CS272 - Theoretical Foundations of Learning, Decisions, and Games

Lecture 3: Statistical learning II

September 4, 2025

Lecturer: Nika Haghtalab
Scribe: Yaowen Ye

Readings: UML Chapter 2

1 Recap

In Lecture 2, we considered PAC learning with a *finite* concept class $\mathcal{C}$ and proved the following theorem of sample complexity:

Theorem 1.1 (Sample complexity for finite $\mathcal{C}$)**.** *Let A be an algorithm that learns $\mathcal{C}$ in a consistency model, then A also learns $\mathcal{C}$ in the PAC model with the number of samples*

$$m_{\mathcal{C}}(\epsilon, \delta) = \mathcal{O}\left(\frac{1}{\epsilon}\left(\ln|\mathcal{C}| + \ln(\frac{1}{\delta})\right)\right).$$

In today's lecture, we consider the case that $\mathcal{C}$ is *infinite*. With infinite $\mathcal{C}$, the union bound we used to prove theorem 1.1 will no longer work. Nevertheless, although $\mathcal{C}$ is infinite, many $c \in \mathcal{C}$ might behave similarly. Therefore, instead of $|\mathcal{C}|$, we should understand the behavior of $c \in \mathcal{C}$ on the sample set or the data distribution and come up with a behavior-dependent "effective" size of the concept class to describe its expressiveness. Specifically, we will leverage *growth functions* and *symmetrization* tricks (double sampling S and S' + random swaps with σ) to achieve the proof.

2 Growth functions

We begin by defining projections and growth functions of concept classes.

Definition 2.1 (Projection)**.** Given sample set $S = (x_1, \ldots, x_m) \in \mathcal{X}^m$ and concept class $\mathcal{C}$, the *projection* of $\mathcal{C}$ on S is

$$\mathcal{C}[S] := \{ (c(x_1), \ldots, c(x_m)) : c \in \mathcal{C} \}.$$

Definition 2.2 (Growth function)**.** For a concept class $\mathcal{C}$, its growth function is defined to be

$$\Pi_{\mathcal{C}}(m) := \max_{S \in \mathcal{X}^m} |\mathcal{C}[S]|$$

Intuitively, the projection is the set of all labelings of S that concepts in $\mathcal{C}$ can produce, while the growth function is the largest number of distinct labelings on any m points that $\mathcal{C}$ can realize. Below we discuss three examples.

Example 2.3 (upper bound). *If $\mathcal{C}$ is the set of all functions $c : \mathcal{X} \to \{0,1\}$, then $\Pi_{\mathcal{C}}(m) = 2^m$ since for each of the m data points we can assign either $+1$ or -1. This is an upper bound for any $\mathcal{C}$.*

Example 2.4 (1-dimensional intervals). *Consider $\mathcal{C}_1 = \{c_{a,b} \ : \ a, b \in \mathbb{R}\}$ where $c_{a,b}(x) = \mathbb{1}(a \leq x \leq b)$. Intuitively, concepts are defined by intervals on the real line. In this setting, we can consider three cases for some S: 1) at least 2 points in S are labeled $+1$, 2) exactly one point in S is labeled $+1$, 3) no point in S is labeled $+1$.*

In case 1, we can consider choosing the left-most and right-most $+1$ points from the m points, which then uniquely determine the labels of every other point. This gives us $\binom{m}{2}$ labelings. Similarly, we only need to choose one $+1$ point ($\binom{m}{1}$) in case 2 and choose no $+1$ point ($\binom{m}{0}$) in case 3 to produce all unique labelings. Combining these three cases gives us

$$\Pi_{\mathcal{C}_1}(m) \ \leq \ \binom{m}{2} + \binom{m}{1} + \binom{m}{0} = \mathcal{O}(m^2).$$

Example 2.5 (2-dimensional axis-aligned rectangles). *Consider $\mathcal{C}_2 = \{c_{a_1,b_1,a_2,b_2} \ : \ a_1, b_1, a_2, b_2 \in \mathbb{R}\}$ where $c_{a_1,b_1,a_2,b_2}(x) = \mathbb{1}(a_1 \leq x_1 \leq b_1$ and $a_2 \leq x_2 \leq b_2)$.*

In this case, concepts are defined by axis-aligned rectangles on a 2-dimensional plane. Note that we can use the previous example to upper bound the number of labelings that axis aligned rectangles can produce. In particular, for any sample set S the labeling produced by c_{a_1,b_1,a_2,b_2} maps to a pair of labelings produced according the x-axis and according to the y-axis, i.e., labels given by c_{a_1,b_1} and c_{a_2,b_2}. Since these are both 1-dimensional intervals that belong to $\mathcal{C}_1$, we have that

$$\Pi_{\mathcal{C}_2}(m) \ \leq \ (\Pi_{\mathcal{C}_1}(m))^2 = \mathcal{O}(m^4).$$

3 Proving the fundamental theorem of PAC learning

Now we can proceed to prove the sample complexity theorem with infinite $\mathcal{C}$, which is also known as the *fundamental theorem of PAC learning*.

Theorem 3.1 (Fundamental theorem of PAC learning). *Let A be an algorithm that learns $\mathcal{C}$ in a consistency model, then A also learns $\mathcal{C}$ in the PAC model with sample complexity m as long as*

$$m \geq \frac{2}{\epsilon}\left(\log(\Pi_{\mathcal{C}}(2m)) + \log(\frac{2}{\delta})\right).$$

This theorem helps us understand whether certain problems are learnable. For example, if $\Pi_{\mathcal{C}}(m) = 2^m$, the inequality $m \geq \frac{2}{\epsilon}(m + 1 + \log(\frac{2}{\delta}))$ cannot hold for any m, which means that this theorem cannot establish the learnability of such $\mathcal{C}$ (indeed such classes are not learnable!) If $\Pi_{\mathcal{C}'}(m) = m^2$, the inequality is equivalent to $m \geq \mathcal{O}(\frac{1}{\epsilon}\log(m))$ which can be achieved for some values of m (this is an arithmatical operation that will formally perform in the future), so $\mathcal{C}'$ is learnable. In general, this theorem shows that is if $\Pi_{\mathcal{C}}(m) = \text{poly}(m)$ then the concept $\mathcal{C}$ is learnable.

To prove theorem 3.1, we first present three definitions of bad events. These bad events are functions of the sample set $S \sim \mathcal{D}^m$, imaginary sample set $S' \sim \mathcal{D}^m$, and a vector of m Bernoulli random variables σ each label 0 or 1 with equal probability. Therefore, in the definitions, we consider S, S', σ as fixed inputs to the bad event functions.

Definition 3.2 (Standard bad event)**.** We define $B(S)\ :\ \exists h \in \mathcal{C}$ s.t. h is consistent with S (i.e., $\mathrm{err}_S(h) = 0$) but $\mathrm{err}_D(h) > \epsilon$.

Definition 3.3 (Bad event with double sampling)**.** We define $B(S, S')\ :\ \exists h \in \mathcal{C}$ s.t. h is consistent with S but $\mathrm{err}_{S'}(h) > \epsilon$.

Definition 3.4 (SWITCH)**.** For some $S = \{x_1, \ldots, x_m\}, S' = \{x'_1, \ldots, x'_m\}, \sigma = (\sigma_1, \ldots, \sigma_m) \in \{0, 1\}^m$, denote $(T, T') = \mathrm{SWITCH}(S, S', \sigma)$ where $T = \{z_1, \ldots, z_m\},\ T' = \{z'_1, \ldots, z'_m\}$. Then, the function SWITCH is defined such that $\forall i = 1, \ldots, m$,

$$z_i = \begin{cases} x_i & \text{if} \quad \sigma_i = 1, \\ x'_i & \text{if} \quad \sigma_i = 0. \end{cases} \qquad z'_i = \begin{cases} x_i & \text{if} \quad \sigma_i = 0, \\ x'_i & \text{if} \quad \sigma_i = 1. \end{cases}$$

Definition 3.5 (Bad event with double sampling and randomness)**.** We define $B(S, S', \sigma)\ :\ \exists h \in \mathcal{C}$ s.t. h is consistent with T but $\mathrm{err}_{T'}(h) > \frac{2}{\epsilon}$, where $(T, T') = \mathrm{SWITCH}(S, S', \sigma)$.

Now, let's start by analyzing $B(S, S')$.

Claim 3.6. *If $m \geq \frac{8}{\epsilon}$, then*

$$\Pr_{S \sim D^m, S' \sim D^m}[B(S, S') | B(S)] \geq \frac{1}{2}.$$

Proof. Assume that $B(S)$ holds and let h be the hypothesis for which $\mathrm{err}_S(h) = 0$ and $\mathrm{err}_{\mathcal{D}}(h) > \epsilon$. Note that $\Pr_{S \sim D^m, S' \sim D^m}[B(S, S') | B(S)] \geq \Pr_{S' \sim \mathcal{D}^m}[\mathrm{err}_{S'}(h) > \frac{\epsilon}{2} | B(S), h]$. We now lower bound this quantity. We have

$$\mathop{\mathbb{E}}_{S' \sim D^m}[\mathrm{err}_{S'}(h) | B(S)] = \mathrm{err}_D(h) > \epsilon,$$

because S' is independent of h. Then, by Chernoff's bound (proof skipped in lecture), we have

$$\Pr\left[\mathrm{err}_{S'}(h) > \frac{\epsilon}{2} \mid B(S)\right] \leq \exp\left(\frac{-m\epsilon}{8}\right) \leq \frac{1}{2}.$$

□

The above claim gives us the following corollary by the definition of conditional probability:

Corollary 3.7. $\Pr_{S \sim \mathcal{D}^m}[B(S)] \leq 2 \Pr_{S, S' \sim \mathcal{D}^m}[B(S, S')]$.

Now, to bound $B(S)$, we can instead consider bounding $B(S, S')$. In fact, bounding $B(S, S')$ is equivalent to bounding $B(S, S', \sigma)$. This is because σ only randomly swaps x's in S and x''s in S', and so (S, S') is identically distributed to (T, T').

Claim 3.8. $\Pr_{S\sim D^m, S'\sim D^m}[B(S,S')] = \Pr_{S\sim D^m, S'\sim D^m, \sigma\sim Ber(\frac{1}{2})}[B(S,S',\sigma)]$.

Here, σ allows us to condition on the additional sample set S' without losing randomness. Introducing S',σ is a common trick called $symmetrization$, where S' is often called shadow samples or ghost samples.

Now we start to derive a bound for $B(S,S',\sigma)$, which will allow us to complete the proof.

Claim 3.9. *For fixed $S,S'\in\mathcal{X}^m$ and $h\in\mathcal{C}$,*

$$\Pr_{\sigma\sim Ber(\frac{1}{2})}\left[\mathrm{err}_T(h)=0 \textit{ and } \mathrm{err}_{T'}(h)>\frac{\epsilon}{2}\right]\le 2^{\frac{-m\epsilon}{2}}.$$

Proof. Let's consider three cases.

Case 1: if $\exists i\in[m]$ s.t. $h(x_i),h(x_i')$ are both wrong, we will have

$$\Pr_{\sigma\sim \mathrm{Ber}(\frac{1}{2})}\left[\mathrm{err}_T(h)=0 \text{ and } \mathrm{err}_{T'}(h)>\frac{\epsilon}{2}\mid \text{Case 1}\right]=0,$$

since $err_T(h)$ cannot be 0 in this case. So, this is not possible.

Case 2: similarly, if more than $(1-\frac{\epsilon}{2})m$ such $i\in[m]$ exist s.t. $h(x_i),h(x_i')$ are both correct, it's impossible to have $\mathrm{err}_{T'}(h)>\epsilon/2$, so we have

$$\Pr_{\sigma\sim \mathrm{Ber}(\frac{1}{2})}\left[\mathrm{err}_T(h)=0 \text{ and } \mathrm{err}_{T'}(h)>\frac{\epsilon}{2}\mid \text{Case 2}\right]=0.$$

Case 3: let the number of $i\in[m]$ s.t. exactly one of $h(x_i),h(x_i')$ is correct be r. Case 1 and 2 indicate that $r\ge\frac{m\epsilon}{2}$. To ensure $\mathrm{err}_T(h)=0$ in $B(S,S',\sigma)$, it must be that on every i within these r indices, σ_i chose the correct one from $h(x_i)$ and $h(x_i')$ which happens with probability 0.5 independently for each i. Therefore

$$\Pr_{\sigma\sim \mathrm{Ber}(\frac{1}{2})}\left[\mathrm{err}_T(h)=0 \text{ and } \mathrm{err}_{T'}(h)>\frac{\epsilon}{2}\mid \text{Case 3}\right]\le 2^{-r}\le 2^{-\frac{m\epsilon}{2}}.$$

□

Now, we are finally ready to prove theorem 3.1. To get $\Pr[B(S)]\le\delta$, it suffices (by corollary 3.7 and claim 3.8) to show that $\Pr[B(S,S',\sigma)]\le\frac{\delta}{2}$. We will proceed to show that A is bounded for any S,S', where A is defined in

$$\Pr_{S\sim D^m,S'\sim D^m \sigma\sim \mathrm{Ber}(\frac{1}{2})}[B(S,S',\sigma)] = \Pr_{S\sim D^m,S'\sim D^m}\left[\underbrace{\Pr_{\sigma\sim \mathrm{Ber}(\frac{1}{2})}[B(S,S',\sigma)\mid S,S']}_{A}\right].$$

Note that when S,S' are given, we can restrict h to the projection $C[S\cup S']$. This is the crucial step, where we no longer depend on the size of $|S|$ and instead are able to bring in the growth function! That is,

$$
\begin{aligned}
A :=& \Pr_{\sigma \sim \mathrm{Ber}(\frac{1}{2})} [B(S, S', \sigma) | S, S'] \\
=& \Pr_{\sigma \sim \mathrm{Ber}(\frac{1}{2})} \left[\exists h \in \mathcal{C},\ \mathrm{err}_T(h) = 0 \ and \ \mathrm{err}_{T'}(h) > \frac{\epsilon}{2} | S, S' \right] \\
=& \Pr_{\sigma \sim \mathrm{Ber}(\frac{1}{2})} \left[\exists h \in \mathcal{C}[S \cup S'],\ \mathrm{err}_T(h) = 0 \ and \ \mathrm{err}_{T'}(h) > \frac{\epsilon}{2} | S, S' \right] \\
=& \sum_{h \in \mathcal{C}[S \cup S']} \Pr_{\sigma \sim \mathrm{Ber}(\frac{1}{2})} \left[\mathrm{err}_T(h) = 0 \ and \ \mathrm{err}_{T'}(h) > \frac{\epsilon}{2} | S, S', h \right] \\
\leq& \Pi_C(2m) \cdot 2^{-\frac{m\epsilon}{2}} \leq \frac{\delta}{2}.
\end{aligned}
$$

by claim 3.8 and the definition of growth function. The transition from the second to the third line holds because when conditioned on fixed S, S', we can restrict our analysis to the set of hypotheses h that produce different labelings for S, S', which by definition is the projection $\mathcal{C}[S \cup S']$.

The last inequality follows from the choice of sample complexity

$$
m \geq \frac{2}{\epsilon} \left(\log(\Pi_C(2m)) + \log(\frac{2}{\delta}) \right).
$$

this completes the proof of theorem 3.1.