Yale University S&DS 669
Statistical Learning Theory

Instructor: Dr. Omar Montasser
Scribe: Anish Lakkapragada

September 10, 2025

Contents

Chapter 1

PAC Learning and VC Theory

1.1 Course Logistics (Lecture 1)

We start the class by introducing our names & majors before getting into objectives of this course. Omar also covers the syllabus (recommended prerequisites, grading, and AI policy) before going over a roadmap of the things we will cover. Okay, let's start!

1.2 Introducing the Statistical Learning Theory Framework (Lecture 1)

We now introduce the statistical learning theory framework where we have the following objects:

- Domain X (e.g. $X = \mathbb{R}^d$) where each $x \in X$ is called an "instance"
- Label Space Y (e.g. $Y = \{\pm 1\}$ or $Y = \mathbb{R}$)
- Unknown source distribution D over $X \times Y$. This is an assumption on the data generating process (formed by "nature" or "reality").
- Goal: find a predictor $h : X \to Y$ achieving small *expected error* $L_D(h) := \mathbb{P}_{(x,y)\sim D}\{h(x) \neq y\}$.
- Access to an oracle: We have an i.i.d training sample $S = \{(x_i, y_i)\}_{i=1}^m$ drawn from D (notated by $S \sim D^m$)

Restated, our goal is to create some learner $A : (X \times Y)^\star \to Y^X$, where the $\star$ denotes a variable-length sequence of $X \times Y$ (i.e. our dataset) and Y^X is the set of all functions mapping from X to Y.

Omar notes that we will first start by assuming that any instance $x \in X$ has a "ground-truth" label, as opposed to a case where D allows for 50% probability mass on $(x, +1)$ and $(x, -1)$ (such a case could happen to to reflect uncertainty in the label of x). More generally, we will start with these strong assumptions in the bulleted list above and relax them later.

It's worth emphasizing **two main assumptions about our data** within this framework:

- We observe i.i.d training samples from (unknown) distribution D.
- Future (*unseen*) examples are drawn from the same distribution D.

The second point is easier to forget.

Expected vs. Empirical Error. Let's look a bit more closely at our objective: minimizing our *expected error*

$$L_D(h) := \mathbb{P}_{(x,y)\sim D}\{h(x) \neq y\} \tag{1.1}$$

Why not minimize this directly? Answer: we don't assume access to the data distribution D. Hence, given some sample S we use the *empirical error*:

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{h(x_i) \neq y_i\} \tag{1.2}$$

as our proxy to D. While this is a typical setup in machine learning, it leads to the following questions:

- How should we use the empirical error?
- Is it a good estimate for the expected error? And how good?

Specifically, we are interested in their difference:

$$|L_D(h) - L_S(h)| = |\mathbb{P}_{(x,y)\sim D}\{h(x) \neq y\} - \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{h(x_i) \neq y_i\}| \tag{1.3}$$

Please recognize that the empirical error $L_S(h)$ is a random variable as it is a function of the randomly drawn dataset $S \sim D^m$ whereas $L_D(h)$ is just a population statistic. The relationship between the two should be more clear from the below quick exercise:

$$\forall\ h : X \to \{\pm 1\} \text{ with } D \text{ over } X \times \{\pm 1\},\ \text{show that } \mathbb{E}_{S\sim D^m}[L_S(h)] = L_D(h) \tag{1.4}$$

But this is not a useful fact as its an asymptotic, and we are more interested in the difference in the case of a finite dataset size m. Thus, we often use tools like *concentration inequalities* (e.g. Hoeffding's) to create bounds like the below for some fixed h and m:

$$\mathbb{P}_{S\sim D^m}[|L_S(h) - L_D(h)| > \epsilon] \leq 2\exp(-2\epsilon^2 m) \tag{1.5}$$

Or restated equivalently (i.e. define $\delta := 2\exp(-2\epsilon^2 m)$ and "invert" the probabilities),

$$\mathbb{P}_{S\sim D^m}[|L_S(h) - L_D(h)| \leq \sqrt{\frac{\ln(2/\delta)}{2m}}] \geq 1 - \delta \tag{1.6}$$

From this expression it should be clear that as $m \to \infty$, our expected difference between expected and empirical error goes to zero.

1.3 Consistent Learning Rule Bound for Finite Hypothesis Class (Lecture 1)

Before actually creating another bound ourselves, we structure our problem even more with some prior knowledge/decisions we make:

- We restrict ourselves to a subset of functions from X to Y called our *hypothesis class* $H \subseteq Y^X$. Examples of H are given below:
  - Linear Predictors
  - Support Vector Machines (SVMs)
  - Neural Networks

H will represent our "prior knowledge" or "expert knowledge". For example, if our domain X is a set of images we would likely consider using a convolutional neural network (CNN) as our H as CNNs perform well on this kind of data.

- Assume some true function $y = f^\star(x)$ where $f^* \in H$. Our learner A will know H but not $f^\star$ (it will have to learn this function!).
- As an implication of the above assumption, we will say a sequence $((x_i, y_i))_{i=1}^m$ is *realizable* by H if the true function $f^\star \in H$ gives matching ground truth predictions[1]

Having established these assumptions, we are now ready to put them to use by creating our own bound!

Warm-up: Finite Classes. Consider the following assumptions:

- H is finite
- D is realizable by H (i.e. $\exists\ f^\star \in H$ s.t. $L_D(f^\star) = \mathbb{P}_{(x,y)\sim D}[f^\star(x) \neq y] = 0$)

Note that, as stated before, we cannot minimize $\min_{h\in H} L_D(h)$ directly and instead must work on our sample $S \sim D^m$. We present the following definition:

Definition 1.1. *We have a* consistent learning rule (CLR) *when for any input $S = \{(x_i, y_i)\}_{i=1}^m$, we can output* any $h \in H$ *s.t.* $\forall_{1\leq i\leq m}, h(x_i) = y_i$.

Then, we have the following question. If $\hat{h} := \text{CLR}_H(S)$ for some consistent learning rule CLR_H on hypothesis class H, what can we say about $L_S(\hat{h})$? It should be zero, but does this imply that $L_D(\hat{h}) = 0$?

Here's a closely-related example: consider some h where $L_D(h) = \frac{1}{2}$. This is a bad function that is correctly 50% of the time in truth. But $\mathbb{P}_{S\sim D^m}[L_S(h) = 0] > 0 \neq 0$, meaning $\exists\ S$ s.t. $L_S(h) = 0$ (i.e. we can be fooled to think h is good on some sample S.) Thus, we now **create a bound for a finite hypothesis class to control $L_D(h)$ on some CLR-learned h.**

Derivation of CLR bound for finite hypothesis class. Fix any function $h \in H$. We define ϵ s.t. $L_D(h) > \epsilon$. We proceed with the following steps:

- We first can find the probability of the bad event ($L_S(h) = 0$) below: [2]:
  $\mathbb{P}_{S\sim D^m}[L_S(h) = 0] = \prod_{i=1}^m \mathbb{P}_{S\sim D^m}\{h(x_i) = y_i\} = \prod_{i=1}^m (1 - L_D(h)) \leq (1-\epsilon)^m \leq \exp(-\epsilon m)$
- But this is just one bad function $\in H$! We can a *group* of bad functions with $B_\epsilon := \{h \in H : L_D(h) > \epsilon\} \subset H$. Then to get the probability that any CLR-learned function is "bad" we can use a *union bound*:

$$\mathbb{P}_{S\sim D^m}\big[\text{CLR}_H(S) \in B_\epsilon\big] \leq \mathbb{P}_{S\sim D^m}\big[\exists\, h \in B_\epsilon : L_S(h) = 0\big] \tag{1.7}$$

$$\leq \sum_{h\in B_\epsilon} \mathbb{P}_{S\sim D^m}\big[L_S(h) = 0\big] \leq |B_\epsilon|\, e^{-m\epsilon} \leq |H|\, e^{-m\epsilon}. \tag{1.8}$$

- We can then set $\delta := |H| \exp(-m\epsilon)$ and invert the expression to arrive at Theorem 1.2. So to ensure $\text{CLR}_H(S) \notin B_\epsilon \iff \text{CLR}_H(S) \leq \epsilon$ with probability $\geq 1 - \delta$ for some predecided $\delta \in (0, 1)$, we will need $m(\epsilon, \delta) = \frac{\ln|H| + \ln(1/\delta)}{\epsilon}$ many samples[3].

[1] Mathematically speaking, this means $\forall x_i, f^\star(x_i) = y_i \implies L_S(f^\star) = 0$. This realizability assumption is non-trivial and we will discuss it further in the course.

[2] The last argument here is done using Bernoulli's Inequality.

[3] To get this expression, solve for m in terms of δ.

Pat yourself on the back! We resummarize this bound in the following theorem:

Theorem 1.2 (CLR Bound with (ϵ, δ) fixed)**.** *For any finite class H, any (realizable) distribution D, any $(\epsilon, \delta) \in (0,1)^2$, with $m = \frac{\ln|H| + \ln(1/\delta)}{\epsilon}$, we have:*

$$\mathbb{P}_{S\sim D^m}[L_D(CLR_H(S)) \leq \epsilon] \geq 1 - \delta \tag{1.9}$$

Choosing to take the perspective that our number of samples m is fixed and so we are interested in the lowest possible error we can achieve w.h.p, we can use the following theorem:

Theorem 1.3 (CLR Bound with m fixed)**.** *For any finite class H, any (realizable) distribution D, any $\delta \in (0,1), m \in \mathbb{N}$:*

$$\mathbb{P}_{S\sim D^m}[L_D(CLR_H(S)) \leq \frac{\ln|H| + \ln(1/\delta)}{m}] \geq 1 - \delta \tag{1.10}$$

This constitutes the first learning guarantee that we have derived. Note that in our derivation we did not pay much attention to the implementation or procedure of the CLR, which will depend on H. We also used the realizability assumption, which has some implications:

- What if there is *no* predictor $h \in H$ s.t. $L_D(h) = 0$?
- In such a case, can we use with $\min_{h\in H} L_D(h)$?

In response to the second point, we will soon look at **empirical risk minimization** where:

$$\text{ERM}_H(S) = \arg\min_{h\in H} \frac{1}{|S|} \sum_{(x,y)\in S} \mathbf{1}\{h(x_i) \neq y_i\} \tag{1.11}$$

So given some function $\hat{h} := \text{ERM}_H(S)$, you might be wondering if it will satisfy our Hoeffding bound:

$$\mathbb{P}_{S\sim D^m}[|L_S(\hat{h}) - L_D(\hat{h})| \leq \sqrt{\frac{\ln(2/\delta)}{2m}}] \geq 1 - \delta \tag{1.12}$$

The answer is no. This is because that bound operates on a fixed h seen *a priori* before our sampled data, whereas $\hat{h}$ is a function of the data (e.g. $\hat{h}$ is a random variable) and so the inequality does not apply.

Note on Overfitting. Furthermore, while the set of cases where $|L_S(h) - L_D(H)| > \sqrt{\frac{\ln(2/\delta)}{2m}}$ may only have δ probability w.r.t. $S \sim D^m$, they all add up and so given many $\{h_i\}_{i=1}^K \subset H, \mathbb{P}_{S\sim D^m}[\exists\ i \text{ s.t. } |L_D(h_i) - L_S(h_i)| \text{ is large}]$ is not small. Thus, we want a stronger guarantee that w.h.p all empirical errors $L_S(h)$ are close to their expected errors $L_D(h)$:

$$\mathbb{P}_{S\sim D^m}[\forall h \in H : |L_S(h) - L_D(h)| > \epsilon] \leq \dots \tag{1.13}$$

This is known as *uniform convergence.*

1.4 Uniform Convergence & the Probably Approximately Correct (PAC) Framework (Lecture 2)

We start by giving our first attempt at a uniform convergence bound:

Theorem 1.4 (Hoeffding-derived Uniform Convergence Bound for finite H)**.** *For a finite hypothesis class $|H| < \infty$, we have the following uniform convergence bound from the a priori Hoeffding bound:*

$$\mathbb{P}_{S\sim D^m}[\exists h \in H : |L_D(h) - L_S(h)| > \epsilon] \leq |H| \cdot \mathbb{P}_{S\sim D^m}[|L_D(h) - L_S(h)| > \epsilon] = 2|H|\exp(-2\epsilon^2 m) \tag{1.14}$$

Defining $\delta := 2|H|\exp(-2\epsilon^2 m)$ and solving for m, we arrive at the data-form of this bound:

Theorem 1.5 (Theorem 1.5 for fixed (ϵ, δ))**.** *For any finite hypothesis clas $|H| < \infty$, any distribution D, any $(\epsilon, \delta) \in (0,1)^2$ we have:*

$$\mathbb{P}_{S\sim D^m}[\forall h \in H : |L_S(h) - L_D(h)| \leq \epsilon] \geq 1 - \delta \tag{1.15}$$

where $m(\epsilon, \delta) = \frac{\ln|H| + \ln(2/\delta)}{2\epsilon^2}$.

Note here that in contrast to our CLR bound Theorem 1.2, we will require $\frac{1}{\epsilon^2}$ samples as opposed to $\frac{1}{\epsilon}$. So removing the realizability assumption means that we will need more samples. Omar notes that we will explore a relaxed realizability assumption that leads to $m = O(\frac{1}{\epsilon})$ in our homework.

From these bounds, we can create some ERM-specific bounds:

Theorem 1.6 (ERM "Post-hoc" Guarantee for finite H)**.** *For any finite class H, any distribution D, any $\delta \in (0,1), m \in \mathbb{N}$, with probability $\geq 1 - \delta$ over $S \sim D^m$, we have:*

$$L_D(ERM_H(S)) \leq L_S(ERM_H(S)) + \sqrt{\frac{\ln|H| + \ln(2/\delta)}{2m}}$$

Proof. We can invoke Theorem 1.5 with $\epsilon = \sqrt{\frac{\ln|H|+\ln(2/\delta)}{2m}}$ to have $\geq 1 - \delta$ probability that $\forall h \in H : |L_D(h) - L_S(h)| \leq \epsilon$. But $\text{ERM}_H(S) \in H \implies L_D(\text{ERM}_H(S)) \leq L_S(\text{ERM}_H(S)) + \epsilon$ with $\geq 1 - \delta$ probability. So we are finished. □

Theorem 1.7 (ERM "A-Priori" Guarantee for finite H)**.** *For any finite class H, any distribution D, any $\delta \in (0,1), m \in \mathbb{N}$, with probability $\geq 1 - \delta$ over $S \sim D^m$,*

$$L_D(ERM_H(S)) \leq \min_{h\in H} L_D(h) + 2\sqrt{\frac{\ln|H| + \ln(2/\delta)}{2m}} \tag{1.16}$$

Proof. Note that by definition of ERM, $\forall\ \tilde{h} \in H, L_S(\text{ERM}_H(S)) \leq L_S(\tilde{h})$. Furthermore, $\forall\ \tilde{h} \in H : L_S(\tilde{h}) \leq L_D(\tilde{h}) + \epsilon$ with $\geq 1 - \delta$ probability. So with $\geq 1 - \delta$ probability we have:

$$\forall\ \tilde{h} \in H, \underbrace{L_D(\text{ERM}_H(S)) \leq L_S(\text{ERM}_H(S))}_{\text{see proof of Theorem 1.6}} + \epsilon \leq L_S(\tilde{h}) + \epsilon \leq L_D(\tilde{h}) + 2\epsilon$$

We apply $\min_{\tilde{h}}$ to both sides of this inequality to arrive at the theorem. □

Before moving forward, we point out the following concepts of approximation and estimation error shown in Theorem 1.7.

- The approximation error $\min_{h\in H} L_D(h)$ reduces with richer/larger hypothesis classes H
- However, these expanded hypothesis classes will demand more samples in order to maintain the same estimation error $2\sqrt{\frac{\ln|H|+\ln(2/\delta)}{2m}}$

We now move onto formally defining **Probability Approximately Correct (PAC)** learning, which Omar notes won a Turing Award. We provide the following two definitions:

Definition 1.8 (Realizably-PAC-Learnable Hypothesis Class)**.** *A hypothesis class H is realizably-PAC-learnable if there exists a learning rule A s.t. $\forall(\epsilon, \delta) \in (0,1)^2, \exists\ m(\epsilon, \delta) \in \mathbb{N}, \forall$ distributions D s.t. $\inf_{h\in H} L_D(h) = 0$,*

$$\mathbb{P}_{S\sim D^{m(\epsilon,\delta)}}[L_D(A(S)) \leq \epsilon] \geq 1 - \delta$$

Note that this type of PAC-learnable hypothesis class is *distribution independent*, meaning the bound applies for any data distribution D that is realizable (i.e. $\inf_{h\in H} L_D(h) = 0$). Dropping the realizability assumption for H, we provide another PAC definition for a hypothesis class:

Definition 1.9 (Agnostically-PAC-learnable Hypothesis Class)**.** *A hypothesis class H is agnostically-PAC-learnable if there exists a learning rule A such that $\forall(\epsilon,\delta) \in (0,1)^2, \exists\ m(\epsilon,\delta) \in \mathbb{N}, \forall$ distributions D,*

$$\mathbb{P}_{S\sim D^{m(\epsilon,\delta)}}\{L_D(A(S)) \leq \inf_{h\in H} L_D(h) + \epsilon\} \geq 1-\delta$$

Looking at Definition 1.8 and Theorem 1.2, we get the following corollary.

Corollary 1.10 (Finite classes H are realizably-PAC-learnable with CLR.)**.** *All finite classes H are realizably-PAC-learnable using CLR with sample complexity*

$$m(\epsilon,\delta) = \frac{\ln|H| + \ln(1/\delta)}{\epsilon}$$

Similarly looking at Definition 1.9 and Theorem 1.7, we arrive at the following corollary:

Corollary 1.11 (Finite classes H are agnostically-PAC-learnable with ERM)**.** *All finite classes H are agnostically-PAC-learnable using ERM with sample complexity*

$$m(\epsilon,\delta) = O\left(\frac{\ln|H| + \ln(1/\delta)}{\epsilon}\right)$$

So we have established that all finite hypothesis classes are realizably and agnostically PAC-learnable. Now what about infinite classes? And can we learn with less samples than log-cardinality (i.e. $m(\epsilon,\delta) \ll \ln|H|$)? The answer is yes, and we now begin our study of the legendary VC Dimension.

1.5 Vapnik-Chervonenkis (VC) Dimension (Lecture 2)

We first start by developing some technology of the **growth function**. For $C = (x_1,\ldots,x_m) \in X^m$, define the restriction (or projection) of H onto C as:

$$H|_C = \{(h(x_1),\ldots,h(x_m)) \mid h \in H\}$$

We can then define the growth function as $\Gamma_H(m) = \max_{C\in X^m} |H|_C|$. We look at a few examples to understand how this growth function works:

- $X = \{1,\ldots,100\}, H = \{\pm 1\}^X$. Then we have $\Gamma_H(m) = \min(2^m, 2^{100})$.
- $X = \{1,\ldots,2^{100}\}, H = \{\mathbf{1}[x \leq \theta] \mid \theta \in \{1,\ldots,2^{100}\}\}$ will have $\Gamma_H(m) = \min(m+1, 2^{100})$

The idea for both these two examples is that when two of the data points are the same (i.e. $m > |X|$), they must be labeled identically and so the growth function hits a limit. Moreover, we should observe that both function classes in these examples have the same cardinality but that the growth function $\Gamma_H(m)$ can distinguish between them (H is a lot more complex in the first example, and hence has the higher growth function.) As a teaser for future results, the growth function gives us the following result which we will later prove:

Theorem 1.12. *(Growth Function Data Bound for Realizable Distribution)*
For any hypothesis class H, any (realizable) distribution D, any $(\epsilon,\delta) \in (0,1)^2$ with sample complexity:

$$m(\epsilon,\delta) = O\left(\frac{\ln[\Gamma_h(2m)] + \ln(1/\delta)}{\epsilon}\right)$$

with probability $\geq 1-\delta$ over $S \sim D^{m(\epsilon,\delta)}$ we have that $\forall h \in H : L_S(h) = 0 \implies L_D(h) \leq \epsilon$.

We also have a similar theorem in the case that D is not realizable, where $m(\epsilon,\delta) \propto \frac{1}{\epsilon^2}$. The main thing to notice here is that we are no longer having our sample complexity $m(\epsilon,\delta)$ tied to $\ln|H|$. So now we do not require our bounds to be finite, and can work with infinite function classes!

Vapnik-Chervonenkis (VC) Dimension. Using the technology we have so far, we are ready to define the VC dimension. We say $C = \{x_1, \ldots, x_m\}$ is *shattered* by H if $|H|_C| = 2^m$, i.e. the projection contains all 2^m possible labelings. The VC-dimension of H, denoted by $\text{vc}(H)$ is the largest number of points that can be shattered by H:

$$\text{vc}(H) = \max\{m \in \mathbb{N} : \Gamma_H(m) = 2^m\}$$

We say $\text{vc}(H)$ is infinite if H is infinite[4] and $\forall m, \Gamma_H(m) = 2^m$. We now practice in class with a few examples of the VC Dimension:

- $X = \{1, \ldots, 100\}, \quad H = \{\pm 1\}^X$.
- $X = \{1, \ldots, 2^{100}\}, \quad H = \{\mathbf{1}[x \leq \theta] \mid \theta \in \{1, \ldots, 2^{100}\}\}$.
- $X = \mathbb{R}, \quad H = \{\mathbf{1}[x \leq \theta] \mid \theta \in \mathbb{R}\}$.
- $X = \mathbb{R}, \quad H = \{\mathbf{1}[a \leq x \leq b] \mid a, b \in \mathbb{R}\}$.
- Axis-aligned rectangles[5] (in $\mathbb{R}^d$).

Note that in order to show that the $\text{vc}(H) = k$, we must show that we can shatter some set of k points but no set of $k + 1$ points.

[4] If H is finite, $\text{vc}(H) \leq \log_2 |H|$ as $\Gamma_H(m) \leq |H|$ (see definition.)

[5] **SPOILER**: Answer is four. Any set of five points will have an "interior point" in the convex hull and so you cannot make all boundary points be class one but the interior point be class zero.