Neurocomputing 192 (2016) 14–28

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Can machine learning explain human learning?

Mehrnoosh Vahdat [a,b], Luca Oneto [a,*], Davide Anguita [c], Mathias Funk [b], Matthias Rauterberg [b]

[a] DITEN – University of Genova, Via Opera Pia 11a, I-16145 Genova, Italy
[b] Department of Industrial Design, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
[c] DIBRIS – University of Genova, Via Opera Pia 13, I-16145 Genova, Italy

ARTICLE INFO

Article history:
Received 10 July 2015
Received in revised form
9 October 2015
Accepted 17 November 2015
Available online 8 March 2016

Keywords:
Machine learning
Human learning
Rademacher Complexity
Algorithmic stability
Exploratory experiments on students

ABSTRACT

Learning Analytics (LA) has a major interest in exploring and understanding the learning process of humans and, for this purpose, benefits from both Cognitive Science, which studies how humans learn, and Machine Learning, which studies how algorithms learn from data. Usually, Machine Learning is exploited as a tool for analyzing data coming from experimental studies, but it has been recently applied to humans as if they were algorithms that learn from data. One example is the application of Rademacher Complexity, which measures the capacity of a learning machine, to human learning, which led to the formulation of Human Rademacher Complexity (HRC). In this line of research, we propose here a more powerful measure of complexity, the Human Algorithmic Stability (HAS), as a tool to better understand the learning process of humans. The experimental results from three different empirical studies, on more than 600 engineering students from the University of Genoa, showed that HAS (i) can be measured without the assumptions required by HRC, (ii) depends not only on the knowledge domain, as HRC, but also on the complexity of the problem, and (iii) can be exploited for better understanding of the human learning process.

1. Introduction

Since the emergence of Technology-Enhanced Learning (TEL) systems and automatic analysis of educational data, many efforts have been carried out to enhance the learning experience [1,2]. For this reason, Learning Analytics (LA) and Educational Data Mining have recently gained a lot of attention as one of their major interest is to explore the way humans learn [3–5]. New advances in LA enable measuring, collecting and analyzing data about learners and their contexts and allow exploring the behavior of people while learning (e.g. through Machine Learning models), opening the door towards optimized and personalized education [6–9]. LA is a multi-disciplinary field which is tightly connected to Statistics and ML on one side and to Cognitive Science (COGS) and Pedagogy on the other side [10,11].

Machine Learning (ML) is a field of research which develops and studies algorithms that can learn from and make predictions on data [12]. Such algorithms, used as tools in LA, build models from data in order to make data-driven predictions or decisions. ML offers tools for solving many real world problems [13–16]: classification, regression, clustering, online learning, semi-supervised learning, reinforcement learning, etc. [12,17–19]. According to [20] there are three ways of using models of educational processes: the first one includes models used as scientific tools to understand an educational situation, such as using models to predict the student academic success [21]. In the second one, models are used as a component of educational artefacts such as student modeling and its application in a TEL environment [22,23] or integrating the model of student problem-solving into a TEL system with the aim to personalize and adapt educational materials to their needs [24]. Finally, the third one includes models used as basis for design of TEL systems [25].

In addition to proposing new algorithms and tools, ML develops different methods for measuring the effectiveness of a learning process. In particular, ML studies the learning ability of an algorithm in order to avoid data memorization and to improve its generalization performance, which is the ability to learn the targeted concept effectively [26]. Examples of techniques for assessing the performance of a learning algorithm are: Hypothesis Space-based methods [27] (based on the VC-Dimension [26], Rademacher Complexity (RC) [28–30], and PAC Bayes Theory [31,32]) and Algorithm-based methods [33] (based on Compression Bounds [34], and Algorithmic Stability (AS) Theory [35,36]). Thanks to these approaches, many valuable parameters that describe how a particular machine learns can be quantified. For

* Corresponding author.
E-mail addresses: mehrnoosh.vahdat@edu.unige.it (M. Vahdat), luca.oneto@unige.it (L. Oneto), davide.anguita@unige.it (D. Anguita), m.funk@tue.nl (M. Funk), g.w.m.rauterberg@tue.nl (M. Rauterberg).

http://dx.doi.org/10.1016/j.neucom.2015.11.100
0925-2312/

example, it is possible to rigorously measure the generalization performance of a learned model.

While ML studies learning algorithms, COGS studies and analyses how learning takes place in humans [37–39]. In this context, humans can be considered as information processing systems (as suggested in [40]) with a high learning potential and learning is a permanent process that is regulated by optimizing the complexity of the learning context, based on actions and mental schemata of humans [41]. Concept Learning (CL) is the area of COGS that explores how concepts are attained in humans (Human Learning – HL). Various approaches exist in categorising concepts and how they are attained [42,43]. One approach is to consider concepts as mental representations which help to identify and separate objects, events, and relationships. Another approach considers that concepts are learned inductively even from sparse and noisy evidences. In addition, concepts can be formed by combining other simpler concepts, and their meanings are derived from the ones of their constituents. Various theories integrate different approaches of CL: for instance, Exemplar Theory [44] suggests that the categorization takes place by the proximity of the new stimulus to the members of the category that one has observed, and by comparison of similarities, the label is assigned to the stimulus. Another theory, called Prototype Theory [45], explains that categorization takes place in a similar way as Exemplar Theory, while the comparison is carried out to the average of category members not to a specific member. In this case, at first, the attributes of members of a category are derived (named prototypes), then categorization is done by considering the similarity to the generated prototypes. In addition to these theories, researchers discovered that rule-based theories are important in the initial formation of categories [37,46]: first the distinguishing attributes of new items are extracted from the category, then Examplar or Prototype theory, for categorizing distinct items, is applied. In this approach, concepts are constructed by combination [47]. In particular, a concept is represented by some rule that determines whether a stimulus belongs to a category [48]. Thus, humans try to find a rule (learn a model) when being confronted with a new example.

The latest approach towards human rule-based learning has been a motivation for CL to benefit from the research of other fields like Artificial Intelligence, Information Theory, and ML. In this context, the cross between HL and ML [49–51] leads to development of sophisticated formal models of CL [48,52–54]. For instance, [55] measures the ability of humans in Category Learning by applying Bayesian approaches in iterative learning. In this context, a human learns a concept and produces a hypothesis on the given data, then, another human learns the previously developed hypothesis and generates a new one. This method was adopted for identifying the inductive biases in humans. In another study [56], the difficulty of concepts in relation with HL is exploited. In this context, the subjective difficulty of boolean concepts for humans is measured, and it is shown that the subjective difficulty is proportional to the complexity of boolean statements (length of the statement). Thus, by knowing the complexity of the logical structure of concepts, it is possible to predict how difficult that concept is for humans. Other examples are [57,58] where ML Theory, which helps to understand the learning ability of ML algorithms, has been used to explore HL.

Our main contribution is to build a connection between ML and HL. In particular, we apply ML methods to measure the capacity of students to find meaningful rules given various problems. Measuring the ability of a human to capture information rather than simply memorizing can be the key to optimize and improve HL. In this sense, the parallelism with ML is straightforward: for example, several approaches in the last decades dealt with the development of measures to assess the generalization ability of learning algorithms in order to minimize risks of overfitting (memorization). As a consequence, merging ML studies on the generalization ability estimation and HL has been proposed by some researchers. In particular, Zhu et al. [57] propose the application of ML approaches [59] to estimate the human capability of extracting knowledge (Human Rademacher Complexity – HRC). Unfortunately, (H)RC requires a set of models to be aprioristically defined, which includes the models to be explored by the learner (being either an algorithm or a human) [33]. While this hypothesis is not always satisfied by ML methods (e.g. k-Nearest Neighbors [60]), aprioristically defining a list of alternative models for humans is an even tougher task [61,62]. This leads to formulating further assumptions [57], which do not often hold in practice. As an alternative, we propose to exploit AS [35,33] in order to compute the Human Algorithmic Stability (HAS), which does not rely on the definition of a set of models and does not require any additional assumptions. In this study, we comparatively benchmark HRC and HAS, by designing experiments to analyze the way a group of students learns the tasks with different difficulties, and we compare the two approaches to verify which one is the most informative for getting more insights into HL. To reach this purpose, three different experiments were performed from October 2014 till May 2015 with 606 students of various engineering majors from the University of Genoa, Italy. We generated unique questionnaires for every student to measure HRC and HAS over 7 groups of students, as described in the next sections. Filled questionnaires were collected, digitized, anonymized, and analyzed. Our results show that HAS is influenced by the nature and the complexity of the problem to learn. Moreover, contrarily to HRC, HAS is also able to capture the fast-learning ability of a human when dealing with simple problems: this allows providing new perspectives with reference to the human tendency to overfit training data depending on the nature of the problem faced. These results can thus function as a bridge between ML and HL, for the measure of the propensity of the learner towards CL versus simple memorization. This work completes and extends the preliminary results reported in [58].

The paper is structured as follows: Section 2 presents the theoretical ML framework, Section 3 relates the ML framework to HL, Section 4 describes our experimental design, Section 5 reports the results of our study and finally the conclusions of the paper are drawn in Section 6.

2. Rademacher Complexity and algorithmic stability in machine learning

Let us consider the classical binary classification framework [26]. Let $\mathcal{X}$ and $\mathcal{Y}=\{\pm 1\}$ be, respectively, an input and an output space. We consider a set of labeled independent and identically distributed (i.i.d.) data $\mathcal{S}_n:\{Z_1,\ldots,Z_n\}$ of size n, where $Z_{i\in\{1,\ldots,n\}}=(X_i,Y_i)$, with $X_i\in\mathcal{X}$ and $Y_i\in\mathcal{Y}$, sampled from an unknown distribution μ over $\mathcal{X}\times\mathcal{Y}$. We also define two modified training sets: $\mathcal{S}_n^{\setminus i}$, where the i-th element is removed and $\mathcal{S}_n^i$, where the i-th element is replaced with Z_i', which is another i.i.d. pattern sampled from μ:

$$\mathcal{S}_n^{\setminus i}:\{Z_1,\ldots,Z_{i-1},Z_{i+1},\ldots,Z_n\},\quad \mathcal{S}_n^i:\{Z_1,\ldots,Z_{i-1},Z_i',Z_{i+1},\ldots,Z_n\}. \tag{1}$$

A learning algorithm $\mathcal{A}$ maps $\mathcal{S}_n$ into a function $f:\mathcal{A}_{\mathcal{S}_n}$ from $\mathcal{X}$ to $\mathcal{Y}$. In particular, $\mathcal{A}$ allows designing $f\in\mathcal{F}$ and defining the hypothesis space $\mathcal{F}$, which is generally unknown.

Even if often not specified [35,33], there are some properties that the algorithm $\mathcal{A}$ must satisfy in order to ensure the validity of the results of the next sections. In particular, we consider only deterministic algorithms. It is also assumed that the algorithm $\mathcal{A}$ is symmetric with respect to $\mathcal{S}_n$, i.e. it does not depend on the order of the elements in the training set.

The accuracy of $\mathcal{A}_{\mathcal{S}_n}$ in representing the hidden relationship μ is measured with reference to a loss function $\ell(\mathcal{A}_{\mathcal{S}_n}, Z) : \mathcal{F} \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$. In particular, since we are dealing with binary classification problems, we use the hard loss function:

$$\ell(\mathcal{A}_{\mathcal{S}_n}, Z) = \frac{1 - Yf(X)}{2} \in \{0, 1\}, \tag{2}$$

which counts the number of misclassified examples. The quantity of interest is defined as the generalization error, namely the error that a model will perform on new data generated by μ and previously unseen:

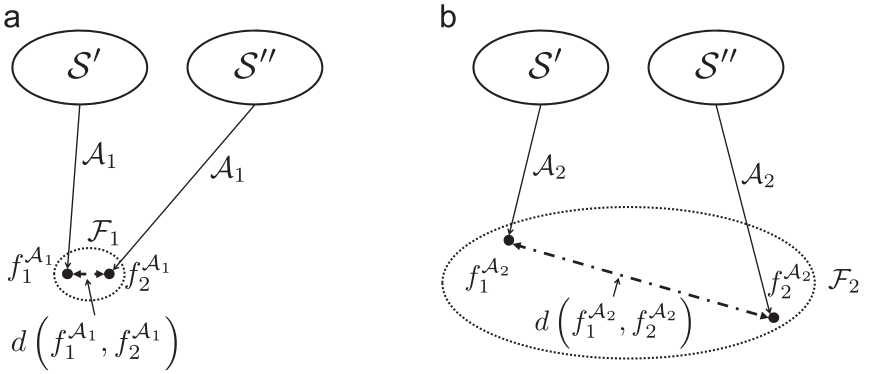
$$R(\mathcal{A}_{\mathcal{S}_n}) = \mathbb{E}_Z \ell(\mathcal{A}_{\mathcal{S}_n}, Z). \tag{3}$$

Unfortunately since μ is unknown, the random variable $R(\mathcal{A}_{\mathcal{S}_n})$ cannot be computed and, consequently, must be estimated. Two common empirical estimators are the Empirical ($\widehat{R}_{\mathrm{EMP}}(\mathcal{A}_{\mathcal{S}_n})$) and Leave-One-Out ($\widehat{R}_{\mathrm{LOO}}(\mathcal{A}_{\mathcal{S}_n})$) errors:

$$\widehat{R}_{\mathrm{EMP}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n) = \frac{1}{n} \sum_{Z \in \mathcal{S}_n} \ell(\mathcal{A}_{\mathcal{S}_n}, Z), \quad \widehat{R}_{\mathrm{LOO}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n) = \frac{1}{n} \sum_{i=1}^{n} \ell(\mathcal{A}_{\mathcal{S}_n^{\setminus i}}, Z_i). \tag{4}$$

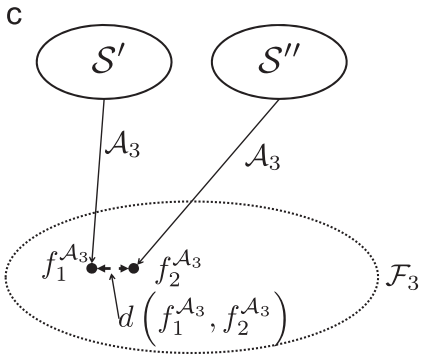
In order to estimate the generalization error, given one of its empirical estimators, we make use of two powerful statistical measures: RC [63,59,27,28,64,65] and AS [35,36,66,33]. The difference between the two approaches can be clarified through the graphical representation, proposed in Fig. 1. Basically, RC allows estimating the size of a class of functions. Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be two algorithms that chooses the model, respectively, in the classes $\mathcal{F}_1$ and $\mathcal{F}_2$, and $\mathcal{S}'$ and $\mathcal{S}''$ be two different training sets, originated from the same distribution μ. The learning phase consists of finding a model from the selected hypothesis space, say $\mathcal{F}_1$, which best fits the data: if $\mathcal{S}'$ is used, then $f_1^{\mathcal{A}_1}$ is obtained, while $f_2^{\mathcal{A}_1}$ is selected if we opt to learn the dataset $\mathcal{S}''$, instead. Since, in this case, the hypothesis space $\mathcal{F}_1$ is small (namely simple) enough, $f_1^{\mathcal{F}_1}$ will be forced to be "close" (with respect to some distance $d(f_1^{\mathcal{A}_1}, f_2^{\mathcal{A}_1})$) to $f_2^{\mathcal{F}_1}$. In other words, the final outcome of the learning phase will not be heavily influenced by the randomness of the process generating the data, so the risk of learning noise, i.e. of overfitting the data, will be small (see Fig. 1(a)). If, instead, we use $\mathcal{A}_2$ which chooses functions from a larger hypothesis set $\mathcal{F}_2$, the model $f_1^{\mathcal{F}_2}$ could end up "far" from $f_2^{\mathcal{F}_2}$ (with respect to the distance $d(f_1^{\mathcal{A}_2}, f_2^{\mathcal{A}_2})$): this means that the hypothesis class is too large for a particular learning task and the risk of overfitting the data is high (see Fig. 1(b)). In practice, RC tries to estimate the largest $d(f_1^{\mathcal{A}}, f_2^{\mathcal{A}})$ given the hypothesis space $\mathcal{F}$ from which the algorithm chooses the model.

The advantage of using AS is that the hypothesis space $\mathcal{F}$ does not need to be known in advance. This means that, even if the algorithm chooses from a large hypothesis space, given a particular μ, the algorithm will choose models that are close to each other. A graphical description is shown in Fig. 1: let us consider, again, algorithms $\mathcal{A}_1$ (see Fig. 1(a)) and $\mathcal{A}_2$ (see Fig. 1(b)), and the two training sets $\mathcal{S}'$ and $\mathcal{S}''$. In this case, $d(f_1^{\mathcal{A}_2}, f_2^{\mathcal{A}_2}) \gg d(f_1^{\mathcal{A}_1}, f_2^{\mathcal{A}_1})$ consequently, we derive the same conclusion of the analysis performed with RC-based approach. Instead, if we apply a stable algorithm $\mathcal{A}_3$ (see Fig. 1(c)) on $\mathcal{S}'$ and $\mathcal{S}''$, we obtain that $d(f_1^{\mathcal{A}_3}, f_2^{\mathcal{A}_3})$ is small even if $\mathcal{F}_3$ is a large hypothesis space. Note that, since we are just looking at $d(f_1^{\mathcal{A}}, f_2^{\mathcal{A}})$, we do not need to explicitly know $\mathcal{F}$: the stability of a particular learning algorithm is simply computed by measuring $d(f_1^{\mathcal{A}}, f_2^{\mathcal{A}})$.



$\mathcal{F}_1$: RC and AS are small (considering the size of $\mathcal{F}_1$ and $d(f_1^{\mathscr{A}_1}, f_2^{\mathscr{A}_1})$).

$\mathcal{F}_2$: RC and AS are large (considering the size of $\mathcal{F}_2$ and $d(f_1^{\mathscr{A}_2}, f_2^{\mathscr{A}_2})$).



$\mathcal{F}_3$: RC is large but AS is small (considering the size of $\mathcal{F}_3$ and $d(f_1^{\mathscr{A}_3}, f_2^{\mathscr{A}_3})$).

Fig. 1. Different approaches to learning. $\mathcal{A}_1$ and $\mathcal{A}_2$ are two algorithms that chooses the model from the classes $\mathcal{F}_1$ and $\mathcal{F}_2$. $\mathcal{S}'$ and $\mathcal{S}''$ are two different training sets, originated from the same distribution μ. $d(f_1^{\mathcal{A}_{1,2,3}}, f_2^{\mathcal{A}_{1,2,3}})$ shows the distance between the chosen models during the two different learning processes.

In the next sections, we recall some results that are useful for the remaining part of the paper. Some of the theoretical results can be retrieved in previous works [59,27,35,36,33], but the adaptation of these approaches to the context of this paper is entirely new.

2.1. Understanding the learning ability of an algorithm through Rademacher Complexity

As detailed in the previous section, the quantity of interest in any learning procedure is the generalization error $R(\mathcal{A}_{\mathcal{S}_n})$ which is not measurable since μ is unknown. We can however measure its empirical estimators which are optimistically biased estimators of $R(\mathcal{A}_{\mathcal{S}_n})$ [26,35]. In order to estimate this bias, we can use RC to bound Uniform Deviation which is the maximum distance between the generalization error and the empirical error [59,27]:

$$\widehat{U}_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} \left[R(f) + \widehat{R}_{\mathrm{EMP}}(f, \mathcal{S}_n) \right], \tag{5}$$

since

$$R(\mathcal{A}_{\mathcal{S}_n}) \leq \widehat{R}_{\mathrm{EMP}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n) + \widehat{U}_n(\mathcal{F}). \tag{6}$$

The expected value of $\widehat{U}_n(\mathcal{F})$ can be defined as:

$$U_n(\mathcal{F}) = \mathbb{E}_{\mathcal{S}_n} \widehat{U}_n(\mathcal{F}). \tag{7}$$

The RC of a class of functions $\mathcal{F}$ and its expected value are defined as:

$$\widehat{C}_n(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \ell(f, Z_i), \quad C_n(\mathcal{F}) = \mathbb{E}_{\mathcal{S}_n} \widehat{C}_n(\mathcal{F}), \tag{8}$$

where $\sigma_1, \ldots, \sigma_n$ are n independent random variables for which $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2$. Note that, since we use the hard loss function, it is possible to prove that:

$$\widehat{C}_n(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \ell(f, Z_i) = 1 - 2\mathbb{E}_{\boldsymbol{\sigma}} \inf_{f \in \mathcal{F}} \widehat{R}_{\mathrm{EMP}}(f, \mathcal{S}_n^{\boldsymbol{\sigma}}), \tag{9}$$

where $\mathcal{S}_n^{\boldsymbol{\sigma}} : \{Z_1^{\boldsymbol{\sigma}}, \ldots, Z_n^{\boldsymbol{\sigma}}\}$ of size n, and $Z_{i \in \{1,\ldots,n\}}^{\boldsymbol{\sigma}} = (X_i, \sigma_i)$. An upper bound of $R(\mathcal{A}_{\mathcal{S}_n})$ in terms of $\widehat{C}(\mathcal{F})$ was proposed in [59] and the proof consists mainly of an application of the McDiarmid's inequality [67]. Since both RC and Uniform Deviation are bounded difference functions [59,27] then:

$$C_n(\mathcal{F}) \leq \widehat{C}_n(\mathcal{F}) + \sqrt{\frac{2 \log\left(\frac{1}{\delta}\right)}{n}}, \quad \widehat{U}_n(\mathcal{L}) \leq U_n(\mathcal{F}) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}}. \tag{10}$$

The bounds hold with probability $(1-\delta)$, where δ is a user-defined level of confidence. Using these results, it is possible to bound $\widehat{U}_n(\mathcal{F})$, by noting that [59]:

$$U_n(\mathcal{F}) \leq C_n(\mathcal{F}). \tag{11}$$

By exploiting Eqs. (10) and (11), we obtain the following relation which holds with probability $(1-\delta)$ [59,27]:

$$R(\mathcal{A}_{\mathcal{S}_n}) \leq \widehat{R}_{\mathrm{EMP}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n) + \widehat{C}_n(\mathcal{F}) + 3\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}. \tag{12}$$

All the quantities involved in the bound of Eq. (12) can be empirically computed by exploiting the available observations $(\mathcal{S}_n)$. Note that $\widehat{C}_n(\mathcal{F})$ requires that an expectation over all σ is performed. This is obviously infeasible in practice, and a Monte Carlo estimation of the quantity can be computed instead [59,27]. A more refined alternative is to exploit a recent result [60,59], which shows that the following quantity can be used:

$$\widehat{\widehat{C}}_n(\mathcal{F}) = 1 - 2 \inf_{f \in \mathcal{F}} \widehat{R}_{\mathrm{EMP}}(f, \mathcal{S}_n^{\boldsymbol{\sigma}}) \tag{13}$$

that is RC computed with just a single draw of $\boldsymbol{\sigma}$. In fact, $\widehat{\widehat{C}}_n(\mathcal{F})$ is also a bounded difference function and consequently, the following bound holds with probability $(1-\delta)$:

$$C_n(\mathcal{F}) \leq \widehat{\widehat{C}}_n(\mathcal{F}) + \sqrt{\frac{8 \log\left(\frac{1}{\delta}\right)}{n}}. \tag{14}$$

By exploiting Eqs. (10), (11), and (14), we can prove that the following bound holds with probability $(1-\delta)$:

$$R(\mathcal{A}_{\mathcal{S}_n}) \leq \widehat{R}_{\mathrm{EMP}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n) + \widehat{\widehat{C}}_n(\mathcal{F}) + 5\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}. \tag{15}$$

The bound of Eq. (15) is slightly looser but it is more computational tractable respect to the one of Eq. (12). It is worth noting that all the constants in the bound can be improved [64] but this is out of the scope of this paper.

Unfortunately, RC requires the hypothesis space to be fixed before seeing the data and even functions that will never be picked up by the learning procedure are taken into account when estimating the Uniform Deviation. This could compromise our ability to understand the learning properties of our algorithm: if an effective algorithm chooses functions that belong to a complex class, the bound of Eq. (12) is loose [36,33] and so we cannot guarantee the performance of the algorithm.

2.2. Understanding the learning ability of an algorithm through algorithmic stability

Stability does not require a class of functions $\mathcal{F}$ to be defined a priori, since the set of models is implicitly derived by the algorithm $\mathcal{A}$ itself. For this reason, the simple deviation $\widehat{D}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n)$ of the generalization error from $\widehat{R}_{\mathrm{EMP}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n)$ or $\widehat{R}_{\mathrm{LOO}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n)$ is analyzed [35,66,36,33]:

$$\widehat{D}_{\mathrm{EMP}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n) = \left| R(\mathcal{A}_{\mathcal{S}_n}) - \widehat{R}_{\mathrm{EMP}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n) \right|, \tag{16}$$

$$\widehat{D}_{\mathrm{LOO}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n) = \left| R(\mathcal{A}_{\mathcal{S}_n}) - \widehat{R}_{\mathrm{LOO}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n) \right|. \tag{17}$$

Consider that the deterministic squared counterpart of $\widehat{D}_{\mathrm{EMP}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n)$ and $\widehat{D}_{\mathrm{LOO}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n)$ can be defined as:

$$D^2_{\mathrm{EMP}}(\mathcal{A}, n) = \mathbb{E}_{\mathcal{S}_n}[\widehat{D}_{\mathrm{EMP}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n)]^2, \tag{18}$$

$$D^2_{\mathrm{LOO}}(\mathcal{A}, n) = \mathbb{E}_{\mathcal{S}_n}[\widehat{D}_{\mathrm{LOO}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n)]^2. \tag{19}$$

In order to study $\widehat{D}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n)$, we can adopt different approaches. The first one consists of using Hypothesis Stability $H(\mathcal{A}, n)$:

$$H_{\mathrm{EMP}}(\mathcal{A}, n) = \mathbb{E}_{\mathcal{S}_n, Z_i} \left| \ell(\mathcal{A}_{\mathcal{S}_n}, Z_i) - \ell(\mathcal{A}_{\mathcal{S}_n^i}, Z_i) \right| \leq \beta_{\mathrm{EMP}}, \tag{20}$$

$$H_{\mathrm{LOO}}(\mathcal{A}, n) = \mathbb{E}_{\mathcal{S}_n, Z} \left| \ell(\mathcal{A}_{\mathcal{S}_n}, Z) - \ell(\mathcal{A}_{\mathcal{S}_n^{\setminus i}}, Z) \right| \leq \beta_{\mathrm{LOO}}. \tag{21}$$

Lemma 3 in [35] proves that:

$$D^2_{\mathrm{EMP}}(\mathcal{A}, n) \leq \frac{1}{2n} + 3H_{\mathrm{EMP}}(\mathcal{A}, n), \quad D^2_{\mathrm{LOO}}(\mathcal{A}, n) \leq \frac{1}{2n} + 3H_{\mathrm{LOO}}(\mathcal{A}, n). \tag{22}$$

By exploiting the Chebyshev inequality [68], for a random variable a, with probability $(1-\delta)$ we obtain:

$$\mathbb{P}[a > t] \leq \mathbb{E}[a^2]/t^2, \quad a < \sqrt{\mathbb{E}[a^2]/\delta}. \tag{23}$$

Then, by combining Eqs. (16), (20), and (22) (or analogously, Eqs. (17), (21), and (22)) with probability $(1-\delta)$ we obtain that:

$$R(\mathcal{A}_{\mathcal{S}_n}) \leq \widehat{R}_{\mathrm{EMP}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n) + \sqrt{\frac{1}{2n\delta} + \frac{3\beta_{\mathrm{EMP}}}{\delta}}, \tag{24}$$

$$R(\mathcal{A}_{\mathcal{S}_n}) \leq \widehat{R}_{\mathrm{LOO}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n) + \sqrt{\frac{1}{2n\delta} + \frac{3\beta_{\mathrm{LOO}}}{\delta}}. \tag{25}$$

In Theorem 11 of [35], it is proved that $H_{\mathrm{EMP}}(\mathcal{A}, n) \leq 2H_{\mathrm{LOO}}(\mathcal{A}, n)$ but unfortunately, as remarked in [66], this proof contains an error and consequently Leave-One-Out Hypothesis Stability does not imply $H_{\mathrm{EMP}}(\mathcal{A}, n)$ Stability. Moreover, in [33] it is proved that $H_{\mathrm{LOO}}(\mathcal{A}, n)$ can be estimated from the data, while this is not possible for $H_{\mathrm{EMP}}(\mathcal{A}, n)$. Consequently, from now on, we only deal with Leave-One-Out Hypothesis Stability since, as described in the rest of this section, it is the only one which leads to a fully empirical stability-based bound.

In order to estimate $H_{\mathrm{LOO}}(\mathcal{A}, n)$ from the data, we need to suppose that for the algorithm under exam ($\mathcal{A}$), Hypothesis Stability does not increase with the cardinality of the training set:

$$H_{\mathrm{LOO}}(\mathcal{A}, n) \leq H_{\mathrm{LOO}}(\mathcal{A}, \sqrt{n}/2). \tag{26}$$

We point out that Property (26) is a desirable requirement for any learning algorithm: in fact, the impact on the learning procedure of removing samples from $\mathcal{S}_n$ should decrease on average, as n grows. Note that Property (26) has already been studied by many researchers in the past. In particular, Property (26) is related to the notion of consistency [69,70] and connections can also be identified with the trend of the learning curves of an algorithm [71–74]. Moreover, such quantities are strictly linked to the concept of Smart Rule [69]. It is worth underlining that, in the above-referenced works, Property (26) is proved to be satisfied by many well-known algorithms (Support Vector Machines, Kernelized Regularized Least Squares, k-Local Rules with $k > 1$).

In order to derive a fully empirical bound we have to study $H_{\mathrm{LOO}}(\mathcal{A}, \sqrt{n}/2)$. For this purpose, the following empirical quantity can be introduced [33]:

$$\widehat{H}_{\mathrm{LOO}}(\mathcal{A}, \sqrt{n}/2, \mathcal{S}_n) = \frac{8}{n\sqrt{n}} \sum_{k=1}^{\sqrt{n}/2} \sum_{j=1}^{\sqrt{n}/2} \sum_{i=1}^{\sqrt{n}/2} \left| \ell\left(\mathcal{A}_{\check{\mathcal{S}}^k_{\sqrt{n}/2}}, \check{Z}^k_j\right) - \ell\left(\mathcal{A}_{(\check{\mathcal{S}}^k_{\sqrt{n}/2})^{\setminus i}}, \check{Z}^k_j\right) \right|, \tag{27}$$

where $\forall k \in \{1, \ldots, \sqrt{n}/2\}$:

$$\check{\mathcal{S}}^k_{\sqrt{n}/2} : \left\{Z_{(k-1)\sqrt{n}+1}, \ldots, Z_{(k-1)\sqrt{n}+\sqrt{n}/2}\right\}, \quad \check{Z}^k_j : Z_{(k-1)\sqrt{n}+\sqrt{n}/2+j}. \tag{28}$$

Note that the quantity of Eq. (27) is the empirical unbiased estimator of $H_{\mathrm{LOO}}(\mathcal{A}, \sqrt{n}/2)$ and then:

$$H_{\mathrm{LOO}}(\mathcal{A}, \sqrt{n}/2) = \mathbb{E}_{\mathcal{S}_n} \widehat{H}_{\mathrm{LOO}}(\mathcal{A}, \sqrt{n}/2, \mathcal{S}_n). \tag{29}$$

Consequently, with probability $(1-\delta)$, the difference between $H_{\mathrm{LOO}}(\mathcal{A}, \sqrt{n}/2)$ and $\widehat{H}_{\mathrm{LOO}}(\mathcal{A}, \sqrt{n}/2, \mathcal{S}_n)$ can be bounded by exploiting, for example, the Hoeffding inequality [75,33]:

$$H_{\mathrm{LOO}}(\mathcal{A}, \sqrt{n}/2) \leq \widehat{H}_{\mathrm{LOO}}(\mathcal{A}, \sqrt{n}/2, \mathcal{S}_n) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{\sqrt{n}}}. \tag{30}$$

Combining Eqs. (25) and (30), the following stability bound, holding with probability $(1-\delta)$, can be derived:

$$R(\mathcal{A}_{\mathcal{S}_n}) \leq \widehat{R}_{\mathrm{LOO}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n) + \sqrt{\frac{2}{\delta}\left[\frac{1}{2n} + 3\left(\widehat{H}_{\mathrm{LOO}}(\mathcal{A}, \sqrt{n}/2, \mathcal{S}_n) + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{\sqrt{n}}}\right)\right]}. \tag{31}$$

The bound of Eq. (31) takes into account only empirical quantities and, in order to compute them, we do not need to know $\mathcal{F}$ but we need to apply the algorithm $\mathcal{A}$ on a series of modified training sets that are all built using $\mathcal{S}_n$. Moreover, only the functions that, given a set of data, could be actually learned by $\mathcal{A}$ are contemplated, differently from RC. When stability bounds are used, we only need to prove that Property (26) holds for the chosen algorithm.

3. From machine learning to human learning

In this section we show how to measure on humans the three different quantities involved in learning as described in Section 2: the generalization error (that we call Human Error), HRC, and HAS. We describe how these quantities should behave from a ML point of view, and through ad-hoc experiments on humans, we check whether these behaviors can be observed for HL as well.

In order to measure the three above-mentioned quantities, it is important to consider that a human is fundamentally different from a ML algorithm in the sense that humans have memory while ML algorithms do not. In other words, we cannot state that humans are deterministic in the sense that, if we give the same problem P to a human at time t and at time $t+\Delta$, the selected models, respectively, M^t and $M^{t+\Delta}$, could be different ($M^t \neq M^{t+\Delta}$). Moreover, if a human is given a problem $P1$ at time t, and another problem $P2$ at time $t+\Delta$, the selected models, respectively, M^t_{P1} and $M^{t+\Delta}_{P2}$, could be different from the one selected by the same person given the problem $P2$ at time t and problem $P1$ at time $t+\Delta$, respectively, M^t_{P2} and $M^{t+\Delta}_{P1}$. In other words, $M^t_{P2} \neq M^{t+\Delta}_{P2}$ and $M^t_{P1} \neq M^{t+\Delta}_{P1}$. Note that this problem is also underlined in [57,58]. These issues, as are shown in the next section, prevent us to measure, for example, $\widehat{C}_n(\mathcal{F})$, $\widehat{R}_{\mathrm{LOO}}(\mathcal{A}_{\mathcal{S}_n}, \mathcal{S}_n)$, and $\widehat{H}_{\mathrm{LOO}}(\mathcal{A}, \sqrt{n}/2, \mathcal{S}_n)$ for a human. This is due to the fact that, contrary to a ML algorithm, a human cannot be interpellated as many times as we need, because we risk to falsify the results. Unfortunately, for measuring the quantity of interest for a single human we should ask her/him to solve several slightly modified instances of the problem as described in Section 2. For this reason, we propose to average these quantities over different humans as in [57,58].

In order to show how to measure Human Error, HRC, and HAS, first we need to introduce some additional quantities. In particular, we consider m sets $\mathcal{D}^{\mathrm{LEARN}}_m : \{\mathcal{L}_{1,n}, \ldots, \mathcal{L}_{m,n}\}$ of n labeled i.i.d. data and other m sets $\mathcal{D}^{\mathrm{TEST}}_m : \{\mathcal{T}_{1,p}, \ldots, \mathcal{T}_{m,p}\}$ of p labeled i.i.d. data all sampled from μ. Moreover, we consider a group of $2m$ humans $\mathcal{G}^{2m} : \{\mathcal{H}^1, \ldots, \mathcal{H}^{2m}\}$ and each $\mathcal{H}^{i \in \{1,\ldots,2m\}}$ chooses functions in the unknown space $\mathcal{F}^{i \in \{1,\ldots,2m\}}$.

3.1. Human Error

The first quantity of interest in learning, as described in Section 2, is the generalization error of an algorithm $\mathcal{A}$. In this case, the algorithm $\mathcal{A}$ is the human $\mathcal{H}$ that takes some sample of the distribution μ, which basically is a task to learn, and returns a model f in an unknown $\mathcal{F}$. Obviously we cannot explicitly access f but we can ask the human $\mathcal{H}$ to label some previously unseen unlabeled samples, and check whether her/his answer is in agreement with the true label of the sample. Here, we are interested in measuring the expected Human Error:

$$R^{\mu}_n = \mathbb{E}_{\mathcal{H}, \mathcal{S}_n, Z} \ell(\mathcal{H}_{\mathcal{S}_n}, Z). \tag{32}$$

In other words, R^{μ}_n is the average ability of the human to learn a binary classification task μ, given n i.i.d. samples $\mathcal{S}_n$ sampled from μ. Since we do not have all the humans but just a finite group of them $\mathcal{G}^{2m}$, and since we cannot give them many times different samples to learn coming from the same μ (they would remember the samples and this leads to compromising the results), we can estimate R^{μ}_n thanks to $\mathcal{D}^{\mathrm{LEARN}}_m$ and $\mathcal{D}^{\mathrm{TEST}}_m$:

$$\widehat{R}^{\mu}_n = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{p} \sum_{Z \in \mathcal{T}_{i,p}} \ell(\mathcal{H}^i_{\mathcal{L}_{i,n}}, Z). \tag{33}$$

Note that, $\widehat{R}_n^{\mu}$ is an unbiased estimator of R_n^{μ} and in particular we can use the Hoeffding inequality to state that the following bound holds, with probability $(1-\delta)$:

$$R_n^{\mu} \leq \widehat{R}_n^{\mu} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}. \tag{34}$$

Consider that the accuracy in estimating R_n^{μ} through $\widehat{R}_n^{\mu}$ depends only on m. We decided to add the parameter p (more data to label for each human) since, as can be seen later in the experimental section, we discovered that this addition leads to a more stable estimate.

Based on these definitions, one can investigate how R_n^{μ} varies by changing task (μ) or the number of samples n available for learning. From a ML perspective, it is expected that R_n^{μ} decreases with n for consistency reasons: the more samples we use for learning, the smaller error we will obtain on previously unseen data. If this behavior is not observed, it would show that the samples have been memorized rather than learned [26,69,73,70,74].

3.2. Human Rademacher Complexity

For what concerns HRC, the average RC of a human needs to be estimated:

$$C_n(\mathcal{F}) = \mathbb{E}_{\mathcal{F},\mathcal{S}_n,\boldsymbol{\sigma}}\left[1 - 2\inf_{f\in\mathcal{F}} \widehat{R}_{\mathrm{EMP}}(f, \mathcal{S}_n^{\boldsymbol{\sigma}})\right]. \tag{35}$$

Note that in this case, $C_n(\mathcal{F})$ cannot be computed since the space $\mathcal{F}$ is unknown. For this reason, an assumption is made: every human is always able to choose the best model of which she/he is capable. Based on this assumption, it is possible to measure the infimum in Eq. (35). Unfortunately, this assumption does not hold in practice because the process of learning of a human is not equivalent to a simple minimization process. Instead, learning may be viewed as a complex process, rather than a collection of factual and procedural knowledge [61].

Based on this assumption, we can reformulate $C_n(\mathcal{F})$ in the following way:

$$C_n^{\mu} = \mathbb{E}_{\mathcal{H},\mathcal{S}_n,\boldsymbol{\sigma}}\left[1 - 2\widehat{R}_{\mathrm{EMP}}(\mathcal{H}_{\mathcal{S}_n^{\boldsymbol{\sigma}}}, \mathcal{S}_n^{\boldsymbol{\sigma}})\right]. \tag{36}$$

Consider that C_n^{μ} (as $C_n(\mathcal{F})$) does not depend on $\mathbb{P}\{Y|X\}$ but just on $\mathbb{P}\{X\}$, since when computing C_n^{μ} the labels are disregarded (see Section 2.1 for details) [76]. In other words, many μ with the same $\mathbb{P}\{X\}$, but different $\mathbb{P}\{Y|X\}$, have the same C_n^{μ}.

Unfortunately, C_n^{μ} cannot be computed for the same reason that R_n^{μ} cannot be computed (see Section 3.1), so we can estimate C_n^{μ} by using $\mathcal{G}^{2m}$ and $\mathcal{D}_m^{\mathrm{LEARN}}$ with the following empirical estimator:

$$\widehat{C}_n^{\mu} = \frac{1}{m}\sum_{i=1}^{m}\left[1 - 2\widehat{R}_{\mathrm{EMP}}(\mathcal{H}^i_{\mathcal{L}^{\boldsymbol{\sigma}^i}_{i,n}}, \mathcal{L}^{\boldsymbol{\sigma}^i}_{i,n})\right]. \tag{37}$$

Note that $\widehat{C}_n^{\mu}$ is an unbiased estimator of C_n^{μ} and in particular, we can use the Hoeffding inequality to state that the following bound holds, with probability $(1-\delta)$:

$$C_n^{\mu} \leq \widehat{C}_n^{\mu} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}. \tag{38}$$

Based on these definitions, one can investigate how C_n^{μ} varies by changing task (μ) or the number of samples n available for learning. From the ML perspective, we expect that RC decreases when n increases. The more data we need to learn, the less $\mathcal{F}$ is able to fit random noise as $\mathbb{P}\{Y|X\}$. If C_n^{μ} does not decrease with n, it shows that the human is not learning but memorizing the information. In other words, $\mathcal{F}$ is too large to be able to learn a particular task [28,59,76,57]. Note that C_n^{μ} is just a possible indicator of the ability of the algorithm to learn the task. There are cases where, even if it is not possible to prove through RC that an algorithm is learning the task, indeed the algorithm is effectively learning [36] or it is learning at much faster rate than we can prove [29,77].

3.3. Human algorithmic stability

The last quantity to estimate is the average Human Leave-One-Out Hypothesis Stability. In particular, we are interested in the quantity of Eq. (21). As explained in Section 2.2, stability does not require the knowledge of the $\mathcal{F}$ from which the humans will choose the model, but it requires that the humans solve many instances of a particular problem. Consequently, the average Human Leave-One-Out Hypothesis Stability can be expressed as:

$$\begin{aligned} H_n^{\mu} &= \mathbb{E}_{\mathcal{H},\mathcal{S}_n,Z}\left|\ell(\mathcal{H}_{\mathcal{S}_n}, Z) - \ell(\mathcal{H}_{\mathcal{S}_n^{\setminus i\in\{1,\ldots,n\}}}, Z)\right| \\ &= \mathbb{E}_{\mathcal{H},\mathcal{S}_n,Z}\left|\ell(\mathcal{H}_{\mathcal{S}_n}, Z) - \ell(\mathcal{H}_{\mathcal{S}_n^{\setminus 1}}, Z)\right|, \end{aligned} \tag{39}$$

Differently from RC, H_n^{μ} depends both on $\mathbb{P}\{X\}$ and $\mathbb{P}\{Y|X\}$. In other words, each μ has, in general, a different H_n^{μ} even if $\mathbb{P}\{X\}$ is the same.

H_n^{μ} cannot be computed for the same reason we cannot compute R_n^{μ} and C_n^{μ}, so we can estimate it by using $\mathcal{G}^{2m}$, $\mathcal{D}_m^{\mathrm{LEARN}}$, and $\mathcal{D}_m^{\mathrm{TEST}}$, and the following unbiased empirical estimator:

$$\widehat{H}_n^{\mu} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{p}\sum_{Z\in\mathcal{T}_{i,p}}\left|\ell\left(\mathcal{H}^i_{\mathcal{L}_{i,n}}, Z\right) - \ell\left(\mathcal{H}^i_{\mathcal{L}^{\setminus i}_{i,n}}, Z\right)\right| \tag{40}$$

By using the Hoeffding inequality again, we can state that the following bound holds, with probability $(1-\delta)$:

$$H_n^{\mu} \leq \widehat{H}_n^{\mu} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}. \tag{41}$$

Note that, as for $\widehat{R}_n^{\mu}$, even if the accuracy of our estimate depends only on m, we introduce p for making $\widehat{H}_n^{\mu}$ a more stable estimator of H_n^{μ} as can be seen later in the experimental section.

A problem in computing H_n^{μ} and $\widehat{H}_n^{\mu}$ is that they do not take into account the ability of humans to memorize. In particular, it is difficult that people will change their mind once they learned a concept [61]. This is a problem when measuring $|\ell(\mathcal{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathcal{H}^i_{\mathcal{L}^{\setminus i}_{i,n}}, Z)|$, since we need to provide the same data-set to the human with one sample removed in two different moments in time. This produces a bias in the estimation, as described in the beginning of Section 3: in fact, our experiments (see Section 5) will show that this bias is noticeable. A solution to this problem is to measure the average cross-human stability instead of the average self-human stability. In other words, we can consider the stability as a property that connects different humans instead of the property of a single human (analogously to the size of the space of function for RC). Consequently, we can measure how stable two humans are, one with respect to the other, averaged over a group of people: this interpretation of stability measures the stability of a group of people while learning a task μ. Based on these considerations, we reformulate H_n^{μ} in order to measure the following notion of stability:

$$\begin{aligned} \overline{H}_n^{\mu} &= \mathbb{E}_{\mathcal{H}',\mathcal{H}'',\mathcal{S}_n,Z}\left|\ell(\mathcal{H}'_{\mathcal{S}_n}, Z) - \ell(\mathcal{H}''_{\mathcal{S}_n^{\setminus i\in\{1,\ldots,n\}}}, Z)\right| \\ &= \mathbb{E}_{\mathcal{H}',\mathcal{H}'',\mathcal{S}_n,Z}\left|\ell(\mathcal{H}'_{\mathcal{S}_n}, Z) - \ell(\mathcal{H}''_{\mathcal{S}_n^{\setminus 1}}, Z)\right|, \end{aligned} \tag{42}$$

Analogously to H_n^{μ}, $\overline{H}_n^{\mu}$ can be estimated by using $\mathcal{G}^{2m}$, $\mathcal{D}_m^{\mathrm{LEARN}}$, and

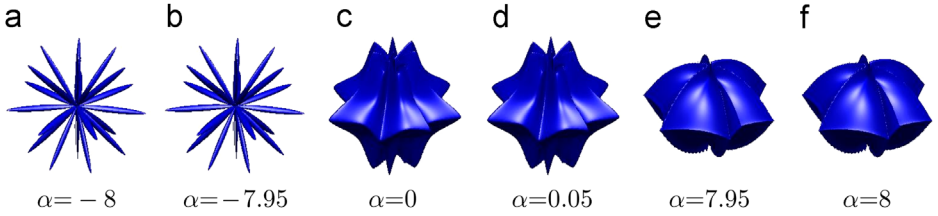


Fig. 2. Samples from Shape domain [57]. The computer-generated shapes are parametrized by $\alpha \in [-8, +8]$ from spiky shapes to smooth ones.

$\mathcal{D}_m^{\text{TEST}}$ with the following unbiased empirical estimator:

$$\widehat{\overline{H}}_n^{\mu} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{p}\sum_{Z\in\mathcal{T}_{i,p}}\left|\ell\left(\mathcal{H}_{\mathcal{L}_{i,n}^{i},Z}\right)-\ell\left(\mathcal{H}_{\mathcal{L}_{i,n}^{\setminus i}}^{i+m},Z\right)\right| \tag{43}$$

Knowing that with probability $(1-\delta)$:

$$\overline{H}_n^{\mu} \le \widehat{\overline{H}}_n^{\mu} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}. \tag{44}$$

Based on these definitions, we investigate how $\overline{H}_n^{\mu}$ and H_n^{μ} vary by changing task (μ) or the number of samples (n) available for learning. From a ML perspective, we expect that stability will decreases as n increases: the more data is available for learning, the less impact there is in removing one single example from the data. If $\overline{H}_n^{\mu}$ and H_n^{μ} do not decrease with n, the reason would be that the machine is not learning and there are not enough examples available to retrieve the information hidden in one example from the others. In other words, $\mathcal{H}$ is not able to find a stable solution or is not able to learn the particular task effectively [35,36,66,33,58].

4. Experimental design

A previous work [57] reports on an experiment targeted towards measuring HRC, while [58] is our first attempt to measure HAS, where we reported the preliminary results. The current work extends and completes both studies [57,58]. Given the drawbacks of RC with respect to AS, as highlighted in the previous sections, we built on the previous efforts and experiments to design and carry out a new experiment, which aims at estimating the average HRC and HAS. Our objective is to compare these two quantities and identify which one is the most informative for getting more insights into HL.

We performed three different experiments EX[1] (a pilot experiment), EX[2], and EX[3]. EX[1] was performed in October 2014, EX[2] in November 2014, and EX[3] in April/May 2015. EX[2] was designed based on the observed results of EX[1] while EX[3] extends EX[2]. The experiments were carried out with a total of 606 students from the University of Genoa (Italy): 70 students of Bioinformatics Engineering participated in EX[1]; EX[2] was carried out with 307 undergraduate students of Electronic Engineering (68 students), Bioinformatics Engineering (136 students), and Computer Engineering (103 students); finally EX[3] was carried out with 229 students of Bioinformatics Engineering (63 students), Electrical Engineering (53 students), Computer Engineering (33 students), and Mechanical Engineering (80 students). Each student participated only one time in the study and was given a unique automatically generated questionnaire. Experiments were carried out with 7 groups of students carefully controlled by the professors and researchers involved in the study together with the help of university employees. Filled questionnaires were collected, digitized, anonymized, and analyzed. We describe the details of the questionnaire design in the next sections.

4.1. *Experimental design:* EX[1]

In this phase, our experiment involved two types of statistical measures: HRC and HAS. For HRC in particular, we followed the approach designed in [57]. Then we designed a new experiment in order to measure HAS. We use the word "task" (or "problem") and "domain" to refer, respectively, to $\mathbb{P}\{Y|X\}$ (or μ) and $\mathbb{P}\{X\}$. The first step consisted of defining the task to be learned by students. Two domains were defined: Shape and Word. Two problems were defined for each domain: simple (linear) and difficult (non-linear).[1] We report the detailed description of these problems as follows.

4.1.1. *Shape domain*

The Shape domain consists of 321 computer-generated 3D shapes, parametrized by $\alpha \in [-8, +8]$, such that a small value of α leads to spiky shapes, while a large α allows us to obtain smooth ones. A label was assigned to each shape, and two problems were defined in accordance with ad-hoc rules to depict tasks of increasing complexity:

- Shape Simple (SS), where $Y=+1$ if $\alpha \le 0$ and $Y=-1$ otherwise.
- Shape Difficult (SD), where $Y=+1$ if $-4 \le \alpha \le 4$ and $Y=-1$ otherwise.

In Fig. 2, the samples from Shape domain are shown, we note that for small changes of α recognizing the label for a human can be difficult since the shapes look similar. The probability distribution over the shapes is uniform.

4.1.2. *Word domain*

The Word domain consists of 321 words,[2] sampled from the Wisconsin Perceptual Attribute Ratings Database [78], which includes words rated by 350 undergraduates based on their emotional valence. Two rules were defined for labeling data, analogously to [57] and to what is done above:

- Word Simple (WS): Words were sorted by their length and the 161 longest ones were labeled with $Y=+1$, while the others were labeled with $Y=-1$.
- Word Difficult (WD): Words were sorted by their emotional valence and the 161 most positive ones were labeled with $Y=+1$, while the others were labeled with $Y=-1$.

The probability distribution over the words is uniform. In Table 1, the samples from the Word domain are shown with their emotional valence.

[1] We thank the authors of [57] and [78] for providing their dataset.

[2] Having to deal with Italian students only, words have been translated into Italian.

Table 1
Samples from the Word domain with their emotional valence (e.v.): negative emotion versus positive emotion [78].

Word	e.v.	Word	e.v.	Word	e.v.
Rape	−5.60	Jeer	−2.20	Smile	4.76
Killer	−5.55	Snub	−2.18	Fun	4.91
Funeral	−5.47	Meal	2.54	Laughter	4.95
Slavery	−5.41	Bunny	2.55	Joy	5.19

4.1.3. Human Rademacher Complexity experimental design

In order to compute HRC, the same procedure of [57] has been adopted. RC does not depend on the problem ($\mathbb{P}\{Y|X\}$) but only on the domain ($\mathbb{P}\{X\}$), consequently we measured HRC for the two above-mentioned domains: Shape and Word.

Thus, for each domain and for a fixed value of n, Eq. (37) was computed. In particular, for each student $\mathcal{H}^i$ with $i \in \{1,\dots,m\}$, we measured $\widehat{R}_{\mathrm{EMP}}(\mathcal{H}^i_{\mathcal{L}^{\boldsymbol{\sigma}^i}_{i,n}}, \mathcal{L}^{\boldsymbol{\sigma}^i}_{i,n})$. Consequently, we needed to build $\mathcal{L}^{\boldsymbol{\sigma}^i}_{i,n}$ with $i \in \{1,\dots,m\}$ for the desired domain as follows:

1. Sample randomly n samples from the desired domain.
2. Discard the labels.
3. Assign random labels to the samples.

After creating $\mathcal{L}^{\boldsymbol{\sigma}^i}_{i,n}$, $\widehat{R}_{\mathrm{EMP}}\left(\mathcal{H}^i_{\mathcal{L}^{\boldsymbol{\sigma}^i}_{i,n}}, \mathcal{L}^{\boldsymbol{\sigma}^i}_{i,n}\right)$ was measured as follows:

1. Each student $\mathcal{H}^i$ was asked to learn the rule by exploiting the dataset $\mathcal{L}^{\boldsymbol{\sigma}^i}_{i,n}$.
2. Each student was asked to label the same samples $\mathcal{L}^{\boldsymbol{\sigma}^i}_{i,n}$ where the labels had been removed.

Based on these considerations we designed the questionnaire for measuring $\widehat{R}_{\mathrm{EMP}}(\mathcal{H}^i_{\mathcal{L}^{\boldsymbol{\sigma}^i}_{i,n}}, \mathcal{L}^{\boldsymbol{\sigma}^i}_{i,n})$, for each n, student, and domain. The questionnaire consists of 4 steps (each student was asked to complete the step within the time given in parentheses):

I Students were asked to learn the underlying rule from $\mathcal{L}^{\boldsymbol{\sigma}^i}_{i,n}$, with a short time limit (3 min).
II Students were asked to perform a filler task consisting of some two-digit addition/subtraction questions, to reduce risks of memorization: this forced the students to learn a rule rather than memorizing the examples. In other words, it forced them to think about an irrelevant concept and reset the short-term memory [57,61] (1 min).
III Students were asked to classify the same samples $\mathcal{L}^{\boldsymbol{\sigma}^i}_{i,n}$ where the labels had been removed, and the order in which the samples were presented was different from Step I. Students were not aware of this fact, and they were encouraged to guess if necessary (without time limit).
IV Students were asked to describe the rule they identified, and to estimate the confidence of their decision (1 min).

Then, for each student $\mathcal{H}^i$ with $i \in \{1,\dots,m\}$, a unique questionnaire was given, the answers were collected, $\widehat{R}_{\mathrm{EMP}}(\mathcal{H}^i_{\mathcal{L}^{\boldsymbol{\sigma}^i}_{i,n}}, \mathcal{L}^{\boldsymbol{\sigma}^i}_{i,n})$ was computed, and finally HRC ($\widehat{C}^{\mu}_n$) was derived by averaging the results over the m students with the same n and domain according to Eq. (37).

This procedure is also detailed in [57] but we varied $n \in \{3, 5, 7, 10, 15, 20, 25\}$ instead of $n \in \{5, 10, 20, 40\}$ because, as reported in the experimental result, this range allows us to better interpret the behavior of the measured quantities. Later, we compare our results with [57]. We report in $\widehat{C}^{\mathrm{SHAPE}}_n$ and $\widehat{C}^{\mathrm{WORD}}_n$ the result for the two domains (Shape and Word) and different values of n.

4.1.4. Human algorithmic stability experimental design

A new experimental protocol was designed, differently from what was done for HRC, in order to measure the average HAS. Since AS changes not only based on the domain but also based on the problem, as it depends on $\mathbb{P}\{Y|X\}$, we measured it over four different problems: SS, SD, WS, and WD.

Thus, for each problem and for a fixed value of n, the value of Eq. (40) was computed. In particular, for each student $\mathcal{H}^i$ with $i \in \{1,\dots,m\}$ we measured $|\ell(\mathcal{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathcal{H}^i_{\mathcal{L}^{\setminus i}_{i,n}}, Z)|$ with $Z \in \mathcal{T}_{i,p}$. Since we use the hard loss function:

$$\left|\ell(\mathcal{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell\left(\mathcal{H}^i_{\mathcal{L}^{\setminus i}_{i,n}}, Z\right)\right| = \left|\mathcal{H}^i_{\mathcal{L}_{i,n}}(X) - \mathcal{H}^i_{\mathcal{L}^{\setminus i}_{i,n}}(X)\right|, \tag{45}$$

which means that in order to compute $|\ell(\mathcal{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathcal{H}^i_{\mathcal{L}^{\setminus i}_{i,n}}, Z)|$, we simply needed to check if the sample Z had been labeled by the students in the same way after learning $\mathcal{L}^{\setminus i}_{i,n}$ or $\mathcal{L}_{i,n}$. In order to measure this quantity, the following steps were performed:

1. Each student $\mathcal{H}^i$ was asked to learn the rule by exploiting the dataset $\mathcal{L}^{\setminus i}_{i,n}$, which was the dataset $\mathcal{L}_{i,n}$ where a random sample $i \in \{1,\dots,n\}$ had been removed.
2. Each student was asked to label the samples $Z \in \mathcal{T}_{i,p}$ where the labels had been removed.
3. Each student $\mathcal{H}^i$ was asked to learn the rule by exploiting the whole dataset $\mathcal{L}_{i,n}$.
4. Each student was asked to label the samples $Z \in \mathcal{T}_{i,p}$ again where the labels had been removed.

Once all these data were collected, one could measure $|\ell(\mathcal{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathcal{H}^i_{\mathcal{L}^{\setminus i}}, Z)|$. As in the previous section we built the questionnaire, consisting of 8 steps:

I Students were asked to learn the underlying rule from $\mathcal{L}^{\setminus i}_{i,n}$ with a short time limit (3 min).
II Students were asked to perform a filler task, as in the HRC experiment (1 min).
III Students were asked to classify the samples in $\mathcal{T}_{i,p}$ where the labels had been removed (without time limit).
IV Students were asked to describe the rule they identified, and to estimate the confidence of their decision (1 min).
V Students were asked to learn the underlying rule from $\mathcal{L}_{i,n}$ (these samples were presented in a random order respect to Step I) with a short time limit (3 min). They were not aware that $n-1$ training instances were the same as in Step I.
VI Students were asked to perform a filler task (1 min).
VII Students were asked to classify the samples in $\mathcal{T}_{i,p}$ (these samples were presented in a random order respect to Step III) where the labels had been removed (without time limit).
VIII Students were asked to describe the rule they identified, and to estimate the confidence of their decision (1 min).

Note that first $\mathcal{L}^{\setminus i}_{i,n}$ and then $\mathcal{L}_{i,n}$ were provided in order to avoid the risk of memorization.

Then, for each student $\mathcal{H}^i$ with $i \in \{1,\dots,m\}$, a unique questionnaire was given, the answers were collected, $|\ell(\mathcal{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathcal{H}^i_{\mathcal{L}^{\setminus i}_{i,n}}, Z)|$ was computed, and finally HAS ($\widehat{H}^{\mu}_n$) was derived by averaging the results over the m students with the same n and problem according to Eq. (40).

The results for two domains and two levels of difficulty were collected, by varying $n \in \{3, 5, 7, 10, 15, 20, 25\}$, in $\widehat{H}^{\mathrm{SS}}_n$, $\widehat{H}^{\mathrm{SD}}_n$, $\widehat{H}^{\mathrm{WS}}_n$, and $\widehat{H}^{\mathrm{WD}}_n$, respectively.

4.1.5. Human Error experimental design

The experiment designed for measuring HAS (see Section 4.1.4) allowed us to additionally measure Human Error.

In particular, for each student $\mathcal{H}^i$ with $i \in \{1,\dots,m\}$, n, and problem, we measured $\ell(\mathcal{H}^i_{\mathcal{L}_{i,n}}, Z)$ with $Z \in \mathcal{T}_{i,p}$. In other words:

1. Each student $\mathcal{H}^i$ was asked to learn the rule by exploiting the dataset $\mathcal{L}_{i,n}$.
2. Each student was asked to label the samples $Z \in \mathcal{T}_{i,p}$ where the labels had been removed.

The questionnaire was designed in the same way as the one of Section 4.1.4 and in particular, Steps V, VI, VII, and VIII: this was advantageous for us since all the quantities were already available and there was no need to build an additional experiment. Moreover, from Steps I, II, III, and IV the data for measuring Human Error on $n-1$ was also available.

Then, for each student $\mathcal{H}^i$ with $i \in \{1,\dots,m\}$, a unique questionnaire was given, the answers were collected, $\ell(\mathcal{H}^i_{\mathcal{L}_{i,n}}, Z)$ with $Z \in \mathcal{T}_{i,p}$ was computed, and finally Human Error ($\widehat{R}^{\mu}_n$) was derived by averaging the results over the m students with the same n and problem according to Eq. (40) (since Human Error depends on $\mathbb{P}\{Y|X\}$).

Based on the designed HAS experiment, it was possible to collect data for $n \in \{3,5,7,10,15,20,25\}$ and $n-1$. The results for two domains and two levels of difficulty were collected, by varying n, in $\widehat{R}^{\mathrm{SS}}_n$, $\widehat{R}^{\mathrm{SD}}_n$, $\widehat{R}^{\mathrm{WS}}_n$, and $\widehat{R}^{\mathrm{WD}}_n$, respectively.

4.1.6. Aggregated questionnaire structure

The aggregated questionnaire consists of 3 sub-questionnaires, 2 for HAS and 1 for HRC:

- A questionnaire of Section 4.1.4 for SS or SD.
- A questionnaire of Section 4.1.4 for WD or WS.
- A questionnaire of Section 4.1.3 for Word or Shape.

We avoided including two simple or two difficult problems in one questionnaire. Note that with two students, there was a complete set of problems: HAS for SS, SD, WS, and WD, as well as HRC for Word and Shape. Human Error was computed for SS, SD, WS, and WD thanks to the HAS questionnaires.

Before giving the questionnaires to the students, we devoted approximately 10 min to explain the experiment procedure and our study goals to the students. In the early trials of this study, we noticed that this kind of explanation increases the motivation of the students to try their best in answering the questions.

The size $n \in \{3,5,7,10,15,20,25\}$ of the sets were randomly chosen. In EX^1, since estimation accuracy of Human Error, AS, and RC depends on m (see Section 3), we set $p=1$. Each experiment took about 30 min in total.

A first trial, conducted on 70 volunteers, showed that students were able to mentally link Steps I and III with Steps V and VII due to the problems described in Section 3.3. This result confirmed our hypothesis that humans easily memorize and they cannot easily forget a previously memorized rule. The detailed results are reported in the experimental result sections. In the next section, we explain a modified experiment which allowed us to successfully measure HAS.

4.2. Experimental design: EX^2

As described above, we had to modify the experiment in order to measure HAS; but for HRC, we kept the design of EX^1.

4.2.1. Human algorithmic stability experiment design

Instead of measuring $\widehat{H}^{\mu}_n$, we needed to measure $\widehat{\widehat{H}}^{\mu}_n$ for each value of n and problem (see Eq. (43)). In particular, for a pair of students $\mathcal{H}^i$ and $\mathcal{H}^{i+m}$ with $i \in \{1,\dots,m\}$, we measured $|\ell(\mathcal{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathcal{H}^{i+m}_{\mathcal{L}^{\setminus i}_{i,n}}, Z)|$ with $Z \in \mathcal{T}_{i,p}$. In order to compute $|\ell(\mathcal{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathcal{H}^{i+m}_{\mathcal{L}^{\setminus i}_{i,n}}, Z)|$, we simply needed to check if the same label had been assigned to Z by both students when one learned the rule through $\mathcal{L}^{\setminus i}_{i,n}$, and the other one through $\mathcal{L}_{i,n}$. In order to measure $|\ell(\mathcal{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathcal{H}^{i+m}_{\mathcal{L}^{\setminus i}_{i,n}}, Z)|$, the following steps were performed:

1. Each student $\mathcal{H}^i$ was asked to learn the rule by exploiting the dataset $\mathcal{L}^{\setminus i}_{i,n}$, while student $\mathcal{H}^{i+m}$ was asked to learn the rule by exploiting the whole dataset $\mathcal{L}_{i,n}$.
2. Student $\mathcal{H}^i$ and student $\mathcal{H}^{i+m}$ were asked to label the samples $Z \in \mathcal{T}_{i,p}$ where the labels had been removed.

The aggregated questionnaire was the same as Section 4.1.4 but split over a pair of students: phases I, II, III, and IV were assigned to one of the students and phases V, VI, VII, and VIII to the other. Thanks to this procedure, all quantities, necessary to compute HAS, could be derived: for every pair of students $\mathcal{H}^i$ and $\mathcal{H}^{i+m}$ with $i \in \{1,\dots,m\}$, a unique questionnaire was given, the answers were collected, $|\ell(\mathcal{H}^i_{\mathcal{L}_{i,n}}, Z) - \ell(\mathcal{H}^{i+m}_{\mathcal{L}^{\setminus i}_{i,n}}, Z)|$ was computed, and finally HAS $\widehat{\widehat{H}}^{\mu}_n$ was derived by averaging the results over m pair of students with same n and problem according to Eq. (43).

As in Section 4.1.4, The results for two domains and two levels of difficulty were collected, by varying n, in $\widehat{\widehat{H}}^{\mathrm{SS}}_n$, $\widehat{\widehat{H}}^{\mathrm{SD}}_n$, $\widehat{\widehat{H}}^{\mathrm{WS}}_n$, and $\widehat{\widehat{H}}^{\mathrm{WD}}_n$, respectively.

4.2.2. Human Error

As for EX^1, the AS questionnaire allows us to measure Human Error as well, although, the data had to be collected from a pair of students instead of a single student. Note that the available data is exactly the same as Section 4.1.4.

4.2.3. Aggregated questionnaire structure

The aggregated questionnaire consists of 5 sub-questionnaires, 4 for HAS and 1 for HRC but split over two students:

- Four questionnaires of Section 4.2.1 for SS, SD, WS, and WD: phases I, II, III, and IV were given to one of the students, while phases V, VI, VII, and VIII were given to the other.
- One questionnaire of Section 4.1.3 for Word or Shape: Word was given to one of the students, while Shape was given to the other.

Analogously to EX^1, for EX^2, n was randomly chosen in $\{3,5,7,10,15,20,25\}$, we set $p=1$ and approximately 10 min were devoted before the experiment, to explain the experiment procedure and our study goals to the students.

We obtained interesting results by providing these questionnaires to 307 volunteers and the results were partly published in [58]. We report here the complete set of results including the extension of this experiment to more students and an additional domain. In the next section we explain how the experiment was extended. In particular, we address a problem raised in [58] showing that it is possible to reduce the oscillation of the results with a better estimate of the quantity of interest (Human Error and HAS).

4.3. Experimental design: EX^3

In EX^3, we decided not to measure HRC for two reasons: firstly, as can be seen later in the experimental results, we managed to replicate and obtain satisfactory results with EX^1 and EX^2 which are consistent with the one reported in [57]. Consequently, our fist goal of being able to reproduce and verify the results of [57] was already achieved. Secondly, by excluding RC, we had the possibility to insert another domain into the experiment without increasing the length

of the experiment, and compromising the level of attention of the students, which could produce artefacts in the final results [61].

In EX^3, we made two major changes respect to EX^2 (apart from the above-mentioned changes). Firstly, instead of setting $p=1$, we decided to set $p=15$ in order to improve the accuracy and statistical robustness of the estimator (see Section 3). In other words, we computed a more accurate estimate of the difference between the rules found by the students in each pair. Secondly, we added a new domain: the Math domain.

In the next sections, these modifications are described in detail.

4.3.1. Math domain

The Math domain consists of 321 numbers $x \in \{0, 1, \ldots, 320\}$. A label was assigned to each number, and two problems were defined in accordance with ad-hoc rules to depict tasks of increasing complexity:

- Math Simple (MS), where $Y=+1$ if $x \leq 159$ and $Y=-1$ otherwise.
- Math Difficult (MD), where $Y=+1$ if $108 \leq x \leq 214$ and $Y=-1$ otherwise.

The probability distribution over the numbers is uniform.

4.3.2. Aggregated questionnaire structure

The aggregated questionnaire consists of 6 HAS sub-questionnaires of Section 4.2.1, one for each problem among SS, SD, WS, WD, MS, and MD. As in EX^2, phases I, II, III, and IV were given to a student of a pair, while phases V, VI, VII, and VIII were given to the other. The order in which the problems were presented was randomized since, in case of crowded classes, it was impossible for students to copy the labels from each other.

Analogously to EX^2, in EX^3, n was randomly chosen in $\{3, 5, 7, 10, 15, 20, 25\}$ but, differently from EX^2, we set $p=15$. As was the case with EX^1 and EX^2, we devoted approximately 10 min before the experiment, to explain the experiment procedure and our study goals to the students. The experiment itself took about 30 min, the same as EX^1 and EX^2.

5. Experimental results

In the following sections, the results of the three experiments (EX^1, EX^2, and EX^3) are presented.[3] In particular, for EX^1, the following quantities are reported:

- Human Error, $\widehat{R}_n^{\mu}$.
- Human Rademacher Complexity, $\widehat{C}_n^{\mu}$ (where we set $p=1$).
- Human Algorithmic Stability, $\widehat{H}_n^{\mu}$ ($p=1$).

In the first experiment $\mu \in \{\mathrm{SS}, \mathrm{SD}, \mathrm{WS}, \mathrm{WD}\}$. Note that, since $\widehat{C}_n^{\mu}$ depends just on the domain, $\widehat{C}_n^{\mathrm{SHAPE}} = \widehat{C}_n^{\mathrm{SS}} = \widehat{C}_n^{\mathrm{SD}}$ and $\widehat{C}_n^{\mathrm{WORD}} = \widehat{C}_n^{\mathrm{WS}} = \widehat{C}_n^{\mathrm{WD}}$ (see Section 4.1.3).

For EX^2 instead, we report the following quantities:

- Human Error, $\widehat{R}_n^{\mu}$.
- Human Rademacher Complexity, $\widehat{C}_n^{\mu}$ ($p=1$).
- Human Algorithmic Stability, $\widehat{H}_n^{\mu}$ ($p=1$).

Note that in the second experiment, $\mu \in \{\mathrm{SS}, \mathrm{SD}, \mathrm{WS}, \mathrm{WD}\}$ as in the first one, but we measure $\widehat{\widehat{H}}_n^{\mu}$ instead of $\widehat{H}_n^{\mu}$ for the reason that $\widehat{H}_n^{\mu}$ cannot be effectively measured (see Section 4.2.1).

Finally, for EX^3, we report the following quantities:

- Human Error, $\widehat{R}_n^{\mu}$ (where we set $p=15$).
- Human Algorithmic Stability, $\widehat{\widehat{H}}_n^{\mu}$ ($p=15$).

In the third experiment, $\mu \in \{\mathrm{SS}, \mathrm{SD}, \mathrm{WS}, \mathrm{WD}, \mathrm{MS}, \mathrm{MD}\}$ as we added another domain (Math) and two problems (simple and difficult). Note that in this case HRC was not measured.

The results are measured for $n \in \{3, 5, 7, 10, 15, 20, 25\}$ in all the experiments. Thanks to the designed experiments, we can also measure $\widehat{R}_n^{\mu}$ for $n-1$ (see Section 4.1.5).

5.1. Results for EX^1

The first experiment was carried out as a pilot test on a small population, a group of 70 students. Therefore, we lack data for some problems and for some values of n. For example, in Fig. 3(c), $\widehat{C}_{20}^{\mathrm{SHAPE}}$ and $\widehat{C}_{25}^{\mathrm{SHAPE}}$ are missing.

Fig. 3 (a) shows the trend of $\widehat{R}_n^{\mu}$, for different values of n and different problems μ: as expected from ML Theory, $\widehat{R}_n^{\mu}$ is generally smaller for simple tasks than for difficult ones and this is confirmed in HL as well. However, analogies end here. While the error of ML models usually decreases with n, results on HL are characterized by oscillations, even for small variations of n. This is due to the fact that a small sample is considered.

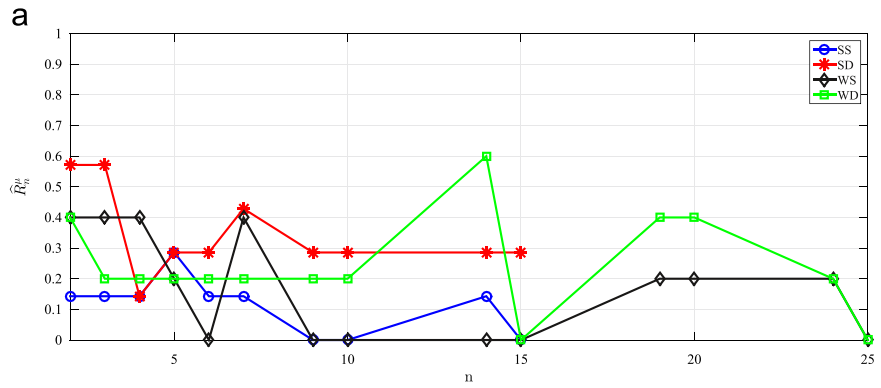
Fig. 3 (c) shows the trend for $\widehat{C}_n^{\mu}$. Contrarily to HAS, HRC is not able to discriminate between the task complexities, since labels are neglected when computing $\widehat{C}_n^{\mu}$. HRC decreases with n (as in ML), and this trend is substantially uncorrelated with the errors for the considered domains. Note that, due to the fact that a small sample of students is considered, HRC is characterized by oscillations as well.

Finally, Fig. 3(b) presents the obtained results when computing $\widehat{H}_n^{\mu}$ (with $p=1$) as n varies. Despite being designed in the ML framework, it is worth highlighting how HAS is able to grasp the nature and peculiarities of HL. As a matter of fact, simple tasks are characterized by smaller values of $\widehat{H}_n^{\mu}$, and HAS for the Shape domain is generally smaller than for the Word domain. Both results are in accordance with the trend of the error, registered in HL, and the nature of the analyzed phenomenon: in this sense, HAS offers interesting insights on HL, because it raises questions about the ability of humans to learn in and across different domains. By analyzing the results and in particular the answers to the questions IV and VIII of the HAS questionnaire (see Section 4.1.4), in a majority of cases, the students did not change the rule when learned during the Step VI respect to the Step I (see Section 4.1.4). Therefore, a lot of zeros in Fig. 3(b) can be seen. After observing this phenomenon, we decided to ask the students about the reason behind their choice. Their answers showed that they recognized that the problems of Steps I and V are the same thus, instead of learning a new rule, they just memorized their previous answer to the problem. This is a quite interesting result since it confirms our initial idea about the learning behavior of humans. Again, the oscillations in HAS results are due to the fact that a small sample is considered.

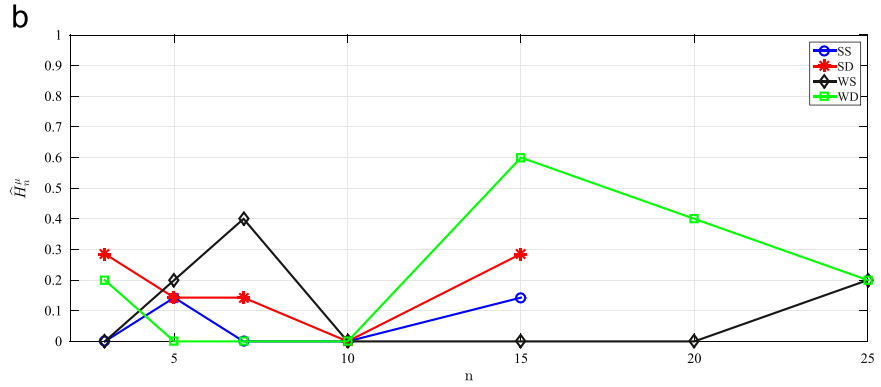
5.2. Results for EX^2

In EX^2 the experiment was carried out over 306 students in order to address the issues of EX^1 related to the lack of data. Moreover, we measure HAS in a different way: we use $\widehat{\widehat{H}}_n^{\mu}$ (with $p=1$) instead of $\widehat{H}_n^{\mu}$ in order to fix the memorization problem encountered during EX^1. These results are partially described in [58].

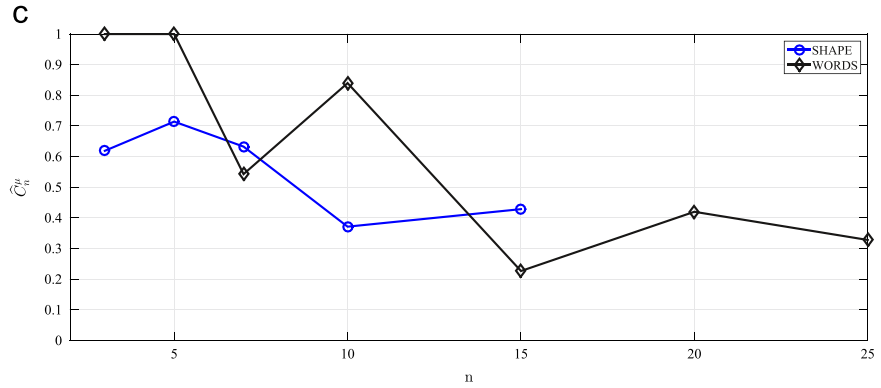
[3] Some examples of the filled and anonimized questionnaires can be retrieved from www.la.smartlab.ws/filled-questionnaires.zip.



Human Error for different values of n and different problems μ (Shape Simple (SS), Shape Difficult (SD), Word Simple (WS), and Word Difficult (WD)).



Human Algorithmic Stability (HAS) for different values of n and different problems μ (Shape Simple (SS), Shape Difficult (SD), Word Simple (WS), and Word Difficult (WD)).



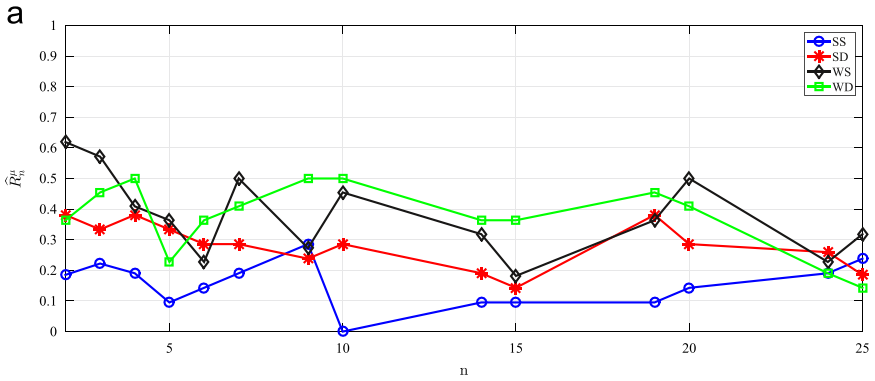
Human Rademacher Complexity (HRC) for different values of n and different domains (Shape and Word).

Fig. 3. Results of EX[1].

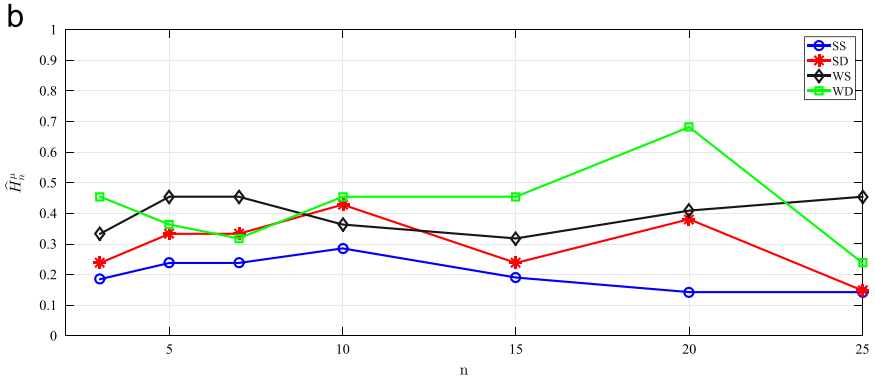
Fig. 4(a) represents Human Error for different problems with varied number of samples available for learning: the results are quite similar to the ones of EX[1] where the oscillations are still a problem. At this point we had the impression that only a subset of students were willing to perform at their best when completing the questionnaire (due to the lack of motivation and concentration) and this could compromise the result. We explored this hypothesis by analyzing the filler tasks, in order to verify the students' level of attention. We discarded the students with low level of attention (the ones with many errors or incomplete filler tasks) but the results remained almost unchanged. Consequently, in EX[3] we set $p=15$ in order to get a more stable estimate. Note that, in any case, the curve is generally higher for difficult problems respect to the simple ones and for the Word domain with respect to the Shape domain. Unfortunately, Human Error still does not decrease with n.

Fig. 4 (b) represents HRC by varying n in different domains. In this case, we significantly reduced the oscillations respect to EX[1] (Fig. 3(b)). It is worth highlighting that the curve is very similar to the one obtained by [57] for the measure of HRC. Note that RC is able to grasp the difference between the domains (Shape domain is a simpler task compared to the Word one). Moreover, it has the same drawbacks depicted in the results of EX[1].

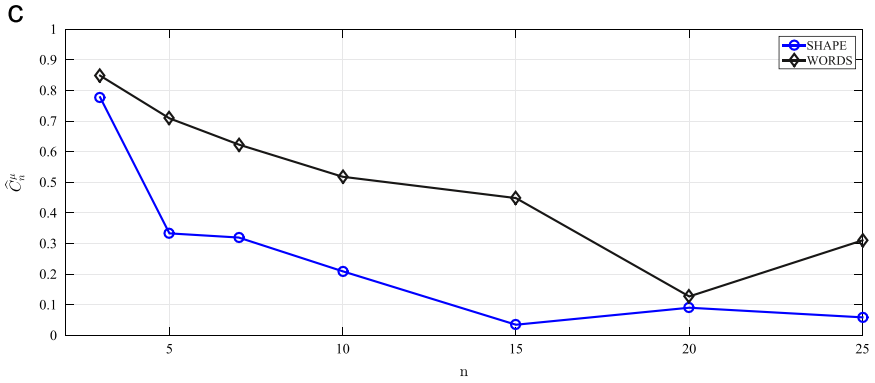
Finally, Fig. 4(c) shows the trend of HAS by varying n. Results are, again, similar to the one of EX[1] (the curve is generally higher



Human Error for different values of n and different problems μ (Shape Simple (SS), Shape Difficult (SD), Word Simple (WS), and Word Difficult (WD)).



Human Algorithmic Stability (HAS) for different values of n and different problems μ (Shape Simple (SS), Shape Difficult (SD), Word Simple (WS), and Word Difficult (WD)).



Human Rademacher Complexity (HRC) for different values of n and different domains (Shape and Word).

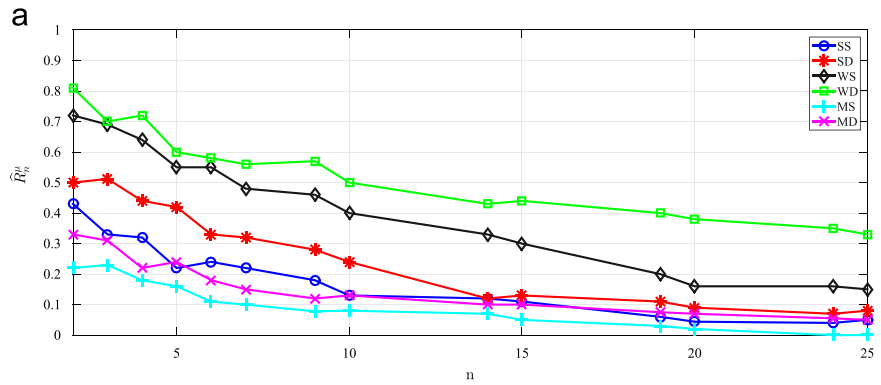
Fig. 4. Results of EX[2].

for difficult problems respect to the simpler ones and for the Word domain with respect to the Shape domain, analogously to what happens with Human Error) although, the oscillations are mostly reduced compared to EX[1]. Also in this case, we tried to discard the students with low level of attention but, again, the results did not noticeably change and for this reason we did not report it. We address this issue in EX[3] by adding the parameter $p=15$ in measuring both Human Error and HAS.

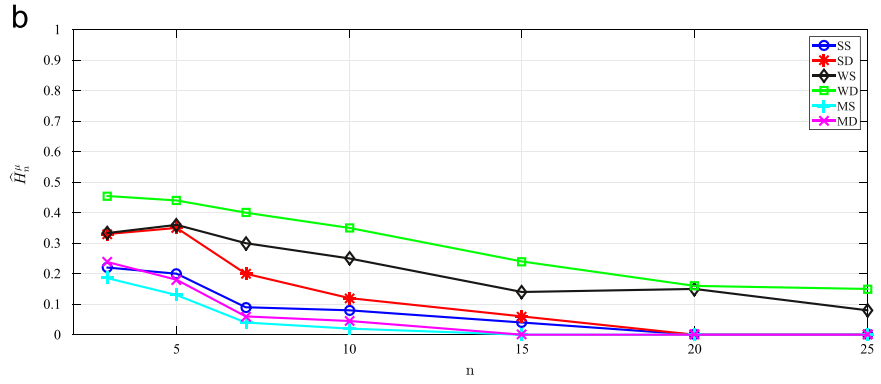
5.3. *Results for* EX[3]

In EX[3] the experiment was carried out with 229 participants in order to address the issues of EX[2] related to stability of the estimators. Moreover, we did not measure HAR, and we use $\widehat{\overline{R}}_n^{\mu}$ and $\widehat{\overline{H}}_n^{\mu}$ with $p=15$ in order to obtain a more stable estimate.

In Fig. 5(a) and (b), it can be seen that the oscillations are substantially reduced. Both Human Error and HAS decreases with n (as expected from ML Theory). Results show that the Math domain is the simplest while the Word domain is the most difficult for the students. This can be due to the fact that all students are from engineering majors, so they are more used to the problems related to Mathematics. Moreover, Human Error and HAS are larger for difficult problems compared to easier ones. The ranking of problems (from difficult to simple) results to be: WD, WS, SD, SS, MD, and MS. While there is a clear distinction between the Word problems (difficult or simple) and the other domains, the distinction



Human Error for different values of n and different problems μ (Shape Simple (SS), Shape Difficult (SD), Word Simple (WS), Word Difficult (WD), Math Simple (MS), and Math Difficult (MD)).



Human Algorithmic Stability (HAS) for different values of n and different problems μ (Shape Simple (SS), Shape Difficult (SD), Word Simple (WS), Word Difficult (WD), Math Simple (MS), and Math Difficult (MD)).

Fig. 5. Results of EX[3].

between Shape and Math problems is not as clear. This effect is more visible as n increases. This can be due to the fact that the Math and the Shape domains leave less space for imagination to find meaningful rules. Instead, the Word domain might open a whole world of possible interpretations and rules which can be discovered and selected. For instance, we observed that some students related the words as part of a common story or image for discovering the rule. This observation is also supported by the HRC experiment which shows how the capacity of HL in case of the Word domain, is larger respect the one of Shape domain.

As a matter of fact, our study results are derived from and limited to the students of engineering majors. As discussed above, these results can be dependent on the background knowledge and the major of the students. For example, students of humanities or literature majors might have better results for the Word domain as opposed to the Math or Shape domain.

In conclusion, we underline the great potentials of HAS to measure the HL capacity in different domains in particular in educational settings. For example, our approach can be integrated as a LA method for improving TEL systems for the purpose of personalization and adaptation of educational materials to the needs of students. Additionally, it can raise awareness of teachers to balance the difficulty of exercises based on the needs of the students with different academic backgrounds.

6. Conclusions

LA has gained a lot of attention in the last decade to gain better insight into HL through the application of ML methods and the use of COGS models. Recently ML tools have been exploited in COGS to understand HL. In addition to providing algorithms for extracting information from the data, ML provides tools to analyze the learning capacity of algorithms.

AS is an effective ML tool for understanding the learning ability of algorithms. In this paper, we propose to exploit this tool for obtaining insight into HL and we show that HAS is more informative towards HL than HRC which was previously studied in [57]. We conducted our experiments with 606 students of various engineering majors from the University of Genoa. Our results showed that: both HAR and HAS are able to detect the difficulty level of different domains for a group of students. However, contrary to HAS, HRC requires some additional assumptions to be measured that are seldom or never satisfied in HL [61]. Additionally, HAS extends these results by detecting the difficulty level of different problems in the same domain. In particular, the difficulty scale of the problems (from difficult to simple) for the students is: WD, WS, SD, SS, MD, and MS.

Our results suggest that ML can offer new opportunities in the study of HL in the fields of LA and COGS. In particular, CL can be enhanced and studied further by application of ML methods and integrated into instructional design in classrooms or TEL systems to improve education. In this context, educational reform has been mentioned as the most important and relevant application of CL [48] such that approaches in manipulating category labels, presentation order, learning strategies, and category variability can optimize CL. Consequently, ML by improving CL opens the doors toward improving HL in educational settings. Recent works in CL [56,79,55,57] highlight how cross fertilization between ML and HL can be extended to better understand how people tackle new

problems and extract knowledge from observations. For instance, HAS can improve the ability of educators to detect when learners are passively memorizing the concepts rather than discovering new knowledge. Note that HAS, unlike HRC, is not only able to explain the difficulty level of the particular domain for learners, but also is able to detect the difficulty of a problem in a domain. Consequently, this ability can lead to better personalization and adaptation of education to the needs of students.

Acknowledgments

This work was supported in part by the Erasmus Mundus Joint Doctorate in Interactive and Cognitive Environments, which is funded by the EACEA Agency of the European Commission under EMJD ICE FPA n 2010-0012. Also, we thank the Professors Giuliano Donzellini, Carla Gambaro, Franco Parodi, Domenico Ponta, Marco Storace, and Eugenia Torello of the University of Genoa who provided their time for running the experiments during their classes, and Remi Brochenin, Emanuele Fumeo, Alessandro Ghio, Ilenia Orlandi, and Jorge Luis Reyes Ortiz for providing their support to our experiment. Finally we thank the authors of [57] and [78] for providing their datasets.

References

[1] M.A. Chatti, A.L. Dyckhoff, U. Schroeder, H. Thüs, A reference model for learning analytics, Int. J. Technol. Enhanc. Learn. 4 (5) (2012) 318–331.
[2] J.I. Lee, E. Brunskill, The impact on individualizing student models on necessary practice opportunities, in: International Conference on Educational Data Mining, 2012.
[3] M. Brown, Learning analytics: moving from concept to practice, in: EDUCAUSE Learning Initiative, 2012.
[4] M. Vahdat, A. Ghio, L. Oneto, D. Anguita, M. Funk, M. Rauterberg, Advances in learning analytics and educational data mining, in: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2015.
[5] Z. Papamitsiou, A. Economides, Learning analytics and educational data mining in practice: a systematic literature review of empirical evidence, Educ. Technol. Soc. 17 (4) (2014) 49–64.
[6] G. Siemens, P. Long, Penetrating the fog: analytics in learning and education, EDUCAUSE Rev. 46 (5) (2011) 30–32.
[7] M. Bienkowski, M. Feng, B. Means, Enhancing teaching and learning through educational data mining and learning analytics: an issue brief, US Department of Education, Office of Educational Technology, 2012, pp. 1–57.
[8] K.R. Koedinger, E. Brunskill, S.S. Baker, E.A. McLaughlin, J. Stamper, New potentials for data-driven intelligent tutoring system development and optimization, AI Mag. 34 (3) (2013) 27–41.
[9] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, D. Koller, Tuned models of peer assessment in moocs, in: arXiv preprint arXiv:1307.2579, 2013.
[10] R. Ferguson, Learning analytics: drivers, developments and challenges, Int. J. Technol. Enhanc. Learn. 4 (5) (2012) 304–317.
[11] T.A. Polk, C.M. Seifert, Cognitive Modeling, MIT Press, London, 2002.
[12] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.
[13] J.M. Lillo-Castellano, I. Mora-Jiménez, C. Figuera-Pozuelo, J.L. Rojo-Álvarez, Traffic sign segmentation and classification using statistical learning methods, Neurocomputing 153 (2015) 286–299.
[14] Y. Yuan, Q. Guo, X. Lu, Image quality assessment: a sparse learning way, Neurocomputing 159 (2015) 227–241.
[15] X. Zhang, Y. Li, Adaptive energy detection for bird sound detection in complex environments, Neurocomputing 155 (2015) 108–116.
[16] Y. Tian, Q. Ruan, G. An, W. Xu, Context and locality constrained linear coding for human action recognition, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2015.04.059, in press.
[17] D.J.C. MacKay, Information Theory, Inference and Learning Algorithms, Cambridge university Press, Cambridge, 2003.
[18] T. Hastie, R. Tibshirani, J. Friedman, Unsupervised Learning, Springer, New York, 2009.
[19] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, 2004.
[20] M.J. Baker, The roles of models in artificial intelligence and education research: a prospective view, J. Artif. Intell. Educ. 11 (2000) 122–143.
[21] S. Kotsiantis, C. Pierrakeas, P. Pintelas, Predicting students' performance in distance learning using machine learning techniques, Appl. Artif. Intell. 18 (5) (2004) 411–426.
[22] P. Brusilovsky, S. Sosnovsky, O. Shcherbinina, User modeling in a distributed e-learning architecture, in: User Modeling, 2005.
[23] M. Rauterberg, S. Schluep, M. Fjeld, How to model behavioural and cognitive complexity in human-computer interaction with petri nets, in: International Workshop on Robot and Human Communication, 1997.
[24] K.E. Arnold, M.D. Pistilli, Course signals at Purdue: using learning analytics to increase student success, in: International Conference on Learning Analytics and Knowledge, 2012.
[25] E. Triantafillou, A. Pomportsis, S. Demetriadis, The design and the formative evaluation of an adaptive educational system based on cognitive styles, Comput. Educ. 41 (1) (2003) 87–103.
[26] V.N. Vapnik, Statistical Learning Theory, Wiley-Interscience, New York, 1998.
[27] D. Anguita, A. Ghio, L. Oneto, S. Ridella, In-sample and out-of-sample model selection and error estimation for support vector machines, IEEE Trans. Neural Netw. Learn. Syst. 23 (9) (2012) 1390–1406.
[28] V. Koltchinskii, Rademacher penalties and structural risk minimization, IEEE Trans. Inf. Theory 47 (5) (2001) 1902–1914.
[29] P.L. Bartlett, O. Bousquet, S. Mendelson, Local Rademacher complexities, Ann. Stat. 33 (4) (2005) 1497–1537.
[30] D. Anguita, A. Ghio, S. Ridella, Maximal discrepancy for support vector machines, Neurocomputing 74 (9) (2011) 1436–1443.
[31] D.A. McAllester, Some PAC-Bayesian theorems, in: Computational Learning Theory, 1998.
[32] G. Lever, F. Laviolette, J. Shawe-Taylor, Tighter PAC-Bayes bounds through distribution-dependent priors, Theor. Comput. Sci. 473 (2013) 4–28.
[33] L. Oneto, A. Ghio, S. Ridella, D. Anguita, Fully empirical and data-dependent stability-based bounds, IEEE Trans. Cybern. (2016), http://dx.doi.org/10.1109/TCYB.2014.2361857, in press.
[34] S. Floyd, M. Warmuth, Sample compression, learnability, and the Vapnik–Chervonenkis dimension, Mach. Learn. 21 (3) (1995) 269–304.
[35] O. Bousquet, A. Elisseeff, Stability and generalization, J. Mach. Learn. Res. 2 (2002) 499–526.
[36] T. Poggio, R. Rifkin, S. Mukherjee, P. Niyogi, General conditions for predictivity in learning theory, Nature 428 (6981) (2004) 419–422.
[37] J.S. Bruner, G.A. Austin, A Study of Thinking, Transaction Publishers, New York, 1956.
[38] S. Watanabe, Pattern Recognition: Human and Mechanical, John Wiley & Sons, Inc., New York, 1985.
[39] H. Pashler, M.C. Mozer, When does fading enhance perceptual category learning? J. Exp. Psychol.: Learn. Memory Cogn. 39 (4) (2013) 1162.
[40] M. Rauterberg, About a framework for information and information processing of learning systems, in: ISCO, 1995.
[41] M. Rauterberg, E. Ulich, Information processing for learning systems: an action theoretical approach, in: IEEE International Conference on Systems, Man, and Cybernetics, 1996.
[42] N.D. Goodman, J.B. Tenenbaum, J. Feldman, T.L. Griffiths, A rational analysis of rule-based concept learning, Cogn. Sci. 32 (1) (2008) 108–154.
[43] D. Vats, C. Studer, A.S. Lan, L. Carin, R. Baraniuk, Test-size reduction for concept estimation, in: International Conference on Educational Data Mining, 2013.
[44] J.K. Kruschke, Alcove: an exemplar-based connectionist model of category learning, Psychol. Rev. 99 (1) (1992) 22.
[45] D.L. Medin, M.M. Schaffer, Context theory of classification learning, Psychol. Rev. 85 (3) (1978) 207.
[46] R.M. Nosofsky, T.J. Palmeri, A rule-plus-exception model for classifying objects in continuous-dimension spaces, Psychon. Bull. Rev. 5 (3) (1998) 345–369.
[47] G.L. Murphy, The Big Book of Concepts, MIT Press, New York, 2002.
[48] R.L. Goldstone, A. Kersten, Concepts and categorization, in: Handbook of Psychology, 2003.
[49] G.O. Deák, M.S. Bartlett, T. Jebara, New trends in cognitive science: integrative approaches to learning and development, Neurocomputing 70 (13) (2007) 2139–2147.
[50] K. Madani, C. Sabourin, Multi-level cognitive machine-learning based concept for human-like artificial walking: application to autonomous stroll of humanoid robots, Neurocomputing 74 (8) (2011) 1213–1228.
[51] T. Matsuka, Y. Sakamoto, A. Chouchourelou, J.V. Nickerson, Toward a descriptive cognitive model of human learning, Neurocomputing 71 (13) (2008) 2446–2455.
[52] T. Joachims, Learning representations of student knowledge and educational content, in: International Conference on Machine Learning Workshop—Machine Learning for Education, 2015.
[53] C. Piech, M. Sahami, D. Koller, S. Cooper, P. Blikstein, Modeling how students learn to program, in: ACM Technical Symposium on Computer Science Education, 2012.
[54] A.S. Lan, A.E. Waters, C. Studer, R.G. Baraniuk, Sparse factor analysis for learning and content analytics, J. Mach. Learn. Res. 15 (1) (2014) 1959–2008.
[55] T.L. Griffiths, B.R. Christian, M.L. Kalish, Using category structures to test iterated learning as a method for identifying inductive biases, Cogn. Sci. 32 (1) (2008) 68–107.
[56] J. Feldman, Minimization of boolean complexity in human concept learning, Nature 407 (6804) (2000) 630–633.
[57] X. Zhu, B.R. Gibson, T.T. Rogers, Human Rademacher complexity, in: Neural Information Processing Systems, 2009.
[58] M. Vahdat, L. Oneto, A. Ghio, D. Anguita, D. Funk, M. Rauterberg, Human algorithmic stability and human Rademacher complexity, in: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2015.

[59] P.L. Bartlett, S. Mendelson, Rademacher and Gaussian complexities: risk bounds and structural results, J. Mach. Learn. Res. 3 (2003) 463–482.
[60] P. Klesk, M. Korzen, Sets of approximating functions with finite Vapnik–Chervonenkis dimension for nearest-neighbors algorithms, Pattern Recognit. Lett. 32 (14) (2011) 1882–1893.
[61] D.L. Schacter, D.T. Gilbert, D.M. Wegner, Psychology, Second ed., Worth Publishers, New York, 2010.
[62] X. Zhu, Machine teaching: an inverse problem to machine learning and an approach toward optimal education, in: AAAI Conference on Artificial Intelligence (Senior Member Track), 2015.
[63] P.L. Bartlett, S. Boucheron, G. Lugosi, Model selection and error estimation, Mach. Learn. 48 (1–3) (2002) 85–113.
[64] L. Oneto, A. Ghio, S. Ridella, D. Anguita, Global Rademacher complexity bounds: from slow to fast convergence rates, Neural Process. Lett. (2016), http://dx.doi.org/10.1007/s11063-015-9429-2, in press.
[65] L. Oneto, A. Ghio, S. Ridella, D. Anguita, Learning resource-aware classifiers for mobile devices: from regularization to energy efficiency, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2014.12.099, in press.
[66] S. Mukherjee, P. Niyogi, T. Poggio, R. Rifkin, Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization, Adv. Comput. Math. 25 (1) (2006) 161–193.
[67] C. McDiarmid, On the method of bounded differences, Surv. Comb. 141 (1) (1989) 148–188.
[68] G. Casella, R.L. Berger, Statistical Inference, vol. 2, Duxbury Pacific Grove, CA, 2002.
[69] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, New York, 1996.
[70] I. Steinwart, Consistency of support vector machines and other regularized kernel classifiers, IEEE Trans. Inf. Theory 51 (1) (2005) 128–142.
[71] R. Dietrich, M. Opper, H. Sompolinsky, Statistical mechanics of support vector networks, Phys. Rev. Lett. 82 (14) (1999) 2975.
[72] M. Opper, W. Kinzel, J. Kleinz, R. Nehl, On the ability of the optimal perceptron to generalise, J. Phys. A: Math. Gen. 23 (11) (1990) L581.
[73] M. Opper, Statistical mechanics of learning: generalization, in: The Handbook of Brain Theory and Neural Networks, 1995.
[74] S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T.R. Golub, J.P. Mesirov, Estimating dataset size requirements for classifying dna microarray data, J. Comput. Biol. 10 (2) (2003) 119–142.
[75] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Am. Stat. Assoc. 58 (301) (1963) 13–30.
[76] D. Anguita, A. Ghio, L. Oneto, S. Ridella, Maximal discrepancy vs. Rademacher complexity for error estimation, in: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2011.
[77] L. Oneto, A. Ghio, S. Ridella, D. Anguita, Local Rademacher complexity: sharper risk bounds with and without unlabeled samples, Neural Netw. 65 (2015) 115–125.
[78] D.A. Medler, A. Arnoldussen, J.R. Binder, M.S. Seidenberg, The Wisconsin perceptual attribute ratings database, ⟨http://www.neuro.mcw.edu/ratings/⟩, 2005.
[79] N. Chater, P. Vitányi, Simplicity: a unifying principle in cognitive science?, Trends Cogn. Sci. 7 (1) (2003) 19–22.

Mehrnoosh Vahdat is currently a Ph.D. student in Interactive and Cognitive Environments (ICE) at DITEN, University of Genoa, Italy. She follows a double-degree program with Eindhoven University of Technology, Netherlands. She is graduated from two international master programs. She received her master's degree in Interactive Media and Knowledge Environments at Tallinn University. Through this program, she had a research internship at LIRIS laboratory in INSA Lyon, France on Technology Enhanced Learning. She also holds a master degree in Digital Library Learning (DILL, offered as an Erasmus Mundus cooperative effort between Oslo University College, Tallinn University and University of Parma). Her major fields of interest are learning analytics, educational data mining, human-computer interactions, and technology enhanced.

Luca Oneto was born in Rapallo, Italy in 1986. He is currently a researcher at University of Genoa with particular interests in Machine Learning and Statistical Learning Theory. He received his bachelor degree in Electronic Engineering at the University of Genoa, Italy in 2008. He subsequently started his master studies in Electronic Engineering in the same university with focus in Intelligent Systems and Statistics. After receiving his M.Sc. degree in 2010, he started to work as a consultant for the DITEN and DIBE Departments at University of Genoa, together with other consultant activities for Mac96 and Ansaldo STS in the context of many European Projects. In 2014 he received his Ph.D. in School of Sciences and Technologies for Knowledge and Information Retrieval (University of Genoa) with the thesis 'Learning Based On Empirical Data'. Today he works as a consultant and teaches in many B.Sc. and M.Sc. courses at University of Genoa as a Researcher.

Davide Anguita received the 'Laurea' degree in Electronic Engineering and a Ph.D. degree in Computer Science and Electronic Engineering from the University of Genoa, Genoa, Italy, in 1989 and 1993, respectively. After working as a research associate at the International Computer Science Institute, Berkeley, CA, on special-purpose processors for neurocomputing, he returned to the University of Genoa. He is currently associate professor of Computer Engineering with the Department of Informatics, BioEngineering, Robotics, and Systems Engineering (DIBRIS). His current research focuses on the theory and application of kernel methods and artificial neural networks.

Mathias Funk since January 2013 is an assistant professor in the Designed Intelligence group of the Industrial Design department at the Eindhoven University of Technology (TU/e). Before that he was postdoctoral researcher in the same group for two years, and from end of 2006 until end of 2010, he did his Ph.D. research at the same university, however, in the department of Electrical Engineering, and there in the Electronic Systems group.

Matthias Rauterberg received a B.S. in Psychology (1978) at the University of Marburg (Germany), a B.A. in Philosophy (1981) and a B.S. in Computer Science (1983), a M.S. in Psychology (1981) and a M.S. in Computer Science (1986) at the University of Hamburg (Germany), and a Ph.D. in Computer Science/Mathematics (1995) at the University of Zurich (Switzerland). He was a Senior Lecturer for 'usability engineering' in Computer Science and Industrial Engineering at the Swiss Federal Institute of Technology (ETH) in Zurich, where later he was heading the Man-Machine Interaction research group (MMI). Since 1998 he is fulltime professor for 'Human Communication Technology' first at IPO, Center for User System Interaction Research, and later at the Department of Industrial Design at the Eindhoven University of Technology (TU/e, The Netherlands). From 1999 till 2001 he was director of IPO. He is now the head of the Designed Intelligence research group at the department of Industrial Design of the TU/e. He was the Swiss representative in the IFIP TC13 on 'Human Computer Interaction' (1994–2002) and the chairman of the IFIP WG13.1 on 'HCI and Education' (1998–2004). He is now the Dutch representative in the IFIP TC14 on 'Entertainment Computing' and the founding vice-chair of this TC14 (since 2006). He is elected as IFIP TC14 chair for the term 2013–2015. He was also the chair of the IFIPWG14.3 on 'Entertainment Theory' (2004–2012). He was appointed as visiting professor at Kwansei Gakuin University (Japan) (2004–2007). He is guest professor of School of Design at Jiangnan University, Wuxi, China (2012–2015). He received the German GI-HCI Award for the best Ph.D. in 1997 and the Swiss Technology Award for the BUILD-IT system in 1998. In 2007 he got the Silver Core Award from IFIP. Since 2004 he is a nominated member of the 'Cream of Science' in The Netherlands (the 200 top-level Dutch researchers) and amongst the 10 top-level TU/e scientists. He has over 400 publications in international journals, conference proceedings, books, etc. He acts also as editor and member of the editorial board of several leading international journals. He is co-editor-in-chief of the international journal 'Entertainment Computing' (Elsevier).