Introduction to Statistical Learning Theory

Matthew Li

July 2024

Table of Contents

Current Scene

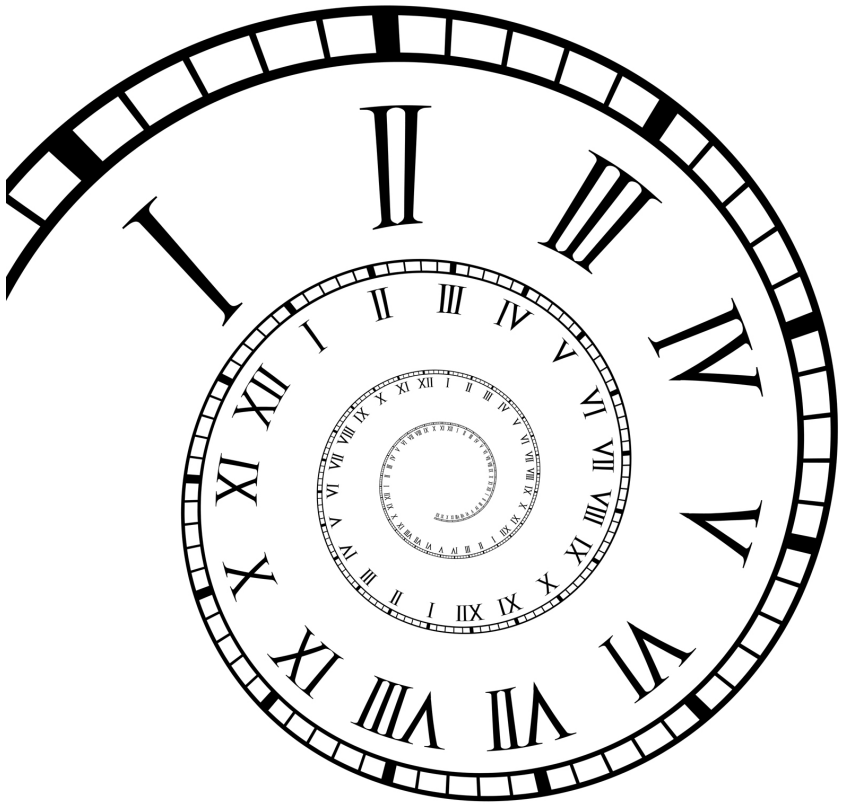
What's trending right now in terms of machine learning?

- Large Language Models (GPT, Claude, Llama)
- Generative Models (GANs, Diffusion)

Takeaway: deep learning is very popular! Think of very big matrix multipliers that "train" on lots of data.

Time Travel

Let's go back to the late 1900's!



Why Bother?

- Leslie Valiant proposes the PAC learning framework (1984)
- Vladimir Vapnik and Alexey Chervonenkis developed VC theory (1970's)



Figure: Left to right: Valiant, Vapnik, Chervonenkis

- Created to answer some very important questions in ML
- What does it mean to learn? What can we learn? When can we learn?

Background

Consider the following scenario: the Ross counselors and peer mentors are sick of grading number theory sets! Decide to train a "learner" to automate the grading process so they can have more time to play ultimate frisbee.

Definition

The **domain set** or **instance space** is an arbitrary set $\mathcal{X}$ of objects we wish to label. The elements of $\mathcal{X}$ are sometimes called instances. Usually the elements are vectors of *features*.

Definition

The **label set** is denoted as $\mathcal{Y}$. For the current scenario, we can restrict $\mathcal{Y}$ to be $\{0, 1\}$.

More Definitions

Definition

The **training set** is a finite sequence $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$ where $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$.

Definition

We assume that all instances are sampled from an arbitrary probability distribution $\mathcal{D}$ that is unknown to the learner. We also assume there is a "true" labeling function $f : \mathcal{X} \to \mathcal{Y}$.

Definition

The learner wishes to output a **hypothesis** $h : \mathcal{X} \to \mathcal{Y}$ that is "close" to the true labeling function. This hypothesis is generated by a **learning algorithm** A that takes in the training set.

Empirical Risk Minimization (ERM)

Definition

The **error** of h is the probability that it does not predict the correct label on a random instance x generated by $\mathcal{D}$.

$$L_{\mathcal{D},f}(h) := \mathbb{P}_{x\sim\mathcal{D}}[h(x) \neq f(x)]$$

Because $\mathcal{D}$ and f are both unknown, we do not know the true error. Instead we calculate the **training error** or **empirical error**.

Definition

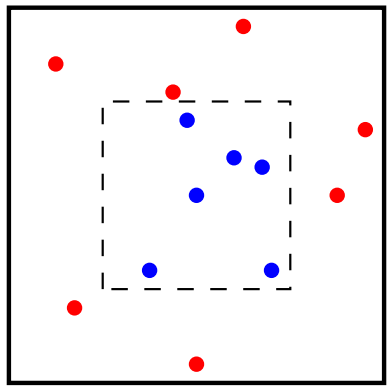
The error the hypothesis incurs over the training set is

$$L_S(h) := \frac{1}{n}\sum_{i=1}^{n}\{h(x_i) \neq y_i\}$$

A learning algorithm that selects $h^\star := \arg\min_h L_S(h)$ is an ERM.

Overfitting Problem

Suppose all instances are in $(0,1) \times (0,1)$ sampled from a probability distribution $\mathcal{D}$ such that the points are uniformly distributed within the square. Also let the labeling function f label any point inside the dashed square of area $\frac{1}{2}$ as 0 and 1 otherwise.



What if:

$$h(x) = \begin{cases} y_i & \text{if } \exists x_i = x \\ 1 & \text{otherwise} \end{cases}$$

Notice $L_S(h) = 0$, so this will be selected based on ERM. However, $L_{\mathcal{D}} = \frac{1}{2}$!

Saving ERM

How do we guarantee ERM does good on training data and doesn't overfit?
Answer: Apply ERM to a restricted search space of hypotheses!

Definition

Before seeing the data, the learner should choose a restricted set of hypotheses, known as a **hypothesis class** $\mathcal{H}$. For a given class $\mathcal{H}$, ERM gives us

$$\mathrm{ERM}_{\mathcal{H}}(S) = h^{\star} \in \arg\min_{h \in \mathcal{H}} L_S(h)$$

Finite Hypothesis Classes

The easiest way to put restrictions on a class is to put a limit on its size. From now on, assume $\mathcal{H}$ is finite. We will show our first important result.

Theorem

If $\mathcal{H}$ is a finite class then $\text{ERM}_{\mathcal{H}}$ will probably not overfit, provided S is large enough (depends on size of $\mathcal{H}$).

It's impossible to guarantee perfect label prediction. Instead introduce **accuracy parameter** ϵ. If $L_{\mathcal{D}}(h_S) < \epsilon$, then the hypothesis is approximately correct. If not, then the learner has overfit or failed.

Suffices to show $\mathbb{P}_{S\sim\mathcal{D}^n}[L_{\mathcal{D}}(h_S) \geq \epsilon]$ is bounded above by some $\delta \in (0,1)$ which implies with probability at least $1-\delta$ that $L_{\mathcal{D}}(h_S) < \epsilon$. We call $1-\delta$ the **confidence parameter**.

Proof - I

Let

$$\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D}}(h) \geq \epsilon\}$$
$$M = \{S \sim \mathcal{D}^n : \exists h \in \mathcal{H}_B, L_S(h) = 0\}.$$

Next, we make the following assumption.

Definition

We assume there exists $h^\star \in \mathcal{H}$ such that $L_{\mathcal{D}}(h^\star) = 0$. This implies with probability 1 over random samples of S that $L_S(h^\star) = 0$. This is known as the **realizability assumption**.

From this, we can see that

$$\{S \sim \mathcal{D}^n : L_{\mathcal{D}}(h_S) \geq \epsilon\} \subseteq M$$

Proof - II

So, we can write

$$\mathbb{P}_{S\sim\mathcal{D}^n}[L_{\mathcal{D}}(h_S) \geq \epsilon] \leq \mathbb{P}_{S\sim\mathcal{D}^n}[\exists h \in \mathcal{H}_B, L_S(h) = 0]$$
$$\leq \mathbb{P}_{S\sim\mathcal{D}^n}\left[\bigcup_{h\in\mathcal{H}_B} L_S(h) = 0\right].$$

Now, by the union bound

$$\mathbb{P}_{S\sim\mathcal{D}^n}[L_{\mathcal{D}}(h_S) \geq \epsilon] \leq \sum_{h\in\mathcal{H}_B} \mathbb{P}_{S\sim\mathcal{D}^n}[L_S(h) = 0]$$

We will bound each summand in the right hand side. Fix some arbitrary $h \in \mathcal{H}_B$. Notice that $L_S(h) = 0$ iff $h(x_i) = f(x_i)$ for all $1 \leq i \leq n$.

Proof - III

We will make one more assumption, namely every element of S is sampled i.i.d. from $\mathcal{D}$. Then, we may write

$$\mathbb{P}_{S\sim\mathcal{D}^n}[L_S(h) = 0] = \prod_{i=1}^{n} \mathbb{P}_{x_i\sim\mathcal{D}}[h(x_i) = f(x_i)].$$

Now notice that $\mathbb{P}_{x_i\sim\mathcal{D}}[h(x_i) = f(x_i)] = 1 - L_{\mathcal{D}}(h) \leq 1 - \epsilon$ and using the inequality $1 - x \leq e^{-x}$

$$\mathbb{P}_{S\sim\mathcal{D}}[L_S(h) = 0] \leq (1-\epsilon)^n \leq e^{-n\epsilon}$$

Then,

$$\mathbb{P}_{S\sim\mathcal{D}^n}[L_{\mathcal{D}}(h_S) \geq \epsilon] \leq |\mathcal{H}_B|e^{-n\epsilon} \leq |\mathcal{H}|e^{-n\epsilon}.$$

Proof - IV

Notice we can guarantee $\mathbb{P}_{S\sim\mathcal{D}^n}[L_{\mathcal{D}}(h_S) \geq \epsilon] \leq \delta$ for some $0 < \delta < 1$ if $|\mathcal{H}|e^{-n\epsilon} \leq \delta$. In other words, we must have

$$n \geq \frac{\log\left(|\mathcal{H}|/\delta\right)}{\epsilon}.$$

Let's summarize all our work into a single result.

Theorem

Let $\mathcal{H}$ be a finite hypothesis class. Then for all $\epsilon, \delta \in (0,1)$, if n is an integer satisfying

$$n \geq \frac{\log\left(|\mathcal{H}|/\delta\right)}{\epsilon},$$

then for any labeling function f and distribution $\mathcal{D}$ for which the realizability hypothesis holds, we have with probability at least $1-\delta$ over $S \sim \mathcal{D}^n$ that every ERM hypothesis h_S satisfies $L_{\mathcal{D}}(h_S) < \epsilon$.

PAC Learning

We can try to generalize this notion of learning we just explored.

Definition

A hypothesis class $\mathcal{H}$ is **PAC learnable** if there exists a learning algorithm A such that for every $\epsilon, \delta \in (0, 1)$, every distribution $\mathcal{D}$ over $\mathcal{X}$ and every f that satisfy the realizability assumption, when the algorithm is given at least $m_{\mathcal{H}}(\epsilon, \delta)$ samples, A will produce a hypothesis h_S such that with probability $1 - \delta$ over the training set, we have $L_{\mathcal{D}}(h_S) < \epsilon$.

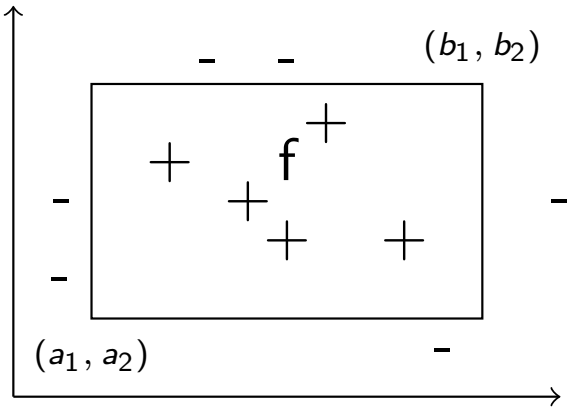
The acronym PAC stands for *probably* and *approximately* correct. The epsilon value corresponds to the "approximate" part and the delta value corresponds to the "probable" part.

Problem Setup

Let's try to prove a different hypothesis class is PAC learnable. Namely, this time we will consider the case where $\mathcal{H}$ is infinite.
In this scenario, let $\mathcal{X}$ be points in $\mathbb{R}^2$ sampled from an arbitrary distribution $\mathcal{D}$ and let $\mathcal{Y} = \{0, 1\}$. Our hypothesis class consists of all

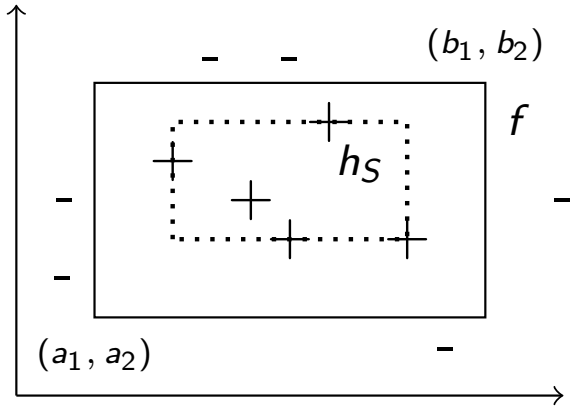
$$h_{(a_1,a_2,b_1,b_2)}(x, y) = \begin{cases} 1 & \text{if } a_1 \leq x \leq b_1 \text{ and } a_2 \leq y \leq b_2 \\ 0 & \text{otherwise} \end{cases}$$

where $a_1, a_2, b_1, b_2 \in \mathbb{R}$ and $a_1 \leq b_1, a_2 \leq b_2$. Our labeling function will be an element of $\mathcal{H}$.



Proof - I

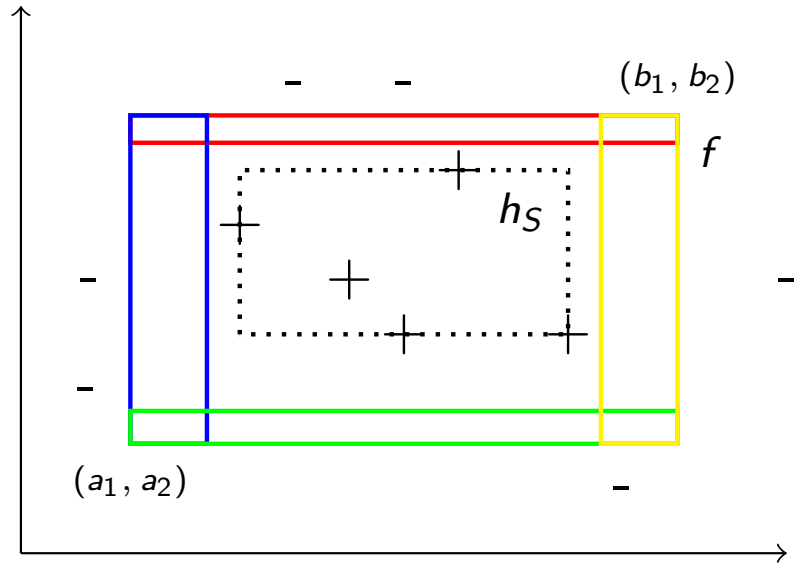
Consider the learning algorithm that chooses $h_{(a_1,b_1,a_2,b_2)}$ such that the rectangle determined by (a_1, a_2) and (b_1, b_2) is the "tightest" rectangle containing all points with value 1.



Fix an arbitrary $\epsilon \in (0, 1)$. We will now "sweep out" 4 rectangular regions $R_1, \ldots, R_4$ along the sides of the target rectangle such that $\mathbb{P}_{x\sim\mathcal{D}}[x \in R_i] \geq \epsilon/4$ for all $1 \leq i \leq 4$.

Proof - II

So we have the following figure.



Thus, we can write $\mathbb{P}_{S\sim\mathcal{D}^n}[L_{\mathcal{D}}(h_S) \geq \epsilon] \leq \mathbb{P}_{S\sim\mathcal{D}^n}[\bigcup_{i=1}^{4}\{R_S \cap R_i = \emptyset\}]$ and by union bound

$$\mathbb{P}_{S\sim\mathcal{D}^n}[L_{\mathcal{D}}(h_S) \geq \epsilon] \leq \sum_{i=1}^{4} \mathbb{P}_{S\sim\mathcal{D}^n}[R_S \cap R_i = \emptyset].$$

Proof - III

Then, we can write

$$\begin{aligned}\mathbb{P}_{S\sim\mathcal{D}^n}[L_{\mathcal{D}}(h_S) \geq \epsilon] &\leq 4\prod_{i=1}^{n}\mathbb{P}_{x_i\in\mathcal{D}}[x_i \cap R_i = \emptyset] \\ &\leq 4\left(1-\frac{\epsilon}{4}\right)^n \\ &\leq 4e^{-\frac{\epsilon}{4}n}\end{aligned}$$

If we want to bound this above by $\delta \in (0,1)$, then we must have

$$n \geq \frac{4}{\epsilon}\log\frac{4}{\delta}.$$

This establishes that $|\mathcal{H}|$ being finite is a sufficient but not necessary condition in proving PAC learnability.

Shattering

Definition

Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$ and let $C \subset \mathcal{X}$. The **restriction of $\mathcal{H}$ to C** is the set

$$\mathcal{H}_C = \{(h(c_1), \ldots, h(c_{|C|})) : h \in \mathcal{H}\}.$$

We may view every function in $\mathcal{H}_C$ as a vector in $\{0,1\}^{|C|}$.

Definition

A hypothesis class $\mathcal{H}$ **shatters** a finite set $C \subset \mathcal{X}$ if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0,1\}^{|C|}$.

VC Dimension

Definition

The **VC-dimension** of a hypothesis class $\mathcal{H}$, denoted $\text{VCdim}(\mathcal{H})$, is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter C with arbitrarily large size, then we say $\text{VCdim}(\mathcal{H}) = \infty$.

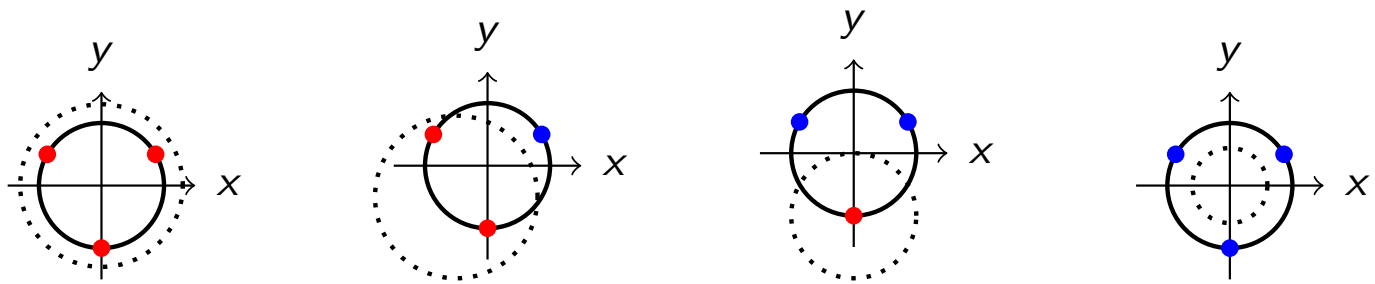
Consider the following example where $\mathcal{X} = \{(x, y) : x^2 + y^2 = 1\}$ and $\mathcal{H} = \{h_{(h,k,r)} : h, k, r \in \mathbb{R}\}$ where

$$h_{(h,k,r)}(x, y) = \begin{cases} 1 & \text{if } (x-h)^2 + (y-k)^2 \leq r^2 \\ 0 & \text{if } (x-h)^2 + (y-k)^2 > r^2 \end{cases}.$$

We claim that $\text{VCdim}(\mathcal{H}) = 3$. Suffices to show $\mathcal{H}$ can shatter C when $|C| = 3$ but not when $|C| = 4$. Consider the set $C = \{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$.

Circle Example

We can show that the function $f : \mathcal{H}_C \to \{0,1\}^{|C|}$ defined by the canonical mapping is surjective through casework on the labels of the elements in C. Rotational symmetry just leaves us with 4 cases, as shown below.



This implies that $|H_C| \geq \left|2^{|C|}\right|$ but we also know that $|H_C| \leq \left|2^{|C|}\right|$ so $|H_C| = \left|2^{|C|}\right|$, and we are done. To show $\mathcal{H}$ can't shatter C when $|C| = 4$, consider an arbitrary arrangement that is labeled in alternating fashion in a clockwise direction. It can be shown geometrically there is no hypothesis that can achieve this labeling.

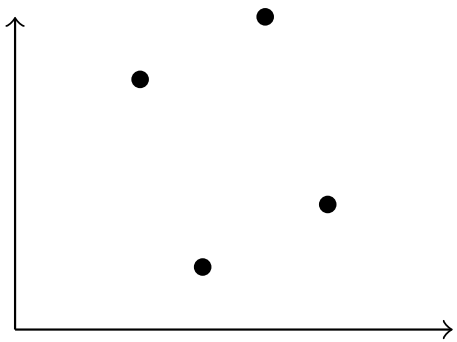
Axis Aligned Rectangles Revisited

There is a very important connection between VC dimension and PAC learnability.

Theorem

A hypothesis class $\mathcal{H}$ is PAC learnable iff $\text{VCdim}(\mathcal{H})$ is finite.

So instead of doing all the probability earlier, we can just compute a single number to determine if $\mathcal{H}_{\text{rectangles}}$ is PAC learnable. We will claim that $\text{VCdim}(\mathcal{H}) = 4$. Let $C = \{(x_1, y_1), \ldots, (x_4, y_4)\}$ such that they have the following arrangement



Calculate Dimension

- It suffices to check all $2^4 = 16$ possible labelings to show that the canonical mapping is indeed surjective. $+$
- To show $\mathcal{H}$ can't shatter when $|C| = 5$, consider an arbitrary arrangement of 5 points.
- Define $S = \{(x, y) : x \notin \left(\max\{x_i\}_{i=1}^5 \vee \min\{x_i\}_{i=1}^5\right) \wedge y \notin \left(\max\{y_i\}_{i=1}^5 \vee \min\{y_i\}_{i=1}^5\right)\}$.
- By the Pigeonhole principle, S is nonempty. Label everything inside S as 0 and everything else as 1.
- The ERM hypothesis $h_{(a_1,a_2,b_1,b_2)}$ must have $a_1 = \min\{x_i\}_{i=1}^5$, $b_1 = \max\{x_i\}_{i=1}^5$, $a_2 = \min\{y_i\}_{i=1}^5$, and $b_2 = \max\{y_i\}_{i=1}^5$
- This contradicts coloring of S, so we are done.

Future Directions

We've shown what it means to learn and we've also answered questions like "what things can be learn" as well as "under what conditions can we learn certain things?" Here are some things you might think about:

- Can we give a definition of PAC learnability without the realizability assumption?
- Can we generalize for a general loss function?
- We've been working with binary classification tasks in a supervised setting. Can we expand the scope of tasks we can analyze?

Thank you for coming!