} \div c$ and $D_{\mathbf{H}, \mathbf{C}} = \sup_{C \in \mathbf{C}} \{D_{\mathbf{H} \div c}\}$. The following properties hold:

Definition 6. For given concept classes \mathbf{C} and \mathbf{H} on \mathbf{X} , \mathbf{H} is *nested* w.r.t. \mathbf{C} if for each $h \in \mathbf{H}$ and for each $c \in \mathbf{C}$, a h' exists such that

- (i) $h \subseteq h'$ and
- (ii) for each pair h_1, h_2 such that $\langle c \div h_1, c \div h_2 \rangle \in up(c \div h)$ and $(c \div h_1) \cup (c \div h_2) \not\subseteq \langle c \div h_1, c \div h_2 \rangle$, $(h_1 \div h) \cup (h_2 \div h) \not\subseteq \langle h'_1 \div h, h'_2 \div h \rangle$.

Example 4. The classes of convex k -gons and the class of circles in a plane are nested on any set of convex figures.

Fact 5. (a) $D_{\mathbf{H}} \leq D_{\mathbf{H}, \mathbf{C} \cup \{\emptyset\}} \leq d_{VC}(\mathbf{H}) + 1$.

(b) *there exists a constant r such that $\langle D_{\mathbf{H}, \mathbf{C}}/D_{\mathbf{H}}, D_{\mathbf{H}, \mathbf{C}}/d_{\mathbf{H}} \rangle \leq r$.*

Proof. The proof passes through the following statements:

- (1) $D_{\mathbf{H}, \{\emptyset\}} = D_{\mathbf{H}}$.
- (2) $(d_{VC}(\mathbf{H}) - 1)/176 \leq \langle d_{\mathbf{H}}, D_{\mathbf{H}}, D_{\mathbf{H} \div c} \rangle \leq d_{VC}(\mathbf{H} \div c) + 1 = d_{VC}(\mathbf{H}) + 1$ for each $c \in \mathbf{C}$.

Fact 6. *For each concept class \mathbf{C} of concepts closed under sum and difference $D_{\mathbf{C}, \mathbf{C}} = D_{\mathbf{C}}$.*

Proof. For each $c, c' \in \mathbf{C}$, $(c' - c) \in \mathbf{C}$, $(c - c') \in \mathbf{C}$ and $(c' - c) \cup (c - c') \in \mathbf{C}$; $\mathbf{C} \div \emptyset = \mathbf{C}$. Then $\mathbf{C} \div \mathbf{C} = \mathbf{C}$.

Example 5. The power set \mathbf{C} of a real line and the class $\mathbf{C} \div \mathbf{C}$ of the symmetric differences of its elements both have detail $= \infty$.

Closure under sum and difference is a sharp property that does not appear to allow suitable results for finite detail classes. On the contrary, denseness and nestedness seem smoother properties of many pairs of wide families of concept and hypothesis classes. For these classes relations between details are specified by the following lemma.

Lemma 1. *Given a hypothesis concept class \mathbf{H} on \mathbf{X} , for each \mathbf{C} such that \mathbf{H} is dense w.r.t. \mathbf{C} , $D_{\mathbf{H}, \mathbf{C}} \leq D_{\mathbf{H}} + d_{\mathbf{H}}$. For each \mathbf{C} such that \mathbf{H} is nested w.r.t. \mathbf{C} , $D_{\mathbf{H}, \mathbf{C}} \leq D_{\mathbf{H}}$.*

Proof. For any $c \in \mathbf{C}$, let \mathbf{S} be a worst outer sentry function for $c \div \mathbf{H}$. For given h let us partition $\mathbf{S}(c \div h)$ in the set ${}^+\mathbf{S}(c \div h) = \mathbf{S}(c \div h) \cap \bar{c}$ and ${}^-\mathbf{S}(c \div h) = \mathbf{S}(c \div h) \cap c$ and $up(h)$ in the sets ${}^+\mathbf{H} \div c$ of hypotheses sentinelled by ${}^+\mathbf{S}(c \div h)$ and ${}^-\mathbf{H} \div c$ of

hypotheses sentinelled by $\neg \mathbf{S}(c \div h)$. Now, for each \mathbf{S} on $\mathbf{H} \div c$ a couple \mathcal{S} and s can be built on \mathbf{H} , such that for any \tilde{h} such that $c \div \tilde{h} \in \text{up}(c \div h)$:

- if \mathbf{H} is dense w.r.t. \mathbf{C} ,
 - (a) if $c \div \tilde{h} \in {}^+ \mathbf{H} \div c$, with reference to h' such that $\tilde{h} \subseteq h'$ mentioned in point (i) of Definition 5, a point belonging to $\tilde{h} - h$ that is used by ${}^+ \mathbf{S}(c \div h)$ to sentinel $\tilde{h} - c$ can be employed by $\mathbf{S}(h)$ to sentinel h' as well, owing to property (ii) of the same definition. Thus ${}^+ \mathbf{S}(h \div c) \subseteq \mathcal{S}(h)$. Moreover:
 - (b) a similar relation occurs between inner sentries of $h \cap c$ w.r.t. $\tilde{h} \cap c$ for $c \div \tilde{h} \in \neg \mathbf{H} \div c$. Therefore $\neg \mathbf{S}(h \div c) \subseteq s(h)$ and $\#\mathbf{S}(h \div c) \leq \#\mathcal{S}(h) + \#s(h)$.
- if \mathbf{H} is nested w.r.t. \mathbf{C} arguments of point (a) extend to the whole $\mathbf{H} \div c$. Thus $\#\mathbf{S}(h \div c) \leq \#\mathcal{S}(h)$.

Corollary 1. *The class \mathbf{C} of convex k -gons and the class \mathbf{C} of circles in a plane have $\mathbf{D}_{\mathbf{C}, \mathbf{C} \cup \{\emptyset\}} = \mathbf{D}_{\mathbf{C}}$.*

3. Learning concepts with a given detail

3.1. Basic results

PAC learning theory [11, 27] gives exact conditions for having that, given:

- a class \mathbf{C} of functions $c : \mathbf{X} \mapsto \{0, 1\}$
- a probability distribution P on \mathbf{X} and a sample $\mathbf{X}_m^c = \{(X_1, c(X_1)), \dots, (X_m, c(X_m))\}$
- a pair of real numbers ε and δ close to zero,

a function h is computable such that

$$P^{(m)}(E[l(h(X), c(X))] \leq \varepsilon) \geq 1 - \delta,$$

where $P^{(m)}$ is the probability measure in the product space \mathbf{X}^m and $l(h(X), c(X))$ is a loss function.

In this paper we will assume a boolean function that detects contradictions between $c(x)$ and $h(x)$ as the basic loss function, and will ground most of our results on the features of the sentries of the symmetric differences $c \div h$.

Definition 7. Given a probabilistic space $(\mathbf{X}, \mathcal{F}, P)$ – where \mathbf{X} is the set of the possible outcomes of a random data source, \mathcal{F} is a σ -algebra on \mathbf{X} , and P is a probability measure defined over \mathcal{F} – and a concept c , we denote by $c(x)$ the characteristic function of concept c and by *labelled sample* \mathbf{X}_m^c a set of m independent random pair $\{(X_1, c(X_1)), \dots, (X_m, c(X_m))\}$. We call *example of size m* any specification \mathbf{x}_m^c of \mathbf{X}_m^c . By a hypothesis H we mean any statistics on \mathbf{X}_m^c which defines a random subset of \mathbf{X} . For any given \mathbf{x}_m^c the specification h of H is said to be *consistent with \mathbf{x}_m^c* if for every $x_i, i = 1, \dots, m$, we have $c(x_i) = h(x_i)$.

Definition 8. Given a concept class \mathbf{C} on \mathbf{X} , by a *full learning algorithm* we mean a function $\mathcal{A} : \{\mathbf{x}_m^c\} \mapsto \{h\}$ such that for every $0 < \delta, \varepsilon < 1$ there is an integer $m^\circ > 0$

such that for every labelled sample X_m^c (i.e. for every P on X and $c \in C$) with $m \geq m^\circ$, denoting $H = \mathcal{A}(X_m^c)$, the probability $P_{\text{error}} \equiv P(c(X) \neq H(X))$ is bounded by the probabilistic inequality:

$$P^{(m)}(P_{\text{error}} \leq \varepsilon) \geq 1 - \delta.$$

If such a function exists, the class C is said to be *fully learnable*, where ε and δ are called *accuracy parameters* of the learning algorithm. We denote m° the sample complexity of the concept class w.r.t. \mathcal{A} and assume it to be a function of the accuracy parameters and possible indices on C . The restriction of \mathcal{A} to the set $\{\{x_m^c \mid m \geq m^\circ\}\}$ is said to be a learning algorithm with accuracy parameters ε and δ for C . \mathcal{A} is consistent if it computes only hypotheses consistent with its input.

An *approximate learning algorithm* with accuracy $\tilde{\varepsilon}$ and $\tilde{\delta}$ is a learning algorithm which works *only* for accuracy parameters $\varepsilon \geq \tilde{\varepsilon}$ and $\delta \geq \tilde{\delta}$.

Remark 3. A consistent learning algorithm \mathcal{A} is essentially made up of two parts: (i) a constrained one that (implicitly) builds a family of sentry functions S_c on $H \div c$ for each c , to frame the hypothesis in the labelled sample, and (ii) a free one that codes the user's preferences in selecting the output within the set of consistent hypotheses. Actually, since a concept class may have more sentry functions, also the first part may be affected by the user's style. Anyway each adopted sentry set must be a subset of the labelled sample X_m^c .

A further specification of the learning algorithm in order to state stringent results on sample complexity is the following:

Definition 9. Given a concept class C on X and a set $W \subseteq X$, let us denote by B a set of subsets of X , and by B_W the quotient set of B with respect to the equivalence relation on the subsets of X defined by having the same intersection with W . Denoting by ${}^m W_c$ the set of labelled examples x_m^c with all the components x_i belonging to W , a function $\mathcal{A} : \{x_m^c\} \mapsto C$ is *strongly surjective (ssu)* if for each subset Y of X_m , \mathcal{A} is a surjection from ${}^m Y_c$ onto C_Y .

Lemma 2 (Basic Lemma, Apolloni and Chiaravalli [5]). *Assume we are given:*

- a set X and its probability measure P ,
- a concept class C with $D_{C,C} = \mu$,
- a labelled sample X_m^c ,
- a consistent ssu function $\mathcal{A} : \{x_m^c\} \mapsto C$,

Consider the family of random sets $\{H_c = \mathcal{A}(X_m^c)\}$, with c varying in C .

For a given c , let us denote by U_c the random variable given by the probability measure of $H_c \div c$.

Let

$$I_x(\mu, m - \mu + 1) = 1 - \sum_{i=0}^{\mu-1} \binom{m}{i} \alpha^i (1 - \alpha)^{m-i}.$$

Then, for each c in \mathbf{C} and $0 < \alpha < 1$

$$P^{(m)}(U_c \leq \alpha) \geq I_x(\mu, m - \mu + 1). \quad (1)$$

Proof (Sketch, see the original proof for the omitted details). Consider a sequence $\mathbf{B}(c \div h^+) = B_1 \subseteq B_2 \subseteq B_3 \subseteq \dots$ of subsets of \mathbf{X} , such that $c \div h \cup \mathbf{S}(c \div h)$ belongs to $\mathbf{B}(c \div h^+)$ and the previous element in the sequence is included in $c \div h$. Consider also the companion sequence $\mathbf{U}(c \div h^+)$ of the probability measures of the sets in $\mathbf{B}(c \div h^+)$.

For a sample \mathbf{X}_m^c , $\mathbf{B}(H_c \div c^+)$ is a random sequence. Thus for a fixed α , the subset B_x of probability measure α in the sequence might include $c \div h$ or not, where $\mathbf{S}(H_c \div c)$ is the witness of this inclusion. In closer detail, sample \mathbf{X}_m^c contains the frontier of $H_c \div c$; thus, if this part of the sample is included in B_x we are sure that also $H_c \div c \subseteq B_x$. The implication chain is completed as follows:

- (i) on the right, by the fact that $\mathbf{S}(H_c \div c) \subseteq B_x$ if and only if $N_x \geq \#\mathbf{S}(H_c \div c)$, where N_x is the number of those from among the sampling points which fall in B_x , and
- (ii) on the left by the fact that the event $U_c \leq \alpha$ is implied by the event $U_c^+ \leq \alpha$, and the latter by $H_c \div c \subseteq B_x$, with the obvious notational extension: $U_c^+ = P(H_c \div c \cup \mathbf{S}(H_c \div c))$

Namely:

$$\begin{aligned} (a) \quad (U_c \leq \alpha) &\Leftarrow (U_c^+ \leq \alpha) \Leftarrow (H_c \div c \subseteq B_x) \Leftarrow (\mathbf{S}(H_c \div c) \subseteq B_x) \\ &\Leftarrow (N_x \geq \#\mathbf{S}(H_c \div c)) \end{aligned} \quad (2)$$

which induces the opposite chain on probabilities, after some technicalities on the value of α ,³

$$(b) \quad P^{(m)}(U_c \leq \alpha) \geq P^{(m)}(N_x \geq \#\mathbf{S}(H_c \div c)) \geq P^{(m)}(N_x \geq \mu). \quad \square \quad (3)$$

This lemma gives rise to the following theorem:

Theorem 2 (Apolloni and Chiaravalli [5]). *Given a concept class \mathbf{C} on \mathbf{X} with $D_{\mathbf{C},\mathbf{C}} = \mu$ and a labelled sample \mathbf{X}_m^c , for $0 < \varepsilon, \delta < \frac{1}{2}$, in case $m \geq \max\{2/\varepsilon \log(1/\delta), 5.5(\mu - 1)/\varepsilon\}$ any ssu function $\mathcal{A} : \{\mathbf{x}_m^c\} \mapsto \mathbf{C}$ outputting consistent hypotheses is a learning algorithm with accuracy parameters ε and δ for \mathbf{C} .*

For any concept class \mathbf{C} on \mathbf{X} , with $d_{\text{VC}}(\mathbf{C}) \geq 17$, and any labelled sample \mathbf{X}_m^c the ratio between maximum and minimum numbers of examples needed to learn \mathbf{C} with

³ Due to the pivoting role of $c \div h \cup \mathbf{S}(c \div h)$, it may be that B_x does not exist. In this case we work with $B_{x'}$, where α' is the first available value over α , and come back to α through the simple statements $(U_c \leq \alpha) \Leftarrow (U_c^+ \leq \alpha')$ and $P^{(m)}(N_{x'} \geq \mu) \geq P^{(m)}(N_x \geq \mu)$.

accuracy parameters $0 < \varepsilon < \frac{1}{8}$, $0 < \delta < \frac{1}{100}$ w.r.t. any probability measure P on \mathbf{X} is bounded by a constant.

Definition 10. Given a concept c within a class \mathbf{C} and a learning algorithm \mathcal{A} , which computes the hypothesis h within a class \mathbf{H} by implementing the family $\{\mathbf{S}_c\}$ of sentry functions on $\mathbf{H} \div \mathbf{C}$ (see Remark 3), the *number of degrees of freedom* $v_{c,\mathcal{A}}$ of a sample with respect to a hypothesis h in output to labelled sample x_m^c (for short, degrees of freedom of h) equals the sample size minus the number of points of the outer frontier of $c \div h$ against $\mathbf{H} \div c$:

$$v_{c,\mathcal{A}} = m - \mathbf{S}_c(c \div h).$$

Fact 7. For each $c \in \mathbf{C}$ and \mathcal{A} , $v_{c,\mathcal{A}} \geq m - \mathbf{D}_{\mathbf{H},\mathbf{C}}$.

Remark 4. What happens if the probability measure on \mathbf{X} is known? Actually all previous arguments remain true, except that some sentinels can be spared, since some subset of the frontier of c is sufficient for preventing *tangible* enlargement of $c \div h$. Therefore, the actual degrees of freedom of a labelled sample increase, with obvious consequences on sample complexity.

3.2. Learning through a subsymbolic device

In spite of the underlying sophisticated theory, the above or similar [11, 13, 20, 22] bounds on sample complexity are used only seldomly by people who want to learn h through a subsymbolical device such as a neural network. The actual size of the training set is of one or more order less than the size m of labelled sample requested by the sample complexity [10], where this discrepancy can be attributed to the following inner limitations of the PAC learning theory [3, 17]:

1. the bounds on m constitute worst case results;
2. the theory is not yet tailored to take into account the peculiarities of neural networks and learning algorithms in respect to the learning instance.

Thus, we generally disregard the bounds, use a short training set to train the neural network and check the adequacy of this set by testing the inferred h on a new set of examples (the test set) [3, 18]. If the network (i.e. h) performs on the test nearby as well as on the training set we are satisfied and declare the learning task successful.

In these pages, we try to narrow the gap between symbolic theory and subsymbolic practice, specialising the degrees of freedom of a labelled sample in some widespread learning instances.

The two usual drawbacks in working with an approximate subsymbolic class of hypotheses \mathbf{H} are: (d.1) we abound in the class elements to be sure to include c or, alternatively, (d.2) we may miss some detail in the approximate hypothesis since c does not belong to \mathbf{H} .

The former drawback generally results in a decreasing of the degrees of freedom of our statistic H , by Fact 3. However, we will see that for some H this inconvenience

may be balanced or even overcome by a saving of frontier points occurring when the permitted error ε is very low.

On the other hand, the effectiveness of the second drawback might either be almost irrelevant or translate into a benefit for ε large enough.

In the following, we state some results which hold for special features of the learnt hypotheses. Generally, we are not able to prove that these are the features of subsymbolical hypotheses, like those provided by neural networks. However, we will show common statistics owning these features and reasonably expect to find them, at least in a weaker form, in the hypotheses supplied by neural networks as well.

To better exploit these results we focus on a *dovetail* [30] issue of PAC learning algorithm described in Definition 11. For this algorithm and some simplified variants we prove a set of results concerning sample complexity that can be easily extended to other usual learning schemes tolerating labelling errors.

Definition 11. Given a concept class \mathbf{C} on \mathbf{X} and a tapering family $\tilde{\mathbf{H}}$ of classes of hypotheses \mathbf{H}_i such that $\mathbf{H}_i \subseteq \mathbf{H}_{i+1}$, a *H-large-as-needs (Hln)-learning procedure* \mathbf{A} is defined by the following steps:

Given \mathbf{x}_m^c

1. **Start** with $k = 0$ and $\mathbf{H} = \mathbf{H}_0$
2. **For** $i = 0$ to k
 3. **Search** for an almost consistent hypothesis $h \in \mathbf{H}$.
 4. **If** the number v of points x such that $c(x) \neq h(x)$ equals i , **then Stop**
5. **Set** $k = k + 1$, $\mathbf{H} = \mathbf{H}_k$ and **go to** step 2.

Let us denote by thresholded procedure \mathbf{A}_t the restricted version of \mathbf{A} where only \mathbf{H} changes during iterations while a fixed threshold t is set to the number of misclassified examples.

Remark 5. The rationale of the dovetailing algorithm relies in a balancing between computational costs, generally growing with the enlargement of \mathbf{H} and decrease of the number of faulty classified examples, and the learning accuracy running in the same direction. A companion learning scheme for \mathbf{A}_t is represented by a PAC procedure learning on the basis of a sample with at most t – malicious or non-malicious [2, 21] – labelling errors. Related operational fields might be constituted by learning with drifting distributions [9] or drifting concepts [19].

When a bounded number of misclassifications is allowed to h we get the following result:

Theorem 3. Given a concept class \mathbf{C} on \mathbf{X} , a hypotheses class \mathbf{H} with $D_{\mathbf{H}, \mathbf{C}} = \mu$ and a labelled sample \mathbf{X}_m^c , let \mathcal{A} be an approximate algorithm which misclassifies at most t points of total probability at most π . For each $0 < \varepsilon$, $\delta < \frac{1}{2}$, in case $m \geq \max\{2/\varepsilon \log(1/\delta), 5.5(\mu + t - 1)/\varepsilon\}$ if \mathcal{A} is a ssu function from $\{\mathbf{x}_m^c\}$ to \mathbf{H} , then \mathcal{A} is a learning algorithm with accuracy parameters $\max\{\pi, \varepsilon\}$ and δ for \mathbf{C} .

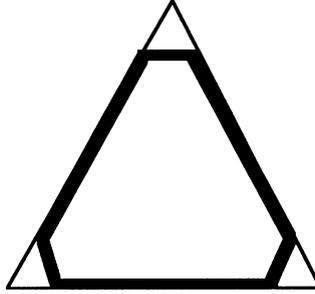


Fig. 6. An item of the class of smoothed triangles vs. triangles, and its companion hypothesis. Uniform distribution results $(3, \gamma)$ -careless.

Proof. Coming back to the sequences \mathbf{B} in the proof of Basic Lemma, and denoting by A_t the set of misclassified points, now we consider a sequence $\mathbf{B}(H_c \div c^{++})$ pivoted on the subset $H_c \div c \cup \mathbf{S}(H_c \div c) \cup A_t$ having $H_c \div c$ again as next antecedent in the sequence. Thus the witness of the inclusion of $H_c \div c$ in B_α is constituted by at most $\mu + t$ points of \mathbf{X}_m^c . Moreover $P(H_c \div c)$ now is lowerbounded by $P(A_t)$. Hence the claim of the theorem follows. \square

Corollary 2. Given a concept class \mathbf{C} on a probability space $(\mathbf{X}, \mathcal{F}, P)$, with P belonging to the family of continuous probability measures, a hypothesis class \mathbf{H} with $D_{\mathbf{H}, \mathbf{C}} = \mu$ and a labelled sample \mathbf{X}_m^c , let \mathcal{A} be an approximate algorithm which misclassifies at most t points. For each $0 < \varepsilon, \delta < \frac{1}{2}$, in case $m \geq \max\{2/\varepsilon \log(1/\delta), 5.5(\mu + t - 1)/\varepsilon\}$ if \mathcal{A} is a ssu function from $\{\mathbf{x}_m^c\}$ to \mathbf{H} , then \mathcal{A} is a learning algorithm with accuracy parameters ε and δ for \mathbf{C} .

Proof. Trivially because now $P(A_t) = 0$. \square

Going deeper in the subsymbolic learning drawbacks, let us start focusing on (d.2).

Definition 12. Given a concept class \mathbf{C} on \mathbf{X} , a hypothesis class \mathbf{H} , such that $\mathbf{C} \not\subseteq \mathbf{H}$, and an outer sentry function \mathbf{S} on \mathbf{C} , a probability distribution is (k, γ) -careless for \mathbf{C} w.r.t. \mathbf{H} if for each c there exist a hypothesis h and at most k sentinelling points of $\mathbf{S}(c)$ such that their violation allows an enlargement of c into an h through a region of probability no higher than γ .

Example 6. Let us consider the concept class \mathbf{C} of exagons constituted by triangles with smoothed angles in a square domain Ω , such that the euclidean measure of the difference between any triangle and a smoothed companion is less than or equal to γ times the measure of Ω (see Fig. 6). Then P uniform on Ω is $(3, \gamma)$ -careless for \mathbf{C} with respect to the class \mathbf{H} of full triangles.

Corollary 3. Given a concept class \mathbf{C} on a probability space $(\mathbf{X}, \mathcal{F}, P)$, a tapering family $\tilde{\mathbf{H}}$ of classes of hypotheses \mathbf{H}_i with $D_{\mathbf{H}_i, \mathbf{C}} = \mu_i$, and a labelled sample \mathbf{X}_m^c , let \mathcal{A}_t be a thresholded procedure. For each ε , if there exist a i and $\gamma < \varepsilon$ such that P is (t, γ) -careless for \mathbf{C} w.r.t. $\mathbf{H}_i \in \tilde{\mathbf{H}}$ and a ssu function \mathcal{A}_t from $\{\mathbf{x}_m^c\}$ to $\tilde{\mathbf{H}}$, then for each $0 < \delta < \frac{1}{2}$, in case $m \geq \max\{2/\varepsilon \log(1/\delta), 5.5(\mu_i + t - 1)/\varepsilon\}$ \mathcal{A}_t it is an approximate learning algorithm with accuracy parameters ε and δ for \mathbf{C} .

Proof. \mathcal{A}_t stops no later than at the i th iteration, since for each c there exists a $h \in \mathbf{H}_i$ such that c enlarges in h violating at most k points by definition. Thus a consistent hypothesis exists in \mathbf{H}_i . If on each sample \mathbf{X}_m^c \mathcal{A}_t stops exactly at the i th iteration, in the sequence $\mathbf{B}(H_{i,c} \div c^+)$ the witness of inclusion is constituted by at most $\mu_i + t$ points of \mathbf{X}_m^c , where at most k of them are inside $c \div H_c$ and the remaining ones on the frontier. Moreover, $P(c \div H_c)$ is lowerbounded at most by γ ; thus the corollary claim comes directly from Theorem 3. If \mathcal{A}_t some times stops before the i th iteration, then the involved detail is $\leq \mu_i$, and the corollary claim holds as well, since all else remains unchanged. \square

Remark 6. Theorem 2 sanctions a sample complexity overhead to lazy learners who use thresholded procedures. By contrast, Corollary 3 rewards sagacious learners who compare the accuracy of the hypothesis class with the accuracy target of the learning task. For a given ε , they possibly point to compatibly approximate and detail cheap hypotheses class \mathbf{H}_i , so that $\mu + t$ in Corollary 3 is less than $D_{\mathbf{C}, \mathbf{C}}$. The dovetailed version of the thresholded algorithm looks exactly for this sagacity, with a suitable balancing between threshold and detail. However, no general results hold in this case, since the threshold is now a random variable as well.

Definition 13. Given subsets a and b of a metric set \mathbf{X} and a natural contiguity⁴ relation between points belonging to \mathbf{X} , we denote by number of intersections χ the number of maximal contiguous alternating homogeneous sets constituting the symmetric difference $a \div b$. With reference to the labels attributed by the characteristic functions $a(x)$ and $b(x)$ to the points of \mathbf{X} , a subset is homogeneous if each point is affected by the same pair of labels; it is maximal if no enlargement through contiguous points has the same property. Two subsets are contiguous if at least one point of the first set is contiguous to a point of the second set; they are alternating if they are homogeneous with different pairs of labels (see Fig. 7)

Definition 14. Given a concept class \mathbf{C} on \mathbf{X} , a *criss-cross* hypothesis class $\mathbf{H}_{cc} = \bigcup \mathbf{H}_{cc; \lambda}$ w.r.t. \mathbf{C} is a set of hypotheses indexed by λ such that: (1) $\mathbf{H}_{cc} \cap \mathbf{C} = \emptyset$ and (2) for uniform probability measure U on \mathbf{X} and each c , $\min(U(c \div h))_\lambda$ is a nonincreasing function of the number χ of intersections of c with h , where the minimum is taken over all $h \in \mathbf{H}_{cc; \lambda}$. A *uniform criss-cross* hypothesis class \mathbf{H}_{ccu} is a criss-cross class

⁴ This relation is a topological counterpart of the concept representation notion used in late definitions of learnability [20].

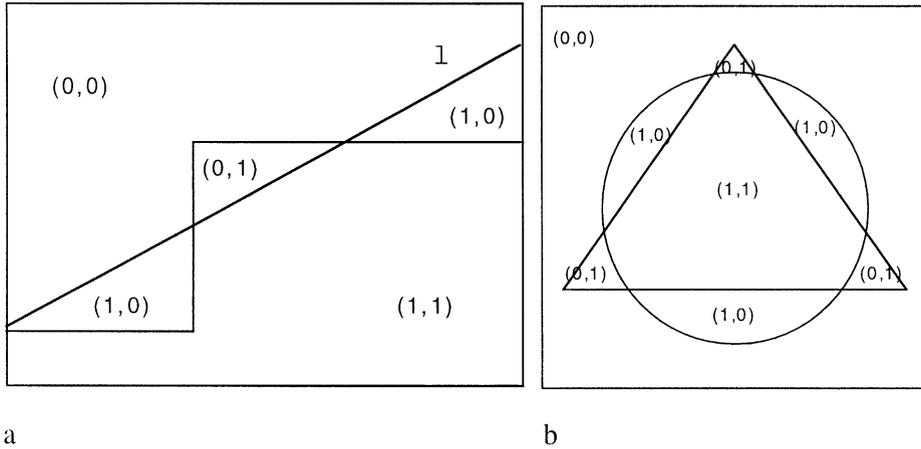


Fig. 7. Criss-cross hypotheses. (a) \mathbf{C} = half-planes, \mathbf{H} = fret delimited regions; (b) \mathbf{C} = circles, \mathbf{H} = triangles. Set label = pair of uniform values of concept and hypothesis characteristic functions. (1, 0) or (0, 1) labelled sets are homogeneous and maximal, and alternate when contiguous. (a) number of intersections = 3; (b) number of intersections = 6.

such that: (i) for each c and ε , a λ and $h \in \mathbf{H}_{cc;\lambda}$ exist such that $U(c \div h) < \varepsilon$, and (ii) there exists a λ^* such that for any $\lambda \geq \lambda^*$ $\min(U(c \div h))_\lambda / \min(U(c \div h))_{\lambda-1} < 1$. Finally, \mathbf{H}_{ccu} is a *smart* class if for any $k \geq 0$ there exist a λ^* and an ε , such that for any $\lambda \geq \lambda^*$ and any $h \in \mathbf{H}_{cc;\lambda}$ with $U(c \div h) \leq \varepsilon$, $\#\mathbf{S}(c \div h) \leq k$.

Example 7. The class of fret delimited regions is uniformly criss-cross w.r.t. the class of half-planes, parametrised in the number of frets per unitary segment of the concept delimiter. The class of the regular polygons, parametrised in the number of their edges, is uniformly criss-cross w.r.t. the class of circles as well (see Fig. 7).

In particular, concerning the first \mathbf{H}_{ccu} , let us denote by 1 the straight line delimiting a concept, by $(c \div h; a)$ the part of $c \div h$ insisting on a segment of 1 of length a and by θ the absolute value of the slope of 1 w.r.t. a coordinate axis. Then

$$\lim_{a \rightarrow \infty} \frac{\min(U(c \div h; a))_\lambda}{a} = \frac{\min\{\theta, 1/\theta\}}{2(\lambda + 1)}.$$

The minimum is reached when all the frets cross 1 and the crossing points are equidistant. Thus λ coincides with the number of these crosses per unitary segment and $\min(U(c \div h))_\lambda / \min(U(c \div h))_{\lambda-1} < \frac{1}{2}$. Denoting by h^* a hypothesis minimising $U(c \div h)$, it is evident by simple inspection that $\#\mathbf{S}(c \div h^*) = 0$.

Definition 15. Given a concept class \mathbf{C} on \mathbf{X} and a uniform criss-cross hypothesis class \mathbf{H}_{cc} , a probability distribution P is *unbiased* if for each c $\min(P(c \div h))_\lambda$ is an increasing function g of $\min(U(c \div h))_\lambda$, with $g(0) = 0$.

Corollary 4. *Given a concept class \mathbf{C} on \mathbf{X} and a hypotheses class \mathbf{H} such that for a given $0 < \varepsilon < \frac{1}{2}$ and each c a hypothesis $h \in \mathbf{H}$ exists such that $P(c \div h) \leq \varepsilon$, let be $(\mathbf{D}_{\mathbf{H},\mathbf{C}})_\varepsilon$ the detail of the symmetric differences $c \div h$ such that $P(c \div h) \leq \varepsilon$. Then, for each $0 < \delta < \frac{1}{2}$, in case $m \geq \max\{2/\varepsilon \log(1/\delta), 5.5((\mathbf{D}_{\mathbf{H},\mathbf{C}})_\varepsilon - 1)/\varepsilon\}$ any consistent ssu function $\mathcal{A} : \{\mathbf{x}_m^c\} \mapsto \mathbf{C}$ outputting consistent hypotheses is a learning algorithm with accuracy parameters ε and δ .*

Proof. The claim comes directly from Basic Lemma, once we state that the witness of inclusion of $c \div H_c$ in B_ε is constituted of at most $(\mathbf{D}_{\mathbf{H},\mathbf{C}})_\varepsilon$ points. \square

Definition 16. Given concept and hypotheses classes \mathbf{C} and \mathbf{H} , respectively, denoting by $ub(\varepsilon, \delta)$ the upperbound on the sample complexity stated in Theorem 2, we say that:

- \mathbf{H} is *cheap* if there exists a ε° such that $(\mathbf{D}_{\mathbf{H},\mathbf{C}})_\varepsilon \leq \mathbf{D}_{\mathbf{H}}$ for each $\varepsilon < \varepsilon^\circ$.
- A *production prize* intervenes for $\varepsilon' < \varepsilon''$ if $ub(\varepsilon', \delta) < ub(\varepsilon'', \delta)$ for each δ .
- h is a *best hypothesis* for c if $\#\mathbf{S}(c \div h) = 0$ for every \mathbf{S} .

Fact 8. *Any hypothesis class \mathbf{H} nested w.r.t. a concept class \mathbf{C} is cheap.*

Theorem 4. *Given a concept class \mathbf{C} and a smart \mathbf{H}_{ccu} on a probability space $(\mathbf{X}, \mathcal{F}, P)$ with P unbiased, then:*

- (1) \mathbf{H}_{ccu} is *cheap*,
- (2) *there exists a ε° such that for some pairs $\varepsilon' < \varepsilon^\circ$ and $\varepsilon'' > \varepsilon^\circ$ a production prize intervenes,*
- (3) *for each c there exists a ε° such that for each $\varepsilon < \varepsilon^\circ$ any h with $U(c \div h) < \varepsilon^\circ$ is a best hypothesis for c .*

Proof. Since $\min(U(c \div h))_\lambda / \min(U(c \div h))_{\lambda-1} < 1$ for each ε^* small enough there exists a λ^* such that for each $\lambda < \lambda^*$ and $h \in \mathbf{H}_{\text{cc};\lambda}$ $U(c \div h) > \varepsilon^*$. From the smartness assumption on \mathbf{H}_{ccu} , for each k there exist ε'^* and λ'^* such that for each $\lambda > \lambda'^*$ and $h \in \mathbf{H}_{\text{cc};\lambda}$ we have that $U(c \div h) < \varepsilon \Rightarrow \#\mathbf{S}(c \div h) \leq k$. Thus, for any k there exists a $\varepsilon^* < \varepsilon'^*$ such that any h with $U(c \div h) < \varepsilon^*$ belongs to a $\mathbf{H}_{\text{cc};\lambda}$ with $\lambda \geq \lambda^*$ and its frontier satisfies: $\#\mathbf{S}(c \div h) \leq k$. Point (1) comes from $k = \mathbf{D}_{\mathbf{H}}$, point (3) from $k = 0$, as well.

Concerning point 2, let us consider the probability inequality (1) of Basic Lemma for $\alpha = \varepsilon$ and μ either equals 2 or equals 1. Namely, meaning by $p_\mu(\alpha) = I_\alpha(\mu, m - \mu + 1)$ the lowerbound on $P^{(m)}(U_c \leq \alpha)$ for given μ :

$$p_2(\varepsilon) = (1 - \varepsilon)^m + m\varepsilon(1 - \varepsilon)^{m-1},$$

$$p_1(\varepsilon) = (1 - \varepsilon)^m.$$

As mentioned in the first part of the proof, both these values of μ are attained by $(\mathbf{D}_{\mathbf{H},\mathbf{C}})_\varepsilon$ with decreasing ε . Let ε° be the switching value inducing $(\mathbf{D}_{\mathbf{H},\mathbf{C}})_\varepsilon$ jumping from 2 to 1.

Then, since $(1 - \varepsilon)^m < (1 - \varepsilon)^m + m\varepsilon(1 - \varepsilon)^{m-1}$ for each m and ε , two $\varepsilon', \varepsilon''$ exist such that $\varepsilon' < \varepsilon^\circ$ and $\varepsilon'' > \varepsilon^\circ$ and $(1 - \varepsilon')^{m-1} < (1 - \varepsilon'')^m + m\varepsilon''(1 - \varepsilon'')^{m-1}$. Thus $ub(\varepsilon', \delta) < ub(\varepsilon'', \delta)$. \square

Remark 7. Note that, in case $\#\mathbf{S}(c \div h) = 0$ we need one witness point, however, to use the implication chain (2) of Basic Lemma. Therefore, in this case we have a *virtual* frontier cardinality = 1.

Fact 9. Given a concept class \mathbf{C} and a smart class \mathbf{H}_{ccu} , a Hln-learning procedure A with $\tilde{\mathbf{H}}$ indexed by the parameter λ of the class, enjoys the decreasing of $D_{\mathbf{H}_{\text{cc}, \lambda}, \mathbf{C}}$ with the procedure iterations. Dovetailing exploits the benefit coming from the monotone reduction of $\min(U(c \div h))_\lambda$ with iterations, looking for hypotheses with few misclassified examples. Thus, from Theorem 3 a twice actual production prize comes from both detail reduction, possibly till 0, and lower sample misclassification overhead.

What about the training of a neural network? This device computes a non-linear function equipped with a lot of free parameters that have to be inferred from the training set.

If we come to learning boolean functions, the training set is exactly a labelled sample \mathbf{X}_m^c and the trained network computes a boolean function as well. A single neuron can be roughly identified by a hyperplane halving the sample space in a crisp or fuzzy way, depending on the shape function computed by the neuron.⁵ Thus, disregarding fuzziness effects, we might upperbind detail $D_{\mathbf{H}_{\text{NN}}}$ of the class of hypotheses \mathbf{H}_{NN} computed by a network with r neurons, each with fan-in n by:

$$D_{\mathbf{H}_{\text{NN}}} \leq r \times n,$$

n being an upperbound to the detail of the single hyperplane.

As a matter of fact, $D_{\mathbf{H}_{\text{NN}}}$ is generally much lower, because dependencies between the hyperplanes induced by the network architecture or direct constraints on the coefficients of the single hyperplane narrow \mathbf{H}_{NN} significantly. In this sense each embedding of formal knowledge (such as in [7, 8, 15] for instance), acting as further constraints, contributes to the above narrowing [1]. This might give rise to a variety of detail reduction amounts, ranging from few units up to the extreme case where $D_{\mathbf{H}_{\text{NN}}} = 0$.

Actually detail = 0 is not a degenerate case when we refer to the class of symmetric differences between concepts and neural hypotheses if the hyperplanes have some constraints which render their composition a best hypothesis according to Definition 13 and Example 7. This would lead to a paradoxical but non-infrequent learning instance where, over a given accuracy learning by subsymbolic tools is more accurate than if we take into account the concept class to which the goal concept belongs (*prejudice remotion*).

⁵ For instance, in case of heavyside function we have a crisp partitioning; in case of sigmoid function we have a fuzzy partitioning.

His learning procedure is a faithful modelisation of usual training stories. We try to succeed with small networks aiming to take zero training errors. Then we enrich the neural architecture, lowering the training accuracy pretences as training becomes more and more computationally expensive. Actually, we cannot state in general that current H_i does not contain a consistent (up to few errors) hypothesis. We only state that such a hypothesis is computationally unfeasible. On each new (enlarged) architecture first we check to see whether a zero error training is feasible, then we relax the accuracy target.

4. What size needs testing?

4.1. Testing boolean functions

What news can testing give us that we did not yet know from training?

If we learn boolean functions, in the case we know the detail of the class of involved symmetric differences and are sure that the training set is representative of the probability measure on \mathbf{X} , we do not expect more hints on the accuracy of the learnt hypothesis than those supplied by the theorems of the previous section.

Actually, if we refer for instance to neural networks, we see that D_{H_m} is generally difficult to compute, and its extension to symmetric differences as well, since it depends in a non-easy way on the architectural and formal constraints and on the target accuracy in connection with the features of the probability measure on \mathbf{X} . In the lack of $D_{H,C}$ the accuracy parameters ε and δ have more of a fuzzy than a probabilistic meaning to the learner, referred to as generalisation capability of the network. For a sample size large enough – so that for instance $I_c(1, m) = 1 - \delta$, with ε and δ low enough – if the learning algorithm produces a hypothesis h consistent with the sample we can count ourselves satisfied. The only doubt is that the actual degrees of freedom are much lower than m or a non-fair sample was drawn, so that our perception “zero errors over m ” is misleading. But if we check our hypotheses on m new examples and count zero errors again, we are confirmed in our perception. In fact, now the degrees of freedom are exactly m , as stated by the following lemma.

Lemma 3. *Given a concept c and a hypothesis h , for a labelled sample X_m^c , if h is consistent with the whole sample, then:*

$$P^{(m)}(U_c \leq \alpha) \geq I_x(1, m). \quad (4)$$

Proof. This is an obvious extension of the proof of Basic Lemma, where the number of witnessing points now is 1, in force of Remark 7. \square

Inequality (4) says that now the degrees of freedom of X_m^c are exactly m . Exchanging [32] the randomness of U_c with the randomness of the event “no sample point falls in $c \div h$ ”, this has two operational meanings.

- (1) In the frame of interval estimates, $(0, \varepsilon)$ is the confidence interval for the measure $P(c \div h)$ at confidence level $I_\varepsilon(1, m) = 1 - (1 - \varepsilon)^m$.
- (2) In the perspective of tests of hypothesis, $1 - I_\varepsilon(1, m) = (1 - \varepsilon)^m$ is the upperbound to the risk β of accepting a neural hypothesis whose symmetrical difference $c \div h$ with the goal concept has a measure $P(c \div h) > \varepsilon$.

As $I_\varepsilon(1, m)$ is the maximum of $I_\varepsilon(D_{\mathbf{H}, \mathbf{C}}, m - D_{\mathbf{H}, \mathbf{C}} + 1)$ in light of Remark 7, we can conclude that:

- in spite of the general use of employing very large test sets mainly because their processing is relatively inexpensive and, of course, more is better than less,
- in spite of the general wisdom [18] that much more examples are suitable for testing than for training a neural network,

vice versa, we conclude that it is generally wasteful to use a test set larger than a training set.

In order to quantify these considerations, we must recall the following result:

Theorem 5 (Apolloni and Chiaravalli [5] and Blumer et al. [11]). *For every concept class \mathbf{C} and $c \in \mathbf{C}$, for any learning algorithm and $0 < \alpha < 1$, there exists a probability distribution P such that*

$$P^{(m)}(U_c \leq \alpha) \leq I_\alpha(1, m).$$

Therefore, in case $m < \lg(1/\delta)1/(-\lg(1 - \varepsilon))$ no function $\mathcal{A} : \{\mathbf{x}_m^c\} \mapsto \mathbf{H}$ is a learning algorithm with accuracy parameters ε and δ for \mathbf{C} .

Remark 8. It is easy to note that sample complexity is exactly $\lceil \lg(1/\delta)1/(-\lg(1 - \varepsilon)) \rceil$ for $D_{\mathbf{H}, \mathbf{C}} = 1$. Moreover, the complexity lower bound is a non-decreasing function of $D_{\mathbf{H}, \mathbf{C}}$ as well. Then the ratio between the sizes of training and testing set can be suitably bounded as follows.

Theorem 6. *For a concept class \mathbf{C} and hypothesis class \mathbf{H} on \mathbf{X} with $D_{\mathbf{H}, \mathbf{C}} = \mu$, any – known or unknown – probability measure P on \mathbf{X} , and any accuracy target (ε, δ) the ratio between the cardinalities of the labelled examples needing to achieve the same accuracy target both in learning and in testing a hypothesis is at most 1.*

Proof. In the distribution free case the claim comes directly from Remark 8. If we are faced by a special distribution law, some sentinels can be spared, as mentioned in Remark 4. Therefore, the difference between sample size and number of degrees of freedom is lower than in the distribution free case but always of the same sign. \square

Remark 9. Let us denote by m and n the upperbounds on the cardinalities of the labelled samples deriving from Lemmas 2 and 3 on a same accuracy target both in learning and in testing a hypothesis, respectively. From an elementary algebra after imposing $P^{(m)}(U_c \leq \alpha)$ in (1) greater than $P^{(n)}(U_c \leq \alpha)$ in (4) we have that the ratio r between m and n obeys the inequality $r \geq 1/(1 + \mu/n)$, with μ as in Theorem 6.

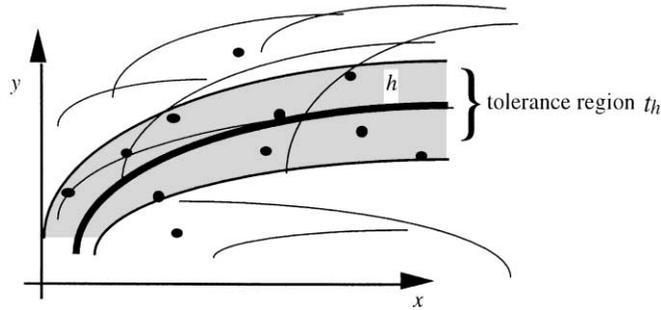


Fig. 8. From among the family \mathbf{F} of functions the learning algorithm selects a hypothesis h (bold line). This corresponds to selecting a tolerance region t_h from among a concept class \mathbf{C} .

Remark 10. In the sentry function perspective, the reader should recognise at the basis of the well-known overfitting phenomena an overdetailed \mathbf{H} in output to \mathcal{A} , leading to an excessive $D_{\mathbf{H},\mathbf{C}}$, rather than an excessive number of training cycles *per se*.

4.2. Extensions

The extension of PAC learning model to real functions did not get, after some basic results (see for instance [16]), the same attention from the scientist community that boolean functions did. Actually regression theory is so rich that the learning approach generally appears as needless as complex.

However, here we will extend our approach to these functions just to argue about the main question of this section. Namely, we will consider two learning schemes that, though not commonly employed, might represent the rationale of many widespread algorithms. Also, within these schemes we can conclude that testing the approximation capability of a neural network generally demands a smaller sample size than training does.

4.2.1. Learning tolerance regions

Given a family \mathbf{F} of functions, from \mathbf{X} to \mathbb{R} and an element f of the family, let us consider a tolerance region in $\mathbf{X} \times \mathbb{R}$ containing f with some slack, like in Fig. 8.⁶ Let us look at the set of tolerance regions around the element of \mathbf{F} as a concept class \mathbf{C} and apply the results of the previous section. Thus we obtain the following:

Lemma 4. *Given a family \mathbf{F} of functions on \mathbf{X} , an associate class \mathbf{C} of tolerance regions t_c with $d_{\mathbf{C}} = \mu$ and a sampling $(\mathbf{X}, \mathbf{Y})_m$, with $y = f(x)$ and $f \in \mathbf{F}$, consider any ssu function $\mathcal{A} : \{(\mathbf{x}, \mathbf{y})_m\} \mapsto \mathbf{F}$ outputting hypotheses whose associate tolerance region t_h contains at least a fraction $1 - \gamma$ of sampled points. Assume π be the total*

⁶ A good perspective for facing these regions is to see them as an extension of tolerance regions introduced by Tukey [25, 26].

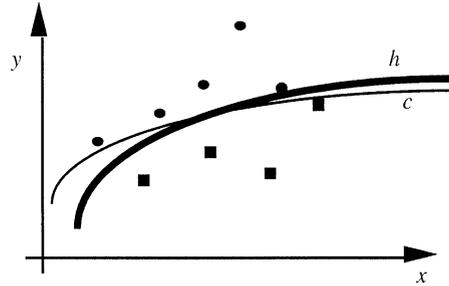


Fig. 9. From the same class of function as in Fig. 8 we draw a hypothesis h based on $\bullet = G$ (reater)E(qual)points and $\blacksquare = L$ (ower) points.

probability of the remaining γ of sampled points, and define

$$\ell_1(h(x), f(x)) = \begin{cases} 1 & \text{if } (x, f(x)) \notin t_h, \\ 0 & \text{otherwise.} \end{cases}$$

Then, for $0 < \varepsilon, \delta < \frac{1}{2}$, in case

$$m \geq \frac{\max\{(2/\varepsilon) \log(1/\delta), 5.5(\mu - 1)/\varepsilon\}}{1 - \gamma},$$

$$P^{(m)}(E[\ell_1(h(X), f(X))] \leq \max\{\pi, \varepsilon\}) \geq 1 - \delta.$$

Proof. It follows directly from Theorem 3, since $E[\ell_1(h(X), f(X))] = P((x, f(x)) \notin t_h)$. \square

4.2.2. GE learning

In the GE learning protocol [6] the training set $(\mathbf{x}, \mathbf{y})_m^c$ is constituted, like in Fig. 9, by triplets $(x, y, \lambda_f(x, y))$ where $\lambda_f(x, y) = 1$ if $y \geq f(x)$, 0 otherwise. Taking into account that the difference between h and f gives rise to a symmetric_difference-like region – let us call *twist concept* – where the label attributed to the points by f and h are different, we can state the following theorem, whose proof is omitted:

Lemma 5. Given a family \mathbf{F} of functions on \mathbf{X} and an associate class \mathbf{C} of twist concepts with $\mu = D_{\mathbf{C}}$, and a GE sample $(\mathbf{X}, \mathbf{Y})_m^c$, consider any ssu function $\mathcal{A} : \{(\mathbf{x}, \mathbf{y})_m^c\} \mapsto \mathbf{F}$ outputting hypotheses consistent with the sample, and define

$$\ell_2(h(x), f(x)) = \begin{cases} 1 & \text{if } \lambda_h(x, y) \neq \lambda_f(x, y), \\ 0 & \text{otherwise.} \end{cases}$$

Then, for $0 < \varepsilon, \delta < \frac{1}{2}$, in case

$$m \geq \frac{\max\{(2/\varepsilon) \log(1/\delta), 5.5(\mu - 1)/\varepsilon\}}{1 - \gamma},$$

$$P^{(m)}(E[\ell_2(h(X), f(X))] \leq \varepsilon) \geq 1 - \delta.$$

Based on of the above lemmas we can state the following extension of Theorem 6.

Theorem 7. *For a function class \mathbf{F} on \mathbf{X} and any accuracy target (ε, δ) on learning \mathbf{F} through either tolerance or twisting regions, the ratio r between the cardinalities of the labelled examples needing to achieve the same accuracy target both in learning and in testing a hypothesis is at most 1.*

From among the variety of further extensions that can be raised from the basic results, we quote only the following instances:

- i. We determine the learning success on the basis of the number of testing errors (not necessarily equal to zero). This causes a further degrees of freedom dropping analogous to what happens in Theorem 3.
- ii. We are interested in other accuracy parameters, different from probabilities ε and δ . In most cases it is just a matter of a different error representation, while the degrees of freedom do not depend on these representations.

5. Conclusions

According to the common experience of any student, learning occurs through relevant examples emerging during the lesson. These examples essentially represent the shattered set at the basis of the Vapnik–Chervonenkis dimension, but their individual management is easier and more useful than the mentioned dimension in respect to the complexity of learning tasks.

Namely, their individual management allows us to deal with various approximate issues of learnability, stating a more robust bridge between theoretical results of computational learning theory and the well-spread practice of subsymbolical learning. Owing to a special production prize effect, we show a learning instance where symbolical knowledge represents an overhead with respect to pure subsymbolic learning, and vice versa we give a rationale to the general improvement introduced by this kind of knowledge.

Finally, we state a clear relation between the two phases of training and generalisation in learning procedures, just in terms of degrees of freedom of the related samples. For equally sized samples these degrees are always larger in number in the generalisation phase.

The key for reading these degrees of freedom in terms of sample complexity of the learning task is provided by Basic Lemma, which mainly constitutes an extension of confidence interval theory to functionally dependent random samples. This is an interesting way to shed light on the connection between randomness, independence and underlying functional relations between data, that will be stressed further in upcoming works. The implication chain in Basic Lemma exploits a monotone relation between model parameters and outcomes statistics in the easy case of a Bernulli distribution

law. The statement of similar monotone relations for more complex models should allow us to assess learning procedures for non-boolean functions as well, in a more direct way than in the tricky extensions of Section 4.2.

References

- [1] Y.S. Abu-Mostafa, Hints and the VC dimension, *Neural Comput.* 5 (1993) 278–288.
- [2] D. Angluin, P.D. Laird, Learning from noisy examples, *Mach. Learning* 2 (2) (1988) 343–370.
- [3] B. Apolloni, Design of algorithms for neural networks, Supervised learning, *Comput. Artificial Intelligence* 5 (1992) 457–480.
- [4] B. Apolloni, F. Baraghini, G. Palmas, PAC meditation on boolean formulas, T.R. University of Milano, 1997.
- [5] B. Apolloni, S. Chiaravalli, PAC Learning of concept classes through the boundaries of their items, *J. Theoret. Comput. Sci.* 172 (1997) 91–120.
- [6] B. Apolloni, C. Ferretti, G. Mauri, Approximation of optimization problems and learnability, *Proc. GAA 92*, Roma, 1992 261–268.
- [7] B. Apolloni, A. Piccolboni, E. Sozio, Hybrid symbolic subsymbolic system for controlling a single link flexible arm, *J. System Eng.* 6 (1996) 208–222.
- [8] B. Apolloni, G. Zamponi, A.M. Zanaboni, Learning fuzzy decision trees, *Neural Networks* 11 (1998) 885–895.
- [9] P. Bartlett, Learning with a slowly changing distribution *Proc. 5th Workshop on Comput. Learning Theory* 1992, pp. 243–252.
- [10] E.B. Baum, D. Haussler, What size net gives valid generalizations? *Neural Comput.* 1 (1989) 151–160.
- [11] A. Blumer, A. Ehrenfeucht, D. Haussler, M. Warmuth, Learnability and the Vapnik-Chervonenkis dimension, *J. ACM* 36 (1989) 929–965.
- [12] D.P. Dobkin, D. Gunopulous, Concept Learning with geometric hypotheses <ftp://ftp.cs.princeton.edu/pub/people/dpd/DobkinGunopulous.ps.Z>.
- [13] A. Ehrenfeucht, D. Haussler, M. Kearns, L.G. Valiant, A general lower-bound on the number of examples needed for learning, *Inform. and Comput.* 82 (3) (1988) 247–251.
- [14] W. Feller, *An Introduction to Probability Theory and its Applications*, Wiley, New York, 1960.
- [15] P. Frasconi, M. Gori, M. Maggini, G. Soda, Unified integration of explicit knowledge and learning by example in recurrent networks, *IEEE Trans. Knowledge Data Eng.* 7 (2) (1995) 340–346.
- [16] D. Haussler, Generalizing the PAC model for neural net and other learning applications, *Res. Rep. UCSC-CRL-89-30*, University of California, Santa Cruz, 1989).
- [17] J. Hertz, A. Krogh, R.G. Palmer, *Introduction to the Theory of Neural Computation*, Addison-Wesley, Reading, MA, 1991.
- [18] P. Hecht-Nielsen, *Neurocomputing*, Addison-Wesley, Reading, MA, New York, 1989.
- [19] D.P. Helmbold, P.M. Long, Tracking drifting concepts by minimizing disagreements, *Mach. Learning* 14 (1994) 27–34.
- [20] M. Kearns, Efficient noise-tolerant learning from statistical queries, *Proc. 25th Annual ACM Symp. on Theory of Comput.*, ACM Press, New York, 1993, pp. 392–401.
- [21] M. Kearns, M. Li, Learning in the presence of malicious errors, *Proc. 20th annual ACM Symp. on Th. of Comput.*, ACM Press, New York, 1988, pp. 267–280.
- [22] B.K. Natarajan, On learning boolean functions, *Proc. 19th ACM Symp. on Theory of Computing*, Ass. Comp. Mach., 1987, NY, New York, pp. 285–295.
- [23] K. Pearson, *Tables of the Incomplete Beta Function*, Cambridge University Press, Cambridge, England, 1934.
- [24] D.E. Rumelhart, J.L. McClelland, the PDP research group, *Parallel Distributed Processing*, MIT Press, Cambridge, MA, 1986.
- [25] J.W. Tukey, Non parametric estimation II, Statistical equivalent blocks and tolerance regions – The continuous case, *Ann. Math. Statist.* 18 (1947) 529–539.
- [26] J.W. Tukey, Non parametric estimation III, Statistical equivalent blocks and multivariate tolerance regions – The discontinuous case *Ann. Math. Statist.* 19 (1948) 30–39.

- [27] L.G. Valiant, A theory of the learnable, *Commun. ACM* 27 (11) (1984) 1134–1142.
- [28] V.V. Vapnik, *Estimation of Dependencies based on Empirical Data*, Springer, New York, 1982.
- [29] V.V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.
- [30] M. Li, P. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*, Springer, Berlin, 1993.
- [31] R.S. Wencour, R.M. Dudley, Some special Vapnik–Chervonenkis classes, *Discr. Math.* 33 (1981) 313–318.
- [32] S.S. Wilks, *Mathematical Statistics*, Wiley, New York, 1962.