Probably Approximately Correct Learning - An Introduction in the Finite Case

Kevin de Haan

Edmonton, Canada
Email: kdehaan@ualberta.ca

***Abstract*—Probably Approximately Correct (PAC) learning is a tool that was first introduced in 1984 by Valiant in order to bridge the gap between computability theory and machine learning [1]. In doing so, Valiant introduced the concept of learnability to a wider group of computer scientists interested in algorithmic efficiency and leading to the combined field of computational learning theory [2].**

***Index Terms*—PAC; Learning Theory; Computability Theory; Computational Learning Theory**

I. INTRODUCTION

The Probably Approximately Correct (hereafter PAC) learning framework, first introduced by Valiant [1], is a method for determining the learnability of a problem for a machine classifier. In other words, using the PAC learning framework one can estimate whether a learning classifier will be able to output an approximately correct prediction, and under what conditions the classifier is able to operate. If a concept is known to be PAC-learnable, it means that the learning algorithm must be able to operate in polynomial time [2], and additionally provides an upper bound on the volume of training samples provided to the classifier in order to produce an acceptable prediction [2]. It can also be useful to prove that a target is not PAC-learnable, for example in the case of a cryptographic function [2].

II. DERIVING THE KEY POINTS OF PAC LEARNING

A. Terminology

PAC and indeed learning theory as a whole has some terminology that varies from traditional machine learning notation. All non-trivial symbols used in this paper can be found in the following list:

- X: the domain space
- Y: the label set space
- f: the mapping function between domain set space to label set space, $f : X \to Y$. Unknowable in its entirety.
- $\mathcal{D}$: the data distribution over X. Unknowable in its entirety.
- S: a sample taken from $\mathcal{D}$
- h: a hypothesis, also referred to as a $model$. A prediction of the real mapping function f
- $\mathcal{H}$: a finite hypothesis space
- L: a $learner$, the algorithm or device that produces a hypothesis h
- *risk*: used interchangeably with 'error rate'
- δ: the probability of getting a nonrepresentative sample S from $\mathcal{D}$

B. Background: Empirical Risk Minimization

To properly understand the theory behind PAC, it is valuable to have an understanding of Empirical Risk Minimization (hereafter ERM). The fundamental theory behind ERM is that given some $h \in \mathcal{H}$ produced from the training set S, there will be some error between h_S and the real f, demonstrated in equation 1. The true error is equal to the probability of sampling x from $\mathcal{D}$ such that prediction of h is different from that of f [3].

$$L_{\mathcal{D},f}(h) = \mathcal{D}(\{x : h(x) \neq f(x)\}) \tag{1}$$

However, because $\mathcal{D}$ and f are unknown, the full error calculation cannot simply be evaluated by the learner L [3]. The next best thing is to determine the *training error*, the error encountered by the classifier in the process of training (equation 2) [3]. The training error is equal to the real number of times that the prediction from h has been different from the real value in y, divided by the number of samples m. This is also referred to as the *empirical error* or *empirical risk* interchangeably [3].

$$L_S(h) = \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}, [m] = \{1, ..., m\} \tag{2}$$

In contrast to the real error (or *actual risk*) the value of empirical risk is available to the learner and is therefore an obvious choice for minimization - hence, the goal of *empirical risk minimization* defined in eq. 3 [3].

$$h_S \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h) \tag{3}$$

Now, what can be done with this empirical risk? By using the *realizability assumption* and the *i.i.d. assumption*, we can establish an upper bound on the error of the model.

*1) **The Realizability Assumption**:* "There exists $h^* \in \mathcal{H}$ s.t. $L_{(\mathcal{D},f)}(h^*) = 0$" [3]. This implies that given enough sets of samples S from the distribution $\mathcal{D}$, there exists some S such that $L_S(h^*) = 0$ with probability 1. This in turn suggests that for every ERM domain and hypothesis there exists an optimal h_S such that $L_S(h_S) = 0$. However, this only applies to the empirical risk - what about the real risk? And what about overfitting? This leads us to the next step:

2) ***The i.i.d. Assumption:*** "Examples in the training set are independently and identically distributed (i.i.d.) according to the distribution" $\mathcal{D}$ [3]. This means that every element $x_i \in S$ is sampled according to $\mathcal{D}$ and in turn labelled according to f, denoted by the term $S \sim \mathcal{D}^m$ such that m is the size of S [3]. While it is guaranteed that there exists some hypothesis h_S such that the empirical risk is 0, as long as S is nonrepresentative this zeroed empirical risk is simply a result of overfitting. The key feature of this assumption is that as m increases, the likelihood that S is properly representing $\mathcal{D}$ (and that h_S reflects f instead of overfitting) also increases [3].

From these two assumptions, we can establish that there is some chance less than 1 for any properly sampled S to be a misrepresentation of $\mathcal{D}$, labelled δ, and consequentially we are able to formally name $(1-\delta)$ as the *confidence parameter* of a prediction [3].

In addition to describing the chance for the sample S to be nonrepresentative, we can also choose a value to describe the quality of the hypothesis h. This is the *accuracy parameter* ϵ, the threshold of error or risk above which the model is considered to be to inaccurate to be useful [3]. I.e, a risk $L_{(\mathcal{D},f)}(h_S) > \epsilon$ is a failure, while $L_{(\mathcal{D},f)}(h_S) \leq \epsilon$ is ***approximately correct*** [3]. We can now describe the chance that a sampled set S will be overfit based on the size m of the sample S, as seen in eq. 4 [3].

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \tag{4}$$

The next step is to set an upper bound on the above value. Let $\mathcal{H}_{\mathcal{B}}$ be the set of hypotheses which will lead to a failure - that is, $\mathcal{H}_{\mathcal{B}} = \{h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \epsilon\}$ [3]. Additionally, we describe the set of misleading samples $M = \{S|_x : \exists h \in \mathcal{H}_{\mathcal{B}}, L_S(h) = 0\}$ as the set of samples which produce zero empirical risk but do not correctly represent f [3]. This can equivalently be represented as $\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\} \subseteq M$, an in turn allows us to rewrite M as in equation 5 [3].

$$M = \bigcup_{h \in \mathcal{H}_{\mathcal{B}}} \{S|_x : L_S(h) = 0\} \tag{5}$$

This leads to the formalized upper bound of equation 6 [3].

$$\begin{gathered} \mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \leq \mathcal{D}^m(M) \\ \mathcal{D}^m(M) = \mathcal{D}^m(\cup_{h \in \mathcal{H}_{\mathcal{B}}}\{S|_x : L_S(h) = 0\}) \end{gathered} \tag{6}$$

Next, using the property of the *union bound* - formally, $\mathcal{D}(A \cup B) \leq \mathcal{D}(A) + \mathcal{D}(B)$ - we can transform the right hand side of the above equation [3]:

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \leq \sum_{h \in \mathcal{H}_{\mathcal{B}}} \mathcal{D}^m(\{S|_x : L_s(h) = 0\}) \tag{7}$$

For each summand in the new right-hand side of the equation, we attempt to find a bound. Notice that the event $L_S(h) = 0$ is the same as $\forall i, h(x_i) = f(x_i)$, and since all training set examples are sampled i.i.d we obtain equation 8 [3].

$$\begin{aligned} \mathcal{D}^m(\{S|_x : L_s(h) = 0\} &= \mathcal{D}^m(\{S|_x : \forall i, h(x_i) = f(x_i)\}) \\ &= \prod_{i=1}^{m} \mathcal{D}(\{x_i : h(x_i) = f(x_i)\}) \end{aligned} \tag{8}$$

Because $h \in \mathcal{H}_{\mathcal{B}}$, it follows that each individual sampling of an element x_i from $\mathcal{D}$ is thus [3]

$$\mathcal{D}(\{x_i : h(x_i) = f(x_i)\}) = 1 - L_{(\mathcal{D},f)}(h) \leq 1 - \epsilon \tag{9}$$

Taking equations 8 and 9 in combination with the inequality $1 - \epsilon \leq e^{-\epsilon}$ (a simple, provable inequality used for simplification), we obtain the following [3]:

$$\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) \leq (1-\epsilon)^m \leq e^{-\epsilon m} \tag{10}$$

And by combining equations 7 and 10 we produce the final upper bound of [3]:

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \leq |\mathcal{H}_{\mathcal{B}}|e^{-\epsilon m} \leq |\mathcal{H}|e^{-\epsilon m} \tag{11}$$

Which leads to the more easily digestible conclusion of ERM:

Corollary 1: Given a finite hypothesis class $\mathcal{H}$, take $\delta \in (0,1)$ and $\epsilon > 0$. Given an integer value m that satisfies

$$m \geq \frac{log(\frac{|\mathcal{H}|}{\delta})}{\epsilon}$$

then for any labeling function f and data distribution $\mathcal{D}$ for which the realizability function holds, as long as the sample S of size m is selected i.i.d then it holds that

$$L_{(\mathcal{D},f)}(h_S) \leq \epsilon$$

with probability of at least $1 - \delta$ [3].

In other words, if the assumptions behind ERM are satisfied it becomes possible to determine if a model will ***probably*** $(1-\delta)$ be ***approximately correct*** $(L_{(\mathcal{D},f)}(h_S) \leq \epsilon)$ [3].

C. *Bringing it Together*

III. Conclusion and Critical Analysis

References

[1] L. G. Valiant, "A theory of the learnable," *Commun. ACM*, vol. 27, no. 11, p. 1134–1142, Nov. 1984. [Online]. Available: https://doi.org/10.1145/1968.1972

[2] D. Haussler, "Part 1: Overview of the probably approximately correct (pac) learning framework," 1995.

[3] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. USA: Cambridge University Press, 2014.