CSCI699: Theory of Machine Learning Fall 2021

Lecture 2: PAC Learnability

Instructor: Vatsal Sharan *Scribes: Ta-Yang Wang & Siddartha Devic*

Today

- PAC bound for finite hypothesis classes
- Impossibility of PAC-learning classes which are too rich (No-free lunch theorem)
- Agnostic PAC learning
- Uniform convergence

Review

Recall our definition of PAC learning.

Definition (PAC-learnability)**.** *A hypothesis class $\mathcal{H}$ is PAC-learnable if there exists a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$, every distribution D over X, and every $h \in \mathcal{H}$, when the algorithm is given $m_{\mathcal{H}}(\epsilon, \delta)$ samples drawn from D and labeled by h, then the algorithm produces a hypothesis $\hat{h}$ such that with probability $1-\delta$, $R(\hat{h}) \leq \epsilon$. (Note that the probability is over randomness in the training set as well as any internal algorithmic randomness).*

1 PAC bound for finite hypothesis classes

Now that we have our definition of learning, we can ask how many samples learning requires. Our first learnability result shows that finite hypothesis classes are PAC learnable with a sample complexity depending logarithmically on the number of hypothesis in the class.

Theorem 1 (PAC bound for finite hypothesis class)**.** *Let $\mathcal{H}$ be a hypothesis class with finite size $|\mathcal{H}|$. Then $\mathcal{H}$ is PAC-learnable with*

$$m_{\mathcal{H}}(\epsilon, \delta) = O\left(\frac{\log(|\mathcal{H}|/\delta)}{\epsilon}\right)$$

samples.

Proof. We will show that an algorithm which finds ERM PAC-learnable $\mathcal{H}$. Recall that we defined

$$R(h) = \mathbb{E}_{(x,y)\sim D} 1(h(x) \neq y)$$
$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} 1(h(x) \neq y)$$

to denote the population and empirical risk respectively.

First, due to realizability, there is $h^* \in \mathcal{H}$ such that $R(h^*) = 0$ (Note this also implies that $\hat{R}_s(h^*) = 0$). We will define two sets:

$$\mathcal{H}_{\text{bad}} = \{h \in \mathcal{H} \mid R(h) \geq \epsilon\} \textbf{ (bad hypothesis)}$$
$$S_{\text{bad}} = \{S \in (\mathcal{X} \times \mathcal{Y})^n \mid \exists h \in \mathcal{H}_{\text{bad}}, \hat{R}_s(h) = 0\} \textbf{ (bad training sets)}.$$

Note that $S = S_{\text{bad}} + S_{\text{good}}$, by disjointness. If we can bound the probability of drawing a training set S from S_{bad}, then w.h.p. it must be the case that low empirical risk $\hat{R}$ will correspond to low population risk R.

So our goal is now to upper bound probability of getting training set from S_{bad}:

$$S_{\text{bad}} = \bigcup_{h \in \mathcal{H}_{\text{bad}}} \left\{S \in (\mathcal{X} \times \mathcal{Y})^n \mid \hat{R}_s(h) = 0\right\}.$$

Now,

$$\begin{aligned}
\mathbb{P}_S[S \in S_{\text{bad}}] &= \mathbb{P}_{S \sim D}\left[\bigcup_{h \in \mathcal{H}_{\text{bad}}} \{\hat{R}_s(h) = 0\}\right] \\
&\leq \sum_{h \in \mathcal{H}_{\text{bad}}} \mathbb{P}_{S \sim D}[\hat{R}_s(h) = 0] \textbf{ (union bound)} \\
&= \sum_{h \in \mathcal{H}_{\text{bad}}} \mathbb{P}_{S \sim D}\left[\{\forall i \in \{1, \ldots, n\} \quad h(x_i) = h^*(x_i)\}\right] \textbf{ (realizability)} \\
&= \sum_{h \in \mathcal{H}_{\text{bad}}} \prod_{i=1}^{n} \mathbb{P}_{x_i \sim D}[h(x_i) = h^*(x_i)] \textbf{ (i.i.d.)} \\
&\leq \sum_{h \in \mathcal{H}_{\text{bad}}} \prod_{i=1}^{n} (1 - \epsilon) \\
&\leq |\mathcal{H}|(1 - \epsilon)^n \leq |\mathcal{H}| e^{-\epsilon n} \text{ (since } 1 - x \leq e^{-x}).
\end{aligned}$$

Assume that the bad event of drawing a training set from S_{bad} happens with bounded probability δ. Then by setting $\delta = |\mathcal{H}| e^{-\epsilon n}$, we can solve for n and say that if $n \geq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$, the probability of failure in PAC learning is at most δ, and hence we succeed w.p. $1 - \delta$.

□

2 Bias-Complexity Tradeoff

Recall the following error decomposition from lecture 1, where we assume that computation error is zero.

$$R(h_{\text{ERM}}) - R(h^*) = \underbrace{R(h_{\text{ERM}}) - \min_{h \in \mathcal{H}} R(h)}_{\text{Estimation error}} + \underbrace{\min_{h \in \mathcal{H}} R(h) - R(h^*)}_{\text{Approximation error}}.$$

A natural question to ask is:

- Can we always make approximation error small?
- Or equivalently, can we learn arbitrarily such classes?

The following result shows that the answer to the above questions is no.

Theorem 2 (No-free-lunch theorem)**.** *Let A be a learning algorithm for binary classification over $\mathcal{X}$ and let $n \leq |\mathcal{X}|/2$. Then there exists a distribution D over $\mathcal{X} \times \{0,1\}$ such that*

1. *$R(h^*) = 0$.*
2. *With probability at least $1/7$ over training set S of size n, we have $R_D(A(S)) \geq 1/8$.*

Proof. Let the set $\mathcal{C}$ of size $2n$ be given, where $\mathcal{C}$ is a subset of the domain: $\mathcal{C} \in \mathcal{X}^{2n}$.

$$\mathcal{C} = x_1 \quad x_2 \quad x_3 \quad \cdots \quad x_{2n}$$

Next, consider all possible $T = 2^{2n}$ labelling functions which map from $\mathcal{C}$ to $\{0,1\}$, where $|\mathcal{C}| = 2n$.

$$\begin{array}{lccccc} f_1 & 0 & 0 & 0 & \cdots & 0 \\ f_2 & 0 & 0 & 0 & \cdots & 1 \\ f_3 & 1 & 0 & 0 & \cdots & 0 \\ \vdots & & & & & \\ f_T & 1 & 1 & 1 & \cdots & 1 \end{array}$$

For each f_i, define a distribution D_i such that D_i is uniform over $\mathcal{C}$, with labels given by f_i. Denote D_i^n to be the distribution of n i.i.d. samples from D_i.

It is simple to show that $R_{D_i}(f_i) = 0$.

We want to show that

$$\max_{i \in [T]} E_{S \sim D_i^n} \left[R_{D_i}(A(S)) \right] \geq 1/4. \tag{1}$$

Exercise: Show that (1) implies $\exists\, i \in [T]$ such that with probability at least $1/7$ over training set $S \sim D_i^n$, we have $R_{D_i}(A(S)) \geq 1/8$.

Proof. This follows by a simple application of Markov's inequality:

$$P(X \geq a) \geq \frac{\mathbb{E}[X] - a}{1 - a}$$

with $X = R_{D_i}(A(S))$, and $a = \dfrac{1}{8}$. □

Let $S_1, \ldots, S_K$ be set of all possible training examples ($K = (2n)^n$). We will use the fact that max is greater than average is greater than min.

$$\begin{aligned}
\max_{i \in [T]} \mathbb{E}_{S_j} \left[R_{D_i}(A(S_j^i)) \right] &\geq \frac{1}{T} \sum_{i=1}^{T} \mathbb{E}_{S_j} \left(R_{D_i}(A(S_j^i)) \right) \\
&= \mathbb{E}_{S_j} \frac{1}{T} \sum_{i=1}^{T} \left(R_{D_i}(A(S_j^i)) \right) \\
&\geq \min_{S_j \in \mathcal{C}^n} \frac{1}{T} \sum_{i=1}^{T} \left(R_{D_i}(A(S_j^i)) \right).
\end{aligned}$$

Next, fix any S_j of size n. Then there are $p \geq n$ samples $v_1, \ldots, v_p \in \mathcal{C}$ that do not appear in S_j.

$$\begin{aligned}
R_{D_i}(h) &= \frac{1}{2n} \sum_{x \in \mathcal{C}} 1(h(x) \neq f_i(x)) \\
&\geq \frac{1}{2p} \sum_{\ell=1}^{p} 1(h(v_\ell) \neq f_i(v_\ell)).
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{1}{T} \sum_{i=1}^{T} R_{D_i}(A(S_j^i)) &\geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{2p} \sum_{\ell=1}^{p} 1 \left(A(S_j^i)(v_\ell) \neq f_i(v_\ell) \right) \\
&\geq \frac{1}{2} \min_{r \in [p]} \frac{1}{T} \sum_{i=1}^{T} 1 \left(A(S_j^i)(v_\ell) \neq f_i(v_\ell) \right).
\end{aligned}$$

Partition f_i into $T/2$ pairs, where each pair $(f_i, f_{i'})$ agrees on everything except v_ℓ. Every pair $(f_i, f_{i'})$ produces same labelled datasets $S_j^i, S_j^{i'}$.

$$1 \left(A(S_j^i)(v_\ell) \neq f_i(v_\ell)) \right) + 1 \left(A(S_j^i)(v_\ell) \neq f_{i'}(v_\ell) \right) = 1.$$

Therefore, we can complete the proof with the following.

$$\frac{1}{T} \sum_{i=1}^{T} R_{D_i}(A(S_j^i)) \geq \frac{1}{2} \cdot \frac{1}{T} \cdot \frac{T}{2} = \frac{1}{4}$$

□

As a corollary, we have that the space of all possible functions on an infinite domain is not PAC-learnable with a finite number of samples.

Corollary 3. *Let $\mathcal{X}$ be an infinite domain set ($\mathbb{R}^d$) and let $\mathcal{H}$ be all possible functions from $\mathcal{X} \to \{0, 1\}$. Then, $\mathcal{H}$ is not PAC-learnable.*

3 Agnostic PAC learning

PAC learning requires the *realizibility* assumption, meaning that it cannot handle noise / error within the labels.

Agnostic PAC learning relaxes the realizability distribution, and also allows the labels to be noisy. Consider an arbitrary distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. We define D with noise η in the following way.

$$\mathcal{D}(y|x) = \begin{cases} 1 & \text{with probability } \eta(x) \\ 0 & \text{with probability } 1 - \eta(x). \end{cases}$$

Note that our previous definitions of population and empirical risk are still preserved.

$$R(h) = \mathbb{E}_{(x,y)\sim D}\ell(h(x), y) = \mathbb{P}_{(x,y)\sim D}[h(x) \neq y]$$
$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} 1(h(x_i) = y_i).$$

Digression (Bayes-Optimal Predictor)

We must also consider what it means to be optimal in this new regime, as we *cannot* always get zero population risk (since there is noise for each datapoint). What is the best possible risk which is achievable now?

$$\begin{aligned} &\min_{h:\mathcal{X}\to\mathcal{Y}} \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] \\ &= \sum_{x\in\mathcal{X}} \mathcal{D}(x) \left(\min_{h(x)\in\{0,1\}} \{\mathbb{P}_D(y = 1|X = x)1\{h(x) = 0\}, \mathbb{P}_D(y = 0|X = x)1\{h(x) = 1\}\} \right) \\ &= \sum_{x\in\mathcal{X}} \mathcal{D}(x) \min\{\mathbb{P}_D(y = 1|X = x), \mathbb{P}_D(y = 0|X = x)\}. \end{aligned}$$

We define the **Bayes-optimal predictor** h^* on a datapoint x as:

$$h^*(x) = \mathbb{1}(\eta(x) \geq 1/2)$$
$$\eta(x) = \mathbb{P}_D(Y = 1 \mid X = x).$$

We can think of this as the "best we can do" given the intrinsic noise. Anagolously, we can also define the **Bayes-optimal risk** as follows.

$$R(h^*) = \mathbb{E}_{x\sim D} \min\{\eta(x), 1 - \eta(x)\}$$