An Introduction to Computational Learning Theory: PAC Learning and VC-Dimensionality

Alana Jaskir
Applied Statistics Reading Group

May 2019

1 Introduction

Probably Approximately Correct (PAC) Learning, proposed by Leslie Valiant in 1984 in his paper "The Theory of the Learnable", laid important groundwork for the theoretical analysis of machine learning. The paper took the first step in identifying models which "shed light on the limits of what can be learned, just as computability does on what can be computed." (Excitingly for us cognitive scientists, the introduction of the paper is motivated heavily by implicit learning in human cognition).

Going forward, we assume that a learner uses an *Empirical Risk Minimzation* rule to learn a "concept", a mapping from domain X to {0,1}. That is, using sample data drawn from an unknown distribution, $\mathcal{S} \sim \mathcal{D}^m$, where m is the number of examples, the learner selects the hypothesis $h_S \in H$ (where $\mathcal{H}$ is the hypothesis class, e.g. a set of halfspaces) where sample loss $\mathcal{L}_S(h_S) \leq \mathcal{L}_S(h_i), \forall h_i \in \mathcal{H}$.

The goal of these notes is to detail PAC Leaning, agnostic PAC Learning, and VC-Dimensionality, the key definitional components of The Fundamental Theorem of Statistical Learning. Definitions and corollaries pulled from "Understanding Machine Learning: From Theory to Algorithms" by Shai Shalev-Shwartz and Shai Ben-David. Content also pulled from CSCI1420: Machine Learning with Stephen Bach.

2 The Fundamental Theorem of Statistical Learning

Qualitative Definition

THEOREM 6.7 (The Fundamental Theorem of Statistical Learning) *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$ and let the loss function be the $0-1$ loss. Then, the following are equivalent:*

1. *$\mathcal{H}$ has the uniform convergence property.*
2. *Any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$.*
3. *$\mathcal{H}$ is agnostic PAC learnable.*
4. *$\mathcal{H}$ is PAC learnable.*
5. *Any ERM rule is a successful PAC learner for $\mathcal{H}$.*
6. *$\mathcal{H}$ has a finite VC-dimension.*

Quantitative Definition

THEOREM 6.8 (The Fundamental Theorem of Statistical Learning – Quantitative Version) *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$ and let the loss function be the $0-1$ loss. Assume that $\mathrm{VCdim}(\mathcal{H}) = d < \infty$. Then, there are absolute constants C_1, C_2 such that:*

1. *$\mathcal{H}$ has the uniform convergence property with sample complexity*
$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m^{UC}_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$
2. *$\mathcal{H}$ is agnostic PAC learnable with sample complexity*
$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq \ m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$
3. *$\mathcal{H}$ is PAC learnable with sample complexity*
$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

3 PAC Learning

3.1 Definition

The hope of any learner is that a learned hypothesis produces low generalization error, $\mathcal{L}_D(h_S)$. However, the world is not so kind. Can we guarantee that we "probably" have a representative sample? Additionally, can we place any guarantees on this generalization error (i.e. that we are "approximately" correct)?

REALIZABILITY We first begin with an assumption of realizability. An informal definition is below. This also implies that each data point has one true labeling.

$$\exists h^* \in \mathcal{H} : L_D(h^*) = 0$$

Formally, we can say that a hypothesis class is PAC Learnable with the following definition:

> DEFINITION 3.1 (PAC Learnability) A hypothesis class $\mathcal{H}$ is PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ over X, and for every labeling function $f : X \to \{0,1\}$, if the realizable assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, then when running the learning algorithm on $m \geq m_h(\epsilon, \delta)$ i.i.d. examples generated by D and labeled by f, the algorithm returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the examples), $L_{(D,f)(h)} \leq \epsilon$.

We refer to the function $m_{\mathcal{H}}$ as the sample complexity. For what hypothesis classes is it possible to construct such a function?

3.2 Finite Hypothesis Classes

We will show that if our hypothesis class is finite, $|\mathcal{H}| < \infty$, we can define a sample complexity function for $\mathcal{H}$. Assuming i.i.d and realizability still hold, finite hypothesis classes are then PAC Learnable.

GOAL Define sample complexity to determine numner of samples to be able to arbitrarily set an upper bound, δ, for the following probability:

$$D^m(\{S|_x : L_D(h_s) > \epsilon\}) \tag{1}$$

where $S|_x = (x_1, x_2, x_3, ..., x_m)$ is the instance of the training set. We arbitrarily specify ϵ, our error tolerance

Since we assume realizability, this implies that $L_S(h_S) = 0$ and $L_D(h_S) > \epsilon$. That is, a "bad" hypothesis H_B ($\forall h \in H_B, L_D(h) > \epsilon$) perfectly predicts the sample data. A way to upperbound our goal then is to ask "What's the probability we get misleading sample data?"

$$M = \bigcup_{h \in H_B} \{S|_x : L_S(h) = 0\} \tag{2}$$

$$\begin{aligned} D^m(\{S|_x : L_D(h_s) > \epsilon\} &\leq D^m(M) && (3)\\ &= D^m(\bigcup_{h \in H_B} \{S|_x : L_S(h) = 0\}) && (4)\\ &\leq \sum_{h \in H_B} D^m(\{S|_x : L_S(h) = 0\}) \qquad \text{(Union Bound)} && (5) \end{aligned}$$

At this point, we are summing over the "Probability of Perfection", or the probability that a bad hypothesis is perfect on each training data. Because examples are i.i.d. we obtain the following:

$$\begin{aligned} D^m(\{S|_x : L_S(h) = 0\}) &= D^m(\{S_x : \forall i \in [m], h(x_i) = y_i) && (6)\\ &= \prod_{i=1}^{m} D(\{x_i : h(x_i) = y_i\}) \qquad \text{(i.i.d. samples)} && (7)\\ &\leq \prod_{i=1}^{m} (1 - \epsilon) \qquad \text{(by definition)} && (8)\\ &\leq e^{-\epsilon m} \qquad (1 - \epsilon \leq e^{-\epsilon}) && (9) \end{aligned}$$

Plugging this into equation 5:

$$\begin{aligned} D^m(\{S|_x : L_D(h_s) > \epsilon\} &\leq \sum_{h \in H_B} e^{-\epsilon m} && (10)\\ &= |H_B| e^{-\epsilon m} && (11)\\ &\leq |H| e^{-\epsilon m} \qquad (H_B \subseteq H) && (12) \end{aligned}$$

We have then that:

$$D^m(\{S|_x : L_D(h_s) > \epsilon\}) \leq |H|e^{-\epsilon m} \leq \delta \tag{13}$$

To find out our sample complexity, we rearrange the last two terms then to solve for m which satisfies this equation:

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

Since we need the number of examples *at least* that of the second side of the equation, we can define the following sample complexity function:

$$m_H(\epsilon, \delta) \leq \left\lceil \frac{log(|H|/\delta)}{\epsilon} \right\rceil \tag{14}$$

Intuitively, this form makes sense - the larger the hypothesis class or the lower the probability of error, the more samples you need. However, the less tolerant you are of error, the more examples you need.

3.3 Agnostic PAC Learner

Now let us consider the case where the realizability assumption does not hold. That is:

$$\nexists h^* \in \mathcal{H} : L_D(h^*) = 0$$

This may be due to ill-suited hypothesis class or label noise (i.e. one data point can have multiple labels with differing probabilities). We still assume i.i.d. data and finite hypothesis class.

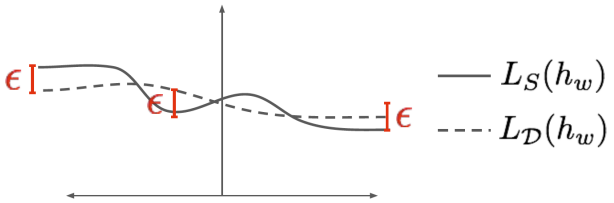
Rather than bounding total error on the distribution $\mathcal{D}$, Agnostic PAC learning instead bounds error relative to the best hypothesis in the proposed hypothesis class.

> DEFINITION (Agnostic PAC Learning) A hypothesis class $\mathcal{H}$ is Agnostic PAC learnable if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$, with i.i.d. examples with probability of at least $1 - \delta$ (over the choice of the examples), the learning algorithm returns h such that $L_{(D,f)(h)} \leq \min_{h' \in \mathcal{H}} L_D(h') + \epsilon$

Having a training set $\mathcal{S}$ that is *epsilon-representative* ensures by *Lemma 4.2* that any output h of ERM is agnostic PAC.

DEFINITION 4.1 (ϵ-representative sample) A training set S is called ϵ-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$) if

$$\forall h \in \mathcal{H}, \quad |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon.$$



We now are set up to define the uniform convergence property of an hypothesis class.

DEFINITION 4.3 (Uniform Convergence) We say that a hypothesis class $\mathcal{H}$ has the *uniform convergence property* (w.r.t. a domain Z and a loss function ℓ) if there exists a function $m_{\mathcal{H}}^{\text{UC}} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{\text{UC}}(\epsilon, \delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then, with probability of at least $1 - \delta$, S is ϵ-representative.

Defining a sample complexity for Uniform Converge is sufficient for proving agnostic PAC learnability.

COROLLARY 4.4 *If a class $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{UC}$ then the class is agnostically PAC learnable with the sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$. Furthermore, in that case, the* $\text{ERM}_{\mathcal{H}}$ *paradigm is a successful agnostic PAC learner for $\mathcal{H}$.*

For any finite hypothesis class, we can define the sample complexity to ensure uniform convergence as follows (see textbook for proof):

$$m_{\mathcal{H}}^{\text{UC}}(\epsilon, \delta) \leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil$$

In conclusion, any finite hypothesis class is agnostic PAC learnable via ERM with respect to a loss function with range [0,1], with the following sample complexity:

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{\text{UC}}(\epsilon/2, \delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

3.4 PAC and No-Free-Lunch Theorem

COROLLARY 5.2 *Let X be an infinite domain set and let H be the set of all functions X to {0,1}. Then H is not PAC Learnable*

Without restricting our hypothesis class, for any learning algorithm, and adversary can construct a distribution for which the

4 VC-Dimensionality

Are infinite hypothesis classes learnable? Consider the following example:

[to be completed]

5 Notes on Learnability and Popular Hypothesis Classes

6 Fun Facts

- Leslie Valiant's son teaches at Brown
- Also Leslie Valiant wrote a book called *Circuits of the Mind*

7 Miscellaneous Links

- Wiki Link on Learning Theory
- Association for Computational Learning