PAC Learnability

Uri Shaham

May 31, 2026

1 The Supervised learning setup

- Domain set: arbitrary $\mathcal{X}$.
- Label set: $\mathcal{Y} = \{0, 1\}$.
- Training data: $S_n = (x_1, y_1, \ldots, (x_n, y_n)$.
- We assume that the elements of S_n are drawn i.i.d from an unknown distribution $\mathcal{D}$.
- A learning algorithm produces a predictor (also called a hypothesis) $h : \mathcal{X} \to \mathcal{Y}$
- The training error (empirical risk) of a hypothesis h is

$$L_{S_n}(h) := \sum_{i=1}^{n} \frac{1}{n} \mathbb{1}_{h(x_i) \neq y_i}.$$

- The generalization error of a hypothesis h is

$$L_D(h) := \mathbb{E}_{(x,y)\sim\mathcal{D}} \mathbb{1}_{h(x)\neq y} = \mathcal{P}_D\, h(x) \neq y = D(\{(x, y) : h(x) \neq y\}).$$

- A collection of hypotheses $\mathcal{H}$ is called a hypothesis class.
- The paradigm of searching hypothesis $h \in \mathcal{H}$ that minimizes the empirical risk is called Empirical Risk Minimization (ERM). That is, given a hypothesis class $\mathcal{H}$, ERM predictor returns $h_{S_n} = \mathrm{ERM}(\mathcal{H}) \in \arg\min_{h\in\mathcal{H}} L_{S_n(h)}$. When the training set size is not of interest we also use the shorter notation h_S.

While sounding natural, ERM might desperately fail, as there may be many hypotheses that minimize the empirical risk, but not all of them are good in terms of generalization error. This is called *overfitting*. We will therefore look for conditions under which the ERM predictor has good performance, also in terms of generalization error. To do so, we aim to characterize hypothesis classes $\mathcal{H}$ that give us guarantees on the connection between the empirical risk and the generalization error.

2 Finite hypothesis classes, realizable case

Definition 2.1 (The realizability assumption)**.** *There exists $h^* \in \mathcal{H}$ such that $L_D(h) = 0$.*

Since S_n is picked by a random process, the predictor h_S is random as well, and consequently the risk $L_D(h_S)$, which we thus treat as a random variable. It is not realistic to expect that a finite set would always suffice to obtain a good predictor, as there is always some chance to obtain a very non-representative sample S under $\mathcal{D}$. We denote this probability of a bad sample by δ, and thus $(1-\delta)$ is the confidence we have in our predictor.

In addition, we cannot guarantee that even a good predictor is always free from mistakes. We thus introduce an accuracy parameter ϵ, and we say that we would like our prediction to have a small enough generalization error, i.e., $L_D(h) \leq \epsilon$.

This leads us to the definition of PAC (Probably Approximately Correct) learnability.

Definition 2.2 (PAC learnability - realizable case)**.** *A hypothesis class $\mathcal{H}$ is PAC learnable if for every $\delta, \epsilon \in [0,1]$ and for every distribution $\mathcal{D}$ over $\mathcal{X} \times Y$, there exist a learning algorithm training set size $n = n(\epsilon, \delta)$ so that when running the learning algoritm on training data of size n from $\mathcal{D}$, the algorithm will return $h \in \mathcal{H}$ such that with probability at least $\mathbb{1} - \delta$ (over the choice of S_n),*

$$L_D(h) \leq \epsilon.$$

We would like to upper bound the probability of a bad sample S_n, that is, a sample S_n drawn from n iid draws from $\mathcal{D}$, which will lead to a choice of a bad hypothesis h_S, i.e., such that

$$L_D(h_S) > \epsilon.$$

We denote this probability by $\mathcal{D}^n(\{S : L_D(h_S) > \epsilon\})$.

Let $\mathcal{H}_B \subseteq \mathcal{H}$ denote the subset of bad hypotheses, i.e.,

$$\mathcal{H}_B = \{h \in \mathcal{H} : L_D(h) > \epsilon\}.$$

Let M denote the set of bad samples, that is,

$$M = \{S : \exists h \in \mathcal{H}_B, \text{Ł}_S = 0\}.$$

That is, for every sample in M, there exists a bad hypothesis h which looks good on S.

Recall that we want to bound the probability of a bad sample, i.e., a sample S such that $L_D(h_S) > \epsilon$. Since we assume realizability, this can only happen if $L_S(h_S) = 0$. Hence, we can write M as

$$M = \bigcup_{h \in \mathcal{H}_B} \{S : L_S(h) = 0\}.$$

We have thus shown that

$$\mathcal{D}^n(\{S : L_D(h_S) > \epsilon\}) \leq \mathcal{D}^n(M) = \mathcal{D}^n(\cup_{h \in \mathcal{H}_B} \{S : L_S(h) = 0\}). \tag{1}$$

Here, we make use of a basic fact from probability, called Union Bound.

Lemma 2.3 (Union Bound)**.** *For any two sets A, B and a distribution $\mathcal{D}$ we have*

$$D(A \cup B) \leq \mathcal{D}(A) + D(B).$$

Applying the union bound to the RHS of equation (1) gives

$$\mathcal{D}^n(\{S : L_D(h_S) > \epsilon\}) \leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^n(\{S : L_S(h) = 0\}). \tag{2}$$

Since the training examples are obtained i.i.d, each summand on the RHS equals

$$\prod_{i=1}^{n} \mathcal{D}(\{(x_i, y_i) : h(x_i) = y_i\}) \tag{3}$$

. For each pair (x_i, y_i) drawn from $\mathcal{D}$ we can write

$$\mathcal{D}(\{(x_i, y_i) : h(x_i) = y_i\}) = 1 - L_D(h) \leq 1 - \epsilon,$$

where the inequality follows since $h \in HH_B$ is bad. Combining this with equation (3), and using the inequality $1 - \epsilon \leq e^{-\epsilon}$, we obtain that for every $h \in \mathcal{H}_B$,

$$\mathcal{D}^n(\{S : L_S(h) = 0\}) \leq e^{-\epsilon n}.$$

combining this with equation 2, we thus get

$$\mathcal{D}^n(\{S : L_D(h_S) > \epsilon\}) \leq |\,\mathcal{H}_B\,| e^{-\epsilon n} \leq |\,\mathcal{H}\,| e^{-\epsilon n} \tag{4}$$

.

Corollary 2.4. *Any finite hypothesis class $\mathcal{H}$ is PAC learnable via ERM.*

Proof. Choosing n such that $n \geq \frac{\log(|\,\mathcal{H}\,|\delta)}{\epsilon}$ gives, by equation (4), $\mathcal{D}^n(\{S : L_D(h_S) \leq \epsilon\}) \geq 1 - \delta$. □

3 Learning via uniform convergence

We begin with extending the definition of PAC learning to cases which are not necessarily realizable, so that there does not have to be a hypothesis $h \in \mathcal{H}$ with $L_S(h) = 0$.

This case is a slight generalization of the realizable case, this time with a more general definition of PAC learnability.

Definition 3.1 (PAC learnability - agnostic case)**.** *A hypothesis class $\mathcal{H}$ is PAC learnable if for every $\delta, \epsilon \in [0, 1]$ and for every distribution $\mathcal{D}$ over $\mathcal{X} \times Y$, there exist a learning algorithm training set size $n = n(\epsilon, \delta)$ so that when running the learning algorithm on training data of size n from $\mathcal{D}$, the algorithm will return $h \in \mathcal{H}$ such that with probability at least $\mathbb{1} - \delta$ (over the choice of S_n),*

$$L_D(h) \leq \min_{h' \in \mathcal{H}} L_D(h') + \epsilon.$$

Next, we introduce the following lemma.

Lemma 3.2. *Let $h_S = \arg\min_{h \in \mathcal{H}} L_{S_n}(h)$ be the hypothesis that minimizes the train error (found via ERM). Let $h^* = \arg\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$. Then $L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(h^*) \leq 2 \sup_{h \in \mathcal{H}} |L_{S_n}(h) - L_{\mathcal{D}}(h)|$.*

Before we prove the lemma, let's understand what it tells us.

- Denote the generalization error $|L_{S_n}(h) - L_{\mathcal{D}}(h)|$ of a hypothesis h by $\epsilon(h)$.

- Assume we can *uniformly* bound this error for all $h \in \mathcal{H}$, i.e., $\epsilon := \sup_{h\in\mathcal{H}} \epsilon(h)$.
- That is, for every choice of $\epsilon, \delta \in (0,1)$ there exists training set size n such that with probability $1-\delta$, S_n is a representative sample and $\epsilon(h) \leq \epsilon$.
- Then $\mathcal{H}$ is (agnostically) PAC-learnable.
- Hence, (once proving the lemma) we will focus on deriving conditions for this uniform convergence.

Proof. Let $\epsilon := \sup_{h\in\mathcal{H}} |L_{S_n}(h) - L_{\mathcal{D}}(h)|$. Then

$$\begin{aligned} l_{\mathcal{D}}(h_S) &\leq L_{S_n}(h_S) + \epsilon, \text{ (since } h_S \in \mathcal{H}) \\ &\leq L_{S_n}(h^*) + \epsilon, \text{ (since } h_S \text{ minimizes } l_{S_n}) \\ &\leq L_{\mathcal{D}}(h^*) + 2\epsilon, \text{ (since } h^* \in \mathcal{H}). \end{aligned}$$

□

4 Finite hypothesis classes, agnostic case

We will now prove that finite hypothesis classes are agnostically PAC learnable, using a similar approach to the realizable case, utilizing a measure concentration argument.

For any choice of $\delta, \epsilon \in (0,1)$, we wish to find a sample size $n = n(\epsilon, \delta)$ such that for any distribution $\mathcal{D}$, with probability at least $1-\delta$ for any $h \in \mathcal{H}$, $|L_{S_n}(h) - L_{\mathcal{D}}(h)| \leq \epsilon)$ (i.e., $\mathcal{H}$ has a uniform convergence property). equivalently, we need to show that

$$\mathcal{D}^n(S_n : \exists h \in \mathcal{H}, |L_{S_n}(h) - L_{\mathcal{D}}(h)| \geq \epsilon) \leq \delta.$$

We can write $S_n : \exists h \in \mathcal{H}, |L_{S_n}(h) - L_{\mathcal{D}}(h)| \geq \epsilon = \bigcup_{h\in\mathcal{H}} \{|L_{S_n}(h) - L_{\mathcal{D}}(h)| \geq \epsilon\}$ and apply the union bound to get

$$\mathcal{D}^n(S_n : \exists h \in \mathcal{H}, |L_{S_n}(h) - L_{\mathcal{D}}(h)| \geq \epsilon) \leq \sum_{h\in\mathcal{H}} \mathcal{D}^n(\{|L_{S_n}(h) - L_{\mathcal{D}}(h)| \geq \epsilon\}). \tag{5}$$

Since $L_{\mathcal{D}}(h)$ is the expectation of the random variable $L_S(h)$, the term $|L_{S_n}(h) - L_{\mathcal{D}}(h)|$ quantifies the deviation of a random variable from its expectation. Measure concentration tools provide probabilistic guarantees about such deviation. Here, we use the Hoeffding inequality.

Theorem 4.1 (Hoeffding inequality)**.** *Let $X_1, \ldots, X_n \in [a,b]$ be n i.i.d random variables with mean μ. Then for any $\epsilon > 0$,*

$$\Pr|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| > \epsilon \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

To utilize Hoeffding's inequality, we write

$$L_{S_n}(h) = \frac{1}{n}\ell(h, x_i),$$

where $\ell(h, x_i) = 1$ if $h(x_i) = y_i$, and $\ell(h, x_i) = 0$ otherwise. Note that $L_{\mathcal{D}}$ is also the expectation of $\ell(h, x_i)$. Plugging this in equation (5) then gives

$$\mathcal{D}^n(S_n : \exists h \in \mathcal{H}, |L_{S_n}(h) - L_{\mathcal{D}}(h)| \geq \epsilon) \leq 2|\mathcal{H}|\exp(-2n\epsilon^2). \tag{6}$$

As before, equating the right-hand side probability to δ, and picking $n \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2}$ gives

$$\mathcal{D}^n(S_n : \exists h \in \mathcal{H}, |L_{S_n}(h) - L_{\mathcal{D}}(h)| \geq \epsilon) \leq \delta, \tag{7}$$

which is uniform convergence. Hence Any finite hypothesis class $\mathcal{H}$ is agnostically PAC learnable.

5 Reading

UML ch. 2-3