Non-Uniform Learnability

Uri Shaham

May 31, 2026

1 Non-Uniform Learnability

Recall: We should that hypothesis classes with finite VC dimension are PAC learnable, via the uniform convergence property: 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) That is, we derived sample sizes which depend on ϵ, δ and hold uniformly for all $h \in \mathcal{H}$.

We now relax this requirement, and allow the sample sizes to depend also on the specific h we compete against. This will allow us to extend the notion of learnability also to broader hypothesis classes.

Definition 1.1. *We say that a hypothesis h is (ϵ, δ)-competitive with hypothesis h' if with probability at least $1-\delta$,*

$$L_{\mathcal{D}}(h) \leq L_{\mathcal{D}}(h') + \epsilon.$$

This leads us to define the concept of nonuniform learnability:

Definition 1.2 (non-uniform learnability)**.** *A hypotesis class $\mathcal{H}$ is non-uniformly learnable if there exist a learning algorithm A and a sample size $m = m(\epsilon, \delta, h)$ such that for every $0 < \epsilon, \delta < 1$, for every $h \in \mathcal{H}$ and for every distribution $\mathcal{D}$, applying A on a training set $S \sim \mathcal{D}^n$, it will hold with probability at least $1-\delta$ that*

$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon.$$

Remark 1.3. *Observe that if $\mathcal{H}$ is agnostic PAC learnable, then it is also nonuniformly learnable.*

2 Structural Risk Minimization

In the SRM setting, we assume that the hypothesis class is a countable union of subclasses $\mathcal{H} = \cup_{n \in \mathbb{N}} \mathcal{H}_n$, and specify a weight function $w : \mathbb{N} \to [0,1]$ assigning a weight $w(n)$ for each class $\mathcal{H}_n$, so that a higher weight reflects a stronger preference (corresponding to the belief of containing the correct hypothesis). We furthermore assume that for each n, $\mathcal{H}_n$ has the uniform convergence property, that is, for all n there exists a sample complexity function $m_{\mathcal{H}_n} : (0,1)^2 \to \mathbb{N}$, so that for all $\epsilon, \delta \in (0,1)$ and for any distribution $\mathcal{D}$, if S is a sample of n iid samples from $\mathcal{D}$, then with probability at least $1-\delta$ for every $h \in \mathcal{H}_n$

$$|L_S(h) - L_{\mathcal{D}}(h)| < \epsilon.$$

That is, fixing ϵ and δ we can get the required sample size m in order for the error to be small (for all $h \in \mathcal{H}_n$). Similarly , we can turn things around and fix δ and m and get the best accuracy ϵ that is achievable for m, δ:

$$\epsilon_n(m, \delta) = \min\{\epsilon \in (0,1) : m_{\mathcal{H}_n}(\epsilon, \delta) \leq m\}.$$

Equipped with this definition, we can state a (non-uniform) generalization bound (that is, such that depends on n).

Theorem 2.1. *Let the weight function w be such that $\sum_{n\in\mathbb{N}} w(n) \leq 1$. Assume that $\mathcal{H} = \cup_{n\in\mathbb{N}} \mathcal{H}_n$ with each $\mathcal{H}_n$ satisfying the uniform convergence property with sample complexity function $m_{\mathcal{H}_n}$. Then for every $\delta \in (0,1)$ and every distribution $\mathcal{D}$, with probability of at least $1-\delta$ over sample S of m iid samples from $\mathcal{D}$, for every $n \in \mathbb{N}$ and every $h \in \mathcal{H}_n$*

$$|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon_n(m, w(n)\delta).$$

Proof. For each n, from uniform convergence of $\mathcal{H}_n$ we have that for each $h \in \mathcal{H}_n$, with probability at least $1 - w(n)\delta$

$$|L_S(h) - L_{\mathcal{D}}(h)| < \epsilon_n(m, w(n)\delta).$$

Then from union bound over all n, and since $\sum_{n\in\mathbb{N}} w(n) \leq 1$ we get that with probability at least $1-\delta$, $|L_S(h) - \text{Ł}_{\mathcal{D}}(h)| \leq \epsilon_n(m, w(n)\delta)$, as required.

□

For each hypothesis $h \in \mathcal{H}$ denote $n(h) = \min\{n : h \in \mathcal{H}_n\}$. Then the above theorem implies that for all $h \in \mathcal{H}$,

$$L_{\mathcal{D}}(h) \leq L_S(h) + \min_{n:h\in\mathcal{H}_n} \epsilon_{n(h)}(m, w(n(h))\delta) \tag{1}$$

The *SRM learning rule* works by minimizing the right-hand side: it selects a hypothesis h that minimizes $\leq L_S(h) + \epsilon_{n(h)}(m, w(n(h))\delta)$. Note that here, along the empirical error, we bias the hypothesis selection towards classes with smaller $\epsilon_{n(h)}(m, w(n(h)))\delta$ (that is, with larger $w(n(h))$ in order to improve the estimation error.

The following theorem states that a countable union of uniformly converging classes is nonuniformly learnable via the SRM rule.

Theorem 2.2. *Let the weight function w be such that $\sum_{n\in\mathbb{N}} w(n) \leq 1$. Assume that $\mathcal{H} = \cup_{n\in\mathbb{N}} \mathcal{H}_n$ with each $\mathcal{H}_n$ satisfying the uniform convergence property with sample complexity function $m_{\mathcal{H}_n}$. Then $\mathcal{H}$ is non-uniformly learnable via the SRM rule, and sample complexity rate*

$$m(\epsilon, \delta, h) \leq m_{\mathcal{H}_n}(\epsilon 2, w(n)\delta).$$

Proof. Pick $h \in \mathcal{H}, \epsilon, \delta$. Let $m \geq m_{\mathcal{H}_{n(h)}}(\epsilon, w(n(n))\delta)$. Then from the previous theorem, with probability at least $1-\delta$ over the choice of sample S of size m, for all $h' \in \mathcal{H}$

$$L_{\mathcal{D}}(h') \leq L_S + \epsilon_{n(h')}(m, w(n(h')\delta).$$

In particular, this holds for $h' = \text{SRM}(S)$ that is returned by the SRM rule. Therefore,

$$L_S(\text{SRM}(S)) \leq L_S(h) + \epsilon_{n(h)(m,w(n(h))\delta)}.$$

Taking $m > m_{\mathcal{H}_{n(h)}}(\epsilon/2, \delta))$ clearly gives (i) $\epsilon_{n(h)}(m, w(n(h))\delta) \leq \epsilon/2$ (from the definition of ϵ_n, and also (ii) $L_S(h) \leq \text{Ł}_{\mathcal{D}}(h) + \epsilon/2$. Together, this implies

$$L_S(\text{SRM}(S) \leq L_{\mathcal{D}}(\text{SRM}(S)) + \epsilon.$$

□

The previous theorem shows that a countable union of uniformly convergent hypothesis classes is non-uniformly learnable. We now prove the opposite direction.

Theorem 2.3. *Let $\mathcal{H}$ ne non-uniformly learnable. Then $\mathcal{H} = \cup_{n\in\mathbb{N}} \mathcal{H}_n$ with each $\mathcal{H}_n$ satisfying the uniform convergence property.*

Proof. Pick some arbitrary ϵ_0 (e.g., 1/8), δ_0 (e.g., 1/7), and define $\mathcal{H}_n = \{h \in \mathcal{H} : m(\epsilon_0, \delta_0, h)\} \leq n\}$. Then clearly $\mathcal{H} = \cup_{n\in\mathbb{N}} \mathcal{H}_n$. Assume by contradiction that for some n, $\mathcal{H}_n$ is not uniformly convergent. Then by the fundamental theorem of PAC learning, it has an infinite VC dimension. IN particular, $\mathcal{H}_n$ shatters some set C of size $2n$. Apply NFL to the domain C: for any algorithm A there is a distribution $\mathcal{D}$ such that with probability at least δ_0 over the choice of a sample S of size n, $L_{\mathcal{D}}(A(s)) \leq \epsilon_0$. Furthermore since C is shattered by $\mathcal{H}_n$, there exists $h \in \mathcal{H}_n$ with $L_{\mathcal{D}}(h) = 0$. This is a contradiction to the definition of $\mathcal{H}_n$, which completes the proof.

□

3 Minimum Description Length

If $\mathcal{H}$ is countable, we can view it as a union of singleton classes (and each such class is clearly uniformly convergent). One way to assign weights in such case to fix some description language (e.g., python, English, some mathematical convention) and use length of the description of each hypothesis in that language. Specifically, let Σ be an alphabet, and let Σ^* be the set of all finite strings over this alphabet. Denote the length of a string (that is, the size of its description) $\sigma \in \Sigma^*$ by $|\sigma|$. We will also restrict our attention to languages which are frefix-free, that is, no string σ is a prefix of another string σ'. The famous Kraft inequality says that if $S \subseteq \{0,1\}^*$ is a prefix-free language, then

$$\sum_{\sigma\in S} \frac{1}{2^{|\sigma|}} \leq 1.$$

(To prove this, consider the generative process of tossing an unibased coin) to get each character, so that the probability to obtain each sequence σ is $\frac{1}{2^{|\sigma|}}$.

This property immediately suggest the description length as a weight function for SRM, as follows. From Hoeffding inequality we get for every singleton $\{h\}$

$$\Pr(|\mathbb{E}_S(h) - \mathbb{E}_{\mathcal{D}}(h)| \geq \epsilon) \leq 2\exp(-2m\epsilon^2).$$

Taking the left hand side as δ and solving for ϵ we get the following expression for ϵ_n:

$$\epsilon_n(\delta, m) = \sqrt{\frac{\ln(2/\delta)}{2m}}.$$

Plugging this in equation (1), we get

$$\begin{aligned} L_{\mathcal{D}}(h) &\leq L_S(h) + \min_{n:h\in\mathcal{H}_n} \epsilon_{n(h)}(m, w(n(h))\delta) \\ &= L_S(h) + \sqrt{\frac{\ln(2/(2^{-|h|}\delta))}{2m}} \\ &\leq L_S(h) + \sqrt{\frac{|h| + \ln(2/\delta)}{2m}}. \end{aligned} \tag{2}$$

Equation (2) gives us what is known as the "MDL learning rule": select h that minimizes the sum of training error a bias term, favoring short descriptions:

$$\text{return } h \in \arg\min_{h \in \mathcal{H}} L_S(h) + \sqrt{\frac{|h| + \ln(2/\delta)}{2m}}.$$

4 Reading

UML ch. 7