STATISTICAL LEARNING THEORY

Summer Term 2024

Resource: Shai Shalev-Schwartz
Understanding Machine Learning

Online Notes: http://rkwitt.org → Teaching
(All my handwritten notes will be available there!)

OTHER COURSE NAMES:

1) Advanced machine learning
2) Machine learning

MOTIVATION

A motivating example:

	Weight (in g)	Color ($\in [0,1]$)	Tasty
Papaya 1	800	0.1	0
⋮	⋮	⋮	⋮
Papaya N	1200	0.9	1

0 ... Not tasty
1 ... Tasty

Lets call the information in this table our **training data**. Based on that we aim to find

$$h: \mathbb{R} \times \mathbb{R} \to \{0,1\}$$

i.e, a function that will take a two-dim. vector as input (weight and color) and output a prediction of wether a papaya is tasty (1) or not (0). We call such a function a **hypothesis**.

More formally, the data available to us comes in the form

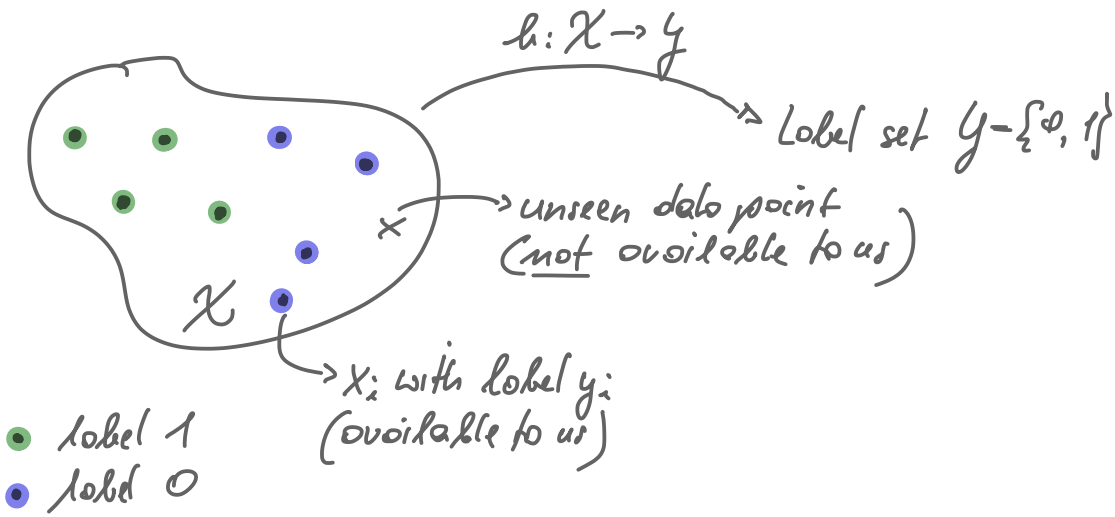
$$((x_1, y_1), \ldots\ldots (x_N, y_N)) = S$$

with $x_i \in \mathbb{R}^2$ and $y \in \{0, 1\}$.

$x_i \in \mathbb{R}^2$ → Domain $X = \mathbb{R}^2$

$y \in \{0, 1\}$ → Label set $Y = \{0, 1\}$

What do we mean by "learning"



We say, a **learner** (some algorithm) receives S and outputs h!

Two assumptions we will make initially:

(1) All the x_i's are drawn identically and independently (iid) from some (unknown) distribution D over the domain X.

(2) The x_i's are labeled by some (unknown) function

$$f: X \to Y$$

which we call the true labeling function, meaning

$$S = \big((x_1, \underbrace{f(x_1)}_{y_1}\big), \ldots\ldots \big(x_N, \underbrace{f(x_N)}_{y_N}\big)$$

What do we care about?

We care about

$$A = \{x \in X : h(x) \neq f(x)\}$$

That is, all the points x in our domain X, where the hypothesis h differs from the true labeling function f.

BACKGROUND

(a really short primer on required concepts)

In the following, $\mathcal{P}(X)$ will denote the power set of X.

Remember that we care about

$$A = \{x \in X : h(x) + f(x)\}$$

↳ domain

In case of $X = \mathbb{R}^n$ and $A \in \mathcal{P}(X)$, we already have a problem.

Measure Problem: find $\mu : \mathcal{P}(\mathbb{R}^n) \to [0, \infty]$ with properties:

1. $A_i \in \mathcal{P}(\mathbb{R}^n)$, $i \in \mathbb{N}$, pairwise disjoint, we want
$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$

2. if $A, B \in \mathcal{P}(\mathbb{R}^n)$ are congruent, we want
$$\mu(A) = \mu(B)$$

3. we want
$$\mu\left([0,1]^n\right) = 1$$

The measure problem is unsolvable for all $n \in \mathbb{N}$!

Solution: We will construct ourselves to elements of some σ-Algebra over the domain X.

Def. (σ-Algebra): Let S be a non-empty set. A family of sets $\mathcal{F} \subset \mathcal{P}(S)$ is called a σ-Algebra over S, if

1. $S \in \mathcal{F}$
2. from $A \in \mathcal{F}$, it follows that $A^c = S \setminus A \in \mathcal{F}$
3. from $A_i \in \mathcal{F}$ with $i \in \mathbb{N}$, it follows that

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F} \qquad \text{(i.e. for any countable collection)}$$

Example(s) (smallest σ-Alg. that contains A):

$$\{\emptyset, A, A^c, S\}$$

$S = \{a, b, c, d\}$ with σ-Algebra $\{\emptyset, \{a, b\}, \{c, d\}, \{a, b, c, d\}\}$

Def. (generator): Given $\mathcal{E} \subset \mathcal{P}(S)$ a family of sets and Σ denoting the set of **all** σ-Algebras that contain $\mathcal{E}$, we call

$$\sigma(\mathcal{E}) = \bigcap_{\mathcal{F} \in \Sigma} \mathcal{F}$$

the σ-Alg. **generated by** $\mathcal{E}$. Further, if it holds that for a σ-Alg. $\mathcal{A}$

$$\sigma(\mathcal{E}) = \mathcal{A}$$

we call $\mathcal{E}$ the **generator** of $\mathcal{A}$.

Example: $\mathcal{E} = \{\{1\}\}$, $S = \{1,2,3\}$

$$\sigma(\mathcal{E}) = \sigma(\{\{1\}\})$$
$$= \{\emptyset, \{1\}, \{2,3\}, \{1,2,3\}\}$$

Def. (Topological Space): A topological space is a tuple (X, τ) with X a set and τ a collection of subsets of X s.t.

1. $\emptyset \in \tau$, $X \in \tau$
2. closed under union, i.e., $\{T_i\}_{i \in I} \subseteq \tau \Rightarrow \bigcup_{i \in I} T_i \in \tau$
3. closed under finite intersection, i.e.,

$$\{T_i\}_{i=1}^{n} \subseteq \tau \longrightarrow \bigcap_{i=1}^{n} T_i \in \tau$$

Remark: The elements of τ are called open sets.

Example: $X = \{1,2,3,4\}$

$$\tau = \{\emptyset, \{1,2,3,4\}, \{2\}, \{1,2\}, \{2,3\}, \{1,2,3\}\}$$

Def. (BOREL σ-Alg.): Given a topological space $(S, \mathcal{Q})$ with $\mathcal{Q}$ denoting the system of open sets, then we call

$$\sigma(\mathcal{Q}) =: \mathcal{B}(S)$$

the BOREL σ-Algebra over S. It's elements are called BOREL sets. For $S = \mathbb{R}^n$, we write

$$\mathcal{B}^n := \mathcal{B}(\mathbb{R}^n)$$

Importantly, each of the following systems of sets are generators for $\mathcal{B}^n$:

- $\{U \subset \mathbb{R}^n : U \text{ open}\}$
- $\{A \subset \mathbb{R}^n : A \text{ closed}\}$
- $\{]a, b] : a, b \in \mathbb{R}^n \text{ with } a \leq b\}$ — meaning $]a, b] =]a_1, b_1] \times \dots \times]a_n, b_n]$ for $a = (a_1 \dots a_n)$, $b = (b_1 \dots b_n)$
- $\{]-\infty, c] : c \in \mathbb{R}^n\}$

Convention: We extend $\mathbb{R}$ by symbols "$+\infty$", "$-\infty$" as

$$\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$$
$$\overline{\mathcal{B}} = \sigma(\mathcal{B} \cup \{-\infty\} \cup \{+\infty\})$$

Def. (Measureable space): if $\mathcal{F}$ is a σ-Algebra over S, we call $(S, \mathcal{F})$ a measurable space.

Example: $(\mathbb{R}^n, \mathcal{B}^n)$

Def. (Measureable function): Given $(S_1, \mathcal{F}_1)$ and $(S_2, \mathcal{F}_2)$ measureable spaces, we call

$$f: S_1 \to S_2$$

a measureable function if

$$\forall E \in \mathcal{F}_2 : f^{-1}(E) \subset \mathcal{F}_1$$

Example: $\mathbb{1}_A : S \to \mathbb{R}$ // characteristic function

$$S \ni \omega \mapsto \mathbb{1}_A(\omega) = \begin{cases} 1, \text{if } \omega \in A \\ 0, \text{else} \end{cases}$$

Consider $\mathbb{1}_A$ as a function mapping to $\mathbb{R}$:

if $B \leq 0$, then $\{\omega : \mathbb{1}_A(\omega) < B\} = \emptyset$

if $B > 1$, then $\{\omega : \mathbb{1}_A(\omega) < B\} = S$

if $0 < B \leq 1$, then $\{\omega : \mathbb{1}_A(\omega) < B\} = S \setminus A = A^c$ (all measurable)

$\Rightarrow \mathbb{1}_A$ is measureable if $A \in \mathcal{F}$.

Def. (Measure): Let $(S, \mathcal{F})$ be a measureable space. A function $\mu: \mathcal{F} \to \overline{\mathbb{R}}$ is called a measure if the following conditions hold:

1. $\mu(\emptyset) = 0$
2. $\mu(A) \geq 0$ for all $A \in \mathcal{F}$
3. for every sequence $(A_n)_{n \in \mathbb{N}}$ of disjoint sets from $\mathcal{F}$, we have

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) \quad \text{// } \sigma\text{-Additivity}$$

Example (counting measure): $\mu_{\text{COUNT}}: \mathcal{F} \to \overline{\mathbb{R}}$

$$A \mapsto \begin{cases} |A|, & \text{if } A \text{ is finite} \\ \infty, & \text{else} \end{cases}$$

Def. (Measure space): If $(S, \mathcal{F})$ is a measurable space and $\mu: \mathcal{F} \to \overline{\mathbb{R}}$ is a measure, we call

$$(S, \mathcal{F}, \mu)$$

a measure space.

Some elementary properties: Let $(S, \mathcal{F}, \mu)$ be a measure space. Also, let $A, B \in \mathcal{F}$ and $A_n \in \mathcal{F}$ with $n \in \mathbb{N}$. Then, it holds that

1. if A and B are disjoint, then

$$\mu(A \cup B) = \mu(A) + \mu(B)$$

2. if $A \subset B$ and $\mu(A) < \infty$, then

$$\mu(\underbrace{B \setminus A}_{\text{set difference}}) = \mu(B) - \mu(A)$$

3. if $A \subseteq B$, then

$$\mu(A) \leq \mu(B)$$

4. $$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mu(A_n)$$ // called σ-Additivity

↳

Def. (Probability space): Given $(S, \mathcal{F}, D)$ a measure space and

$$D(S) = 1,$$

then we call $(S, \mathcal{F}, D)$ a *probability space*, and D a *probability measure*.

Remarks: S ... interpreted as the set of outcomes of a random experiment

$\mathcal{F}$ ... interpreted as the set of events to which we want to assign probabilities to.

D ... the prob. measure assigns to each event $E \in \mathcal{F}$ a probability $D(A) \in [0, 1]$

Another remark: Based on $(S_1, \mathcal{F}_1, \mu)$ a measure space, a measurable space $(S_2, \mathcal{F}_2)$ and a measurable function $f: S_1 \to S_2$, we can easily construct a *new* measure via:

$$\mu_f : \mathcal{F}_2 \to [0, \infty]$$

$$B \mapsto \mu_f(B) = \mu\left(\underbrace{f^{-1}(B)}_{\text{pre-image of } B \text{ under } f}\right)$$

We call μ_f the "push-forward" measure of μ under f.

Def. (Random Variable): Given $(S_1, \mathcal{F}_1, \mathbb{D})$ a prob. space and $(S_2, \mathcal{F}_2)$ a measurable space, we call a measurable function

$$X : S_1 \longrightarrow S_2$$

a <u>random variable</u>. If $S_2 = \mathbb{R}$, we say X is a <u>real random var.</u>

Also, the push-forward measure $\mathbb{P}_X$ on S_2 is a prob. measure, since

$$\mathbb{P}_X(S_2) = \mathbb{D}\big(\underbrace{X^{-1}(S_2)}_{S_1}\big) = \mathbb{D}(S_1) = 1$$

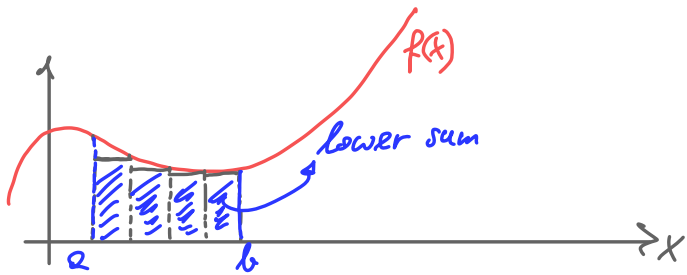
we call $\mathbb{P}_X$ the <u>distribution of X.</u>

CONVENTION(S) on notation: X... random variable

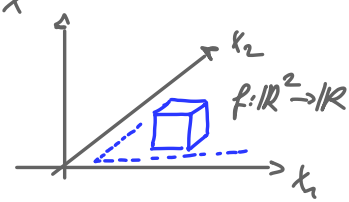
$$\{X \in A\} = \{\omega \in S : X(\omega) \in A\}$$
$$\{X = c\} = \{\omega \in S : X(\omega) = c\}$$
$$\mathbb{P}_X(A) = \mathbb{D}\big(\{\omega \in S : X(\omega) \in A\}\big)$$
$$\mathbb{P}_X(\{c\}) = \mathbb{D}\big(\{\omega \in S : X(\omega) = c\}\big)$$

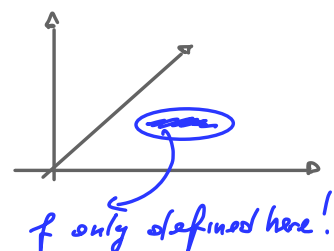
A very short primer on the LEBESGUE INTEGRAL

Key idea of Riemann integral: $f: \mathbb{R} \to \mathbb{R}$



1) Difficult to extend to higher dimensions
2) Reliance on continuity
3) ⋮

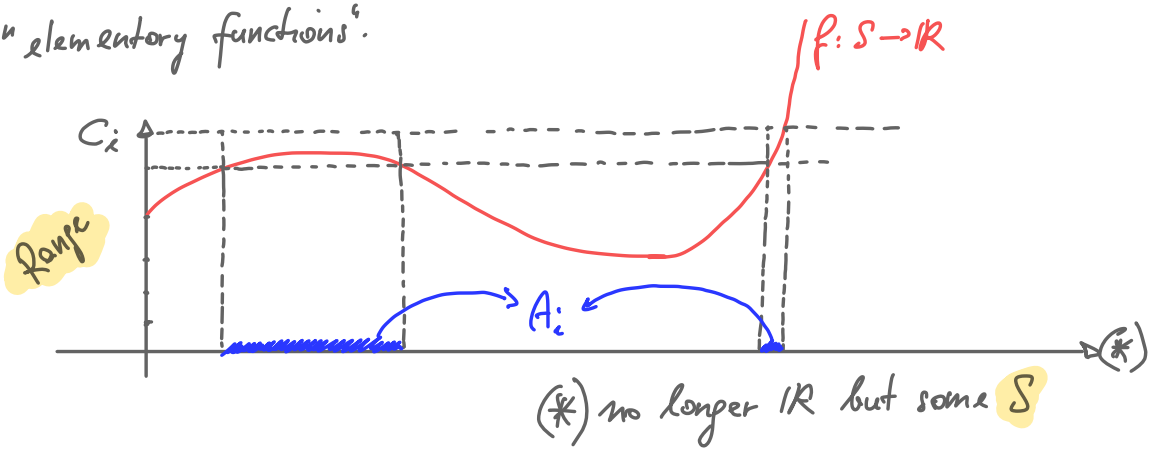




Question: How to partition this domain?

B.P.

Idea of LEBESGUE integration: Partition the range of a function $f: S \to \mathbb{R}$ into intervals to get to an approximation by so called "elementary functions".



We need a way to measure A_i. If we can do that, we can write

$$c_i \cdot \mu(A_i)$$

μ → measure

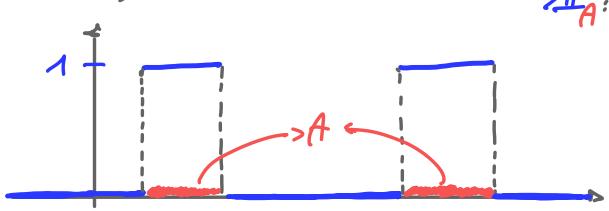
..... this allows us to eventually write

$$\sum_i c_i \cdot \mu(A_i) \quad \underline{\text{OR}} \quad \int_S f \, d\mu$$

A little bit more detail: Lets look at the case of "simple" functions.

Say, we have $(S, \mathcal{F}, \mu)$, $\mu: \mathcal{F} \to [0, \infty]$

$\mathbb{1}_A: S \to \mathbb{K}$, $A \in \mathcal{F}$



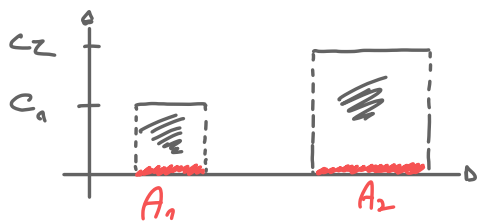
Any reasonable notion of integration of $\mathbb{1}_A$ should return $\mu(A)$:

$$I(\mathbb{1}_A) = \mu(A)$$

Integral

finite here!

Now, a "simple" function can be written as $f(x) = \sum_{i=1}^{n} c_i \cdot \mathbb{1}_{A_i}(x)$ with $A_i \in \mathcal{F}$ and $c_i \in \mathbb{R}$. Visually:



We can define the integral of f via: $I(f) = \sum_{i=1}^{n} c_i \mu(A_i)$. But, since $\mu: \mathcal{F} \to [0, \infty]$ and $c_i \in \mathbb{R}$, we could get something like

$$10 \cdot \infty - 8 \cdot \infty$$

Solution: $\{f: S \to \mathbb{R} \mid f \text{ is "simple" and } f \geq 0\} = T^+$. For $f \in T^+$ we have a well-defined representation of f as

$$f(x) = \sum_{i=1}^{n} c_i \cdot \mathbb{1}_{A_i}(x) \quad \text{with } c_i \geq 0$$

and we define the LEBESGUE integral (for this class) as:

$$I(f) := \sum_{i=1}^{n} c_i \cdot \mu(A_i) \quad \text{or} \quad \int f \, d\mu \quad \text{or} \quad \int_{x \in S} f(x) \, d\mu(x)$$

Some properties: (that follow immediately) IMPORTANT!

for $f, g \in T^+$ and $\alpha, \beta \geq 0$

(A) $\int (\alpha f + \beta g)\, d\mu = \alpha \cdot \int f\, d\mu + \beta \cdot \int g\, d\mu$ LINEARITY

for $f, g \in T^+$ and $f \leq g$

(B) $\int f\, d\mu \leq \int g\, d\mu$ MONOTONICITY

Overall, even though we just looked at "simple" functions, the construction extends to all measurable functions $f: S \to \mathbb{R}$.

(also (A) and (B) hold for all measurable functions)

Def. (Expected value of a RV): Given a prob. space $(S, \mathcal{F}, \mathbb{P})$ and a (quasi)-integrable real RV $X: S \to \mathbb{R}$, then

$$\mathbb{E}[X] = \int X \, d\mathbb{P}$$

is called the expected value of X.

We have:

- If $X \geq 0$, then $\mathbb{E}[X] \geq 0$
- Linearity: $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ for RVs X and Y
  $\mathbb{E}[aX] = a\,\mathbb{E}[X]$ for constant a

Def. (MARKOV inequality): Given prob. space $(S, \mathcal{F}, \mathbb{P})$ and a non-negative RV X (≥ 0), we have for $a > 0$

$$\underbrace{\mathbb{P}[X \geq a]}_{\mathbb{P}(\{\omega \in S: X(\omega) \geq a\})} \leq \frac{\mathbb{E}[X]}{a}$$

Proof: Define $\varphi: S \to \mathbb{R}$ with $\varphi(x) = \begin{cases} a, & \text{if } X \geq a \\ 0, & \text{else } (X < a) \end{cases}$

$\Rightarrow 0 \leq \varphi(x) \leq X(x)$. Hence, by monotonicity we have $\int X \, d\mathbb{P} \geq \int \varphi \, d\mathbb{P} = a \cdot \mathbb{P}(\{\omega \in S: X(\omega) \geq a\})$.

$\Rightarrow$ as $a > 0$, we have $\frac{1}{a} \cdot \int X \, d\mathbb{P} \geq \mathbb{P}(\{\dots\})$

and since $\mathbb{P}(\{\omega \in S : X(\omega) \geq a\}) = \mathbb{P}[X \geq a]$, we get

$$\mathbb{P}[X \geq a] \leq \frac{\int X \, d\mathbb{P}}{a} = \frac{\mathbb{E}[X]}{a}$$

□

CONTINUING ML MATERIAL

1. Domain set $\mathcal{X}$; we call $x \in \mathcal{X}$ an instance
2. Label set $\mathcal{Y}$, e.g., $\mathcal{Y} = \{0, 1\}$
3. Training data set $S = ((x_1, y_1), \ldots, (x_m, y_m))$ with $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$
4. A learner that receives S and outputs $h: \mathcal{X} \to \mathcal{Y}$ which we call a hypothesis.

Assumption: For now, we assume the x_i's are drawn iid from some probability measure D over the domain $\mathcal{X}$ and labeled by some unknown function $f: \mathcal{X} \to \mathcal{Y} : y_i = f(x_i)$

We are interested in:

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

"Generalization error"

The empirical version of this is

$$\frac{1}{m} \cdot \left|\{i \in [m] : h(x_i) \neq f(x_i)\}\right| = L_S(h)$$

"Empirical error" (empirical risk)

$[m] = \{1, \ldots, m\}$

Convention: $S|_x = (x_1, \dots, x_m)$

Claim: $\mathop{\mathbb{E}}_{S|_x \sim D^m}[L_S(h)] = L_{D,f}(h)$ ($\mathbb{E}$ ... expected value)

$$\mathop{\mathbb{E}}_{S|_x \sim D^m}[\underbrace{L_S(h)}_{\text{empirical error}}] = \mathop{\mathbb{E}}_{S|_x \sim D^m}\left[\frac{1}{m}\cdot\sum_{i=1}^{m} \mathbb{1}_{h(x_i)\neq f(x_i)}\right] \quad \text{// by def.}$$

$$= \frac{1}{m}\cdot\sum_{i=1}^{m} \mathop{\mathbb{E}}_{x_i\sim D}\left[\mathbb{1}_{h(x_i)\neq f(x_i)}\right] \quad \text{// by linearity}$$

$$= \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{x\sim D}\left[\mathbb{1}_{h(x)\neq f(x)}\right] \quad \text{// as all } x_i\text{'s are drawn i.i.d from } D$$

$$= \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{P}}_{x\sim D}\left[h(x)\neq f(x)\right]$$

$$= \frac{1}{\cancel{m}}\cdot\cancel{m}\cdot \mathop{\mathbb{P}}_{x\sim D}\left[h(x)\neq f(x)\right]$$

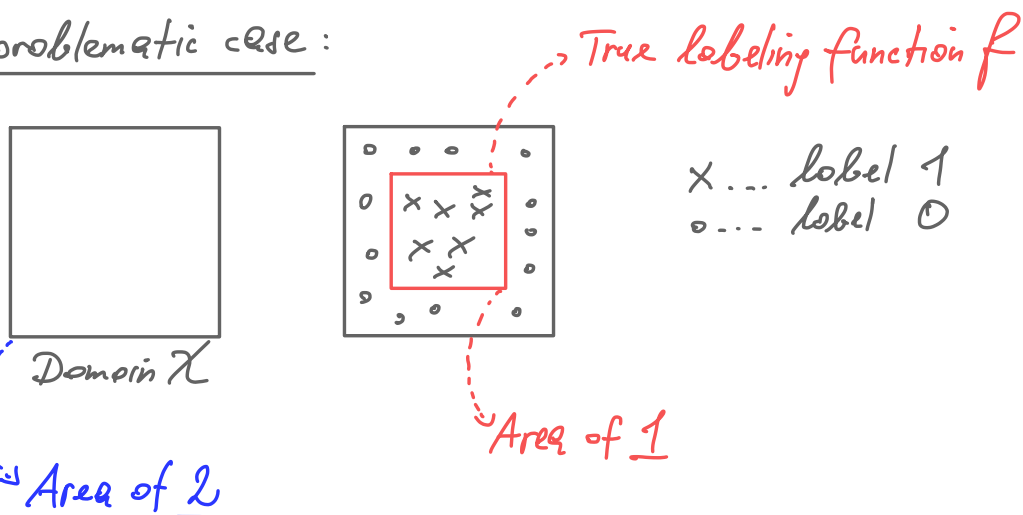
$$= L_{D,f}(h) \quad \text{generalization error}$$

This establishes the claim.

Our first learning paradigm

Empirical risk minimization (ERM): As we only have access to the training data (S), it's natural to try to select h such that the empirical risk (empirical error) is minimized. We call such a h an **empirical risk minimizer**.

A problematic case:



- Say the distribution on X is uniform
- Say we have an ERM algorithm that returns h_S s.t.

$$h_S(x) = \begin{cases} y_i, & \text{if } \exists i \in \{1, \dots, m\} : x_i = x \\ 0, & \text{else} \end{cases}$$

// i.e., a lookup table

Obviously, h_S is correct on all instances in $S \rightarrow h_S$ is an empirical risk minimizer, meaning

$$L_S(h_S) = 0 !$$

But, on unseen instances from our distribution on $\mathcal{X}$ (uniform), h_S is only correct 50% of the time, i.e.,

$$L_{D,f}(h_S) = \frac{1}{2} !$$

This is called overfitting.

Hypothesis class ($\mathcal{H}$): We restrict searching for h to $\mathcal{H}$, i.e., a class of functions from $\mathcal{X} \rightarrow \mathcal{Y}$ and we write

$$ERM_{\mathcal{H}}(S) \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h)$$

ERM over finite hypothesis classes ($|H| < \infty$)

Assumption (realizability): $\exists h^* \in H$ with $L_{D,f}(h^*) = 0$.

Now, any ERM hypothesis h_S will attain 0 empirical error ($L_S(h_S) = 0$), as h_S competes with h^* (which obviously has 0 empirical error).

Hence, $L_{D,f}(h_S) > \varepsilon$ can only happen if we select a hypothesis with $L_S(h_S) = 0$ but $L_{D,f}(h_S) > \varepsilon$.

We define

$$H_{BAD} = \{h \in H : L_{D,f}(h) \geq \varepsilon\}$$ // set of BAD hypothesis

Further, we define

$$M = \{S|_x : \exists h \in H_{BAD}, L_S(h) = 0\}$$

those are the ones with generalization error of $> \varepsilon$

Observation:

$$\{S|_x : L_{D,f}(h_S) \geq \varepsilon\} \subseteq \underbrace{\{S|_x : \exists h \in H_{BAD}, L_S(h) = 0\}}_{M \text{ (from before)}}$$

(A) empirical risk minimizer

$$M = \bigcup_{h \in H_{BAD}} \{S|_x : L_S(h) = 0\}$$

We get (upon measuring with $\mathcal{D}$):

$$\mathcal{D}^m\left(\{S|_x : L_{\mathcal{D},f}(h_S) \geq \varepsilon\}\right) \leq \mathcal{D}^m\left(\bigcup_{h \in H_{BAD}} \{S|_x : L_S(h) = 0\}\right)$$

"union bound" (σ-sub-additivity)

$$\leq \sum_{h \in H_{BAD}} \mathcal{D}^m\left(\{S|_x : L_S(h) = 0\}\right)$$

Let's fix some $h \in H_{BAD}$:

$$\mathcal{D}^m\left(\{S|_x : L_S(h) = 0\}\right) = \mathcal{D}^m\left(\{S|_x : \forall i \in [m] : h(x_i) = f(x_i)\}\right)$$

as all x_i's are iid

$$= \prod_{i=1}^{m} \mathcal{D}\left(\{x_i : h(x_i) = f(x_i)\}\right)$$

$$= \prod_{i=1}^{m} \mathcal{D}\left(\{x : h(x) = f(x)\}\right)$$

$$= \prod_{i=1}^{m} \left(1 - L_{\mathcal{D},f}(h)\right)$$

$$\leq \prod_{i=1}^{m} (1 - \varepsilon)$$ as h is a bad hypothesis $h \in H_{BAD}$

$$= (1-\varepsilon)^m$$

$$\leq e^{-\varepsilon m}$$ (without proof)

$[m] = \{1, \ldots, m\}$

$$\Rightarrow D^m\left(\{S|_x : L_{D,f}(h_S) \geq \varepsilon\}\right) \leq \sum_{h \in H_{BAD}} \underbrace{D^m\left(\{S|_x : L_S(h) = 0\}\right)}_{\leq e^{-\varepsilon m}}$$

$$\leq |H_{BAD}| \cdot e^{-\varepsilon m}$$

$$\leq |H| \cdot e^{-\varepsilon m} \quad \text{// because } H_{BAD} \subseteq H$$

If we want $|H| \cdot e^{-\varepsilon m}$ to be less than some $\delta \in (0,1)$, we can solve for m and get:

$$|H| \cdot e^{-\varepsilon m} < \delta$$

$$\Rightarrow m > \frac{1}{\varepsilon} \cdot \log\left(\frac{|H|}{\delta}\right)$$

(ε: error, δ: confidence)

<u>Corollary:</u> Let $|H| < \infty$ and $\varepsilon, \delta \in (0,1)$. Further, let m be an integer such that $m > \frac{1}{\varepsilon} \cdot \log\left(\frac{|H|}{\delta}\right)$. Then, for each labeling $f: X \to Y$ and any distribution D over X (for which we have that with probability of at least $1-\delta$ over the $S|_x$ (of size m) every ERM hypothesis h_S satisfies

$$L_{D,f}(h_S) \leq \varepsilon.$$

<u>Interpretation:</u> For sufficiently large m, ERM_H returns h_S that is <u>PROBABLY APPROXIMATELY CORRECT (PAC)</u>.

This leads to:

Def. (PAC learnability): A hypothesis class H is PAC learnable, if there exists a function $m_H : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A with the following properties: (I) for every $\varepsilon, \delta \in (0,1)$ and (II) every distribution D over the domain X, and (III) every labeling function $f: X \to \{0,1\}$, if (IV) realizability holds (with respect to H, D, f), then running A on $m \geq m_H(\varepsilon, \delta)$ iid instances from D (labeled by f) returns a hypothesis h such that with probability of at least $1-\delta$ (over the choice of S)

$$L_{D,f}(h) \leq \varepsilon$$

Def. (Sample complexity): $m_H : (0,1)^2 \to \mathbb{N}$ is called the sample complexity function.

04/16/24

We will now move to a more general setting.

1. We will first release the realizability assumption.
(in this setting, the best thing we can hope for are guarantees relative to the "best possible" hypothesis in the class $\mathcal{H}$, i.e., relative to $\min_{h \in \mathcal{H}} L_{\mathcal{D},f}(h)$).

Def. (Höffding inequality): Let $X_1, \dots, X_m$ be iid random variables taking values in $[a_i, b_i]$ for $i \in [m]$. Then, given that

$$S_m = \sum_{i=1}^{m} X_i$$

it holds that

$$\mathbb{P}\left[S_m - \mathbb{E}[S_m] > \varepsilon\right] \leq e^{-2\varepsilon^2 / \sum_{i=1}^{m} (b_i - a_i)^2}$$

$$\mathbb{P}\left[S_m - \mathbb{E}[S_m] < -\varepsilon\right] \leq e^{-2\varepsilon^2 / \sum_i (b_i - a_i)^2}$$

Also:
$$\mathbb{P}\left[\left|S_m - \mathbb{E}[S_m]\right| > \varepsilon\right] \leq 2 \cdot e^{-2\varepsilon^2 / \sum_i (b_i - a_i)^2}$$

Another useful form of this inequality is

$$\mathbb{P}\left[\left|\frac{1}{m}\cdot S_m - \mu\right| > \varepsilon\right] \leq 2\cdot e^{-2\varepsilon^2 m/(b-a)^2} \qquad (*)$$

with $\mu = \mathbb{E}[X_i]$ and $\mathbb{P}[a \leq X_i \leq b] = 1$ for all $i \in [m]$.

As a consequence of $(*)$, we can say the following: fix $\varepsilon > 0$; then for a single $h : X \to Y$, we have

$$\mathbb{P}_{S|_X \sim D^m}\left[\left|\underbrace{L_S(h)}_{\text{corresponds to } S_m \text{ from above}} - L_{D,f}(h)\right| > \varepsilon\right] \leq 2\cdot e^{-2\varepsilon^2 m} \quad (\text{as } b=1, a=0)$$

(If we would set $2e^{-2\varepsilon^2 m} = \delta$ and solve for ε, we would get

$$\varepsilon = \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2m}}$$

$\Rightarrow L_{D,f}(h) \leq L_S(h) + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2m}}$; this holds with probability of at least $1-\delta$ over the choice of S!)

Remark: This result only holds for a single h. However, we can easily get a bound that holds uniformly for all $h \in H$, with $|H| < \infty$:

$$\mathbb{P}_{S|_x \sim D^m}\left[\exists h \in H : |L_S(h) - L_{D,f}(h)| > \varepsilon\right] \overset{\text{union bound}}{\leq} \sum_{h \in H} 2e^{-2\varepsilon^2 m}$$

$$\leq 2 \cdot |H| \cdot e^{-2\varepsilon^2 m}$$

(no realizability needed for that result)

2 Next, we release our requirement of a "true labeling function" f. We do this by letting D be a distribution over $X \times Y - Z$. Hence, we need to adjust our definitions of the empirical error and the generalization error.

(1) $L_D(h) = \mathbb{P}_{(x,y) \sim D}\left[h(x) \neq y\right] = D\left(\{(x,y) \in X \times Y : h(x) \neq y\}\right)$

(generalization error)

(2) $L_S(h) = \frac{1}{m} \cdot \left|\{i \in [m] : h(x_i) \neq y_i\}\right|$

(empirical error)

Def. (Agnostic PAC learnability): A hypothesis class H is agnostic PAC learnable if there exists $m_H : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A with the following properties: (I) for every $\varepsilon, \delta \in (0,1)$ and (II) every distribution D over $X \times Y = Z$, when running A on $m \geq m_H(\varepsilon, \delta)$ iid instances from D, A returns a hypothesis h such that with probability of at least $1-\delta$ over the choice of S

$$L_D(h) \leq \min_{h' \in H} L_D(h') + \varepsilon.$$

General 'loss' functions

$$\ell : H \times (X \times Y) \longrightarrow \mathbb{R}_+$$

Example: 0-1 loss

$$\ell^{0-1}(h, (x,y)) = \begin{cases} 1, & \text{if } h(x) \neq y \\ 0, & \text{else} \end{cases}$$

(Square loss

$$\ell^{SQ}(h, (x,y)) = (h(x) - y)^2 \quad \text{// used in regression problems}$$

$\hookrightarrow \in \mathbb{R}$)

Now, even with more generality:

$$L_D(h) = \underset{z \sim D}{\mathbb{E}} [\ell(h, z)], \quad L_S(h) = \frac{1}{m} \cdot \sum_{i=1}^{m} \ell(h, z_i) \qquad z = (x, y)$$

1st exam date: July, 2, 2024 (10 am)

2nd exam date July, 16, 2024 (10 am)

Remark: $\mathbb{E}_{z\sim D}\left[\ell^{0-1}(h,z)\right] = 0 \cdot \mathbb{P}_z\left[h(x)=y\right] + 1 \cdot \mathbb{P}_z\left[h(x)\neq y\right]$

$$= \mathbb{P}_z\left[h(x)\neq y\right]$$

Uniform convergence

Def. (ε-representative sample): A sample S is called ε-representative with respect to $Z = X \times Y$, hypothesis class H, loss function ℓ and distribution D, if

$$\forall h \in H: \quad |L_S(h) - L_D(h)| \leq \varepsilon.$$

Lemma: Assume that S is $\varepsilon/2$-representative wrt. Z, H, ℓ and D. Then, any hypothesis h_S returned by $ERM_H(S) \in \operatorname{argmin}_{h' \in H} L_S(h')$ satisfies

$$L_D(h_S) \leq \min_{h \in H} L_D(h) + \varepsilon.$$

Proof: for any $h \in H$:

$$L_D(h_S) \leq L_S(h_S) + \varepsilon/2 \quad \text{// by definition of } \varepsilon\text{-rep.}$$

ERM hyp.

$$\leq L_S(h) + \varepsilon/2 \quad \text{// as } h_S \text{ is an ERM hypothesis}$$

$$\leq L_D(h) + \varepsilon/2 + \varepsilon/2 \quad \text{// by def. of } \varepsilon\text{-rep.}$$

$$= L_D(h) + \varepsilon$$

$$\Rightarrow L_D(h_S) \leq L_D(h) + \varepsilon$$

As this inequality chain holds for any $h \in H$, we can conclude that

$$L_D(h_S) \leq \min_{h \in H} L_D(h) + \varepsilon$$

□

Def. (uniform convergence): A hyp. class H has the uniform convergence property (UC) with respect to Z and loss function ℓ, if there exists $m_H^{UC}: (0,1)^2 \to \mathbb{N}$, such that for any $\varepsilon, \delta \in (0,1)$ and every distribution D over $Z = X \times Y$, if S is an iid sample of size $m \geq m_H^{UC}(\varepsilon, \delta)$ from D, then with probability of at least $1-\delta$ (over the choice of S), S is ε-representative.

Corollary: If H has the UC property with m_H^{UC}, then H is agnostic PAC learnable with

$$m_H(\varepsilon, \delta) \leq m_H^{UC}\left(\frac{\varepsilon}{2}, \delta\right)$$

We also know that ERM is a successful agnostic PAC learner.

Claim: Finite hypothesis classes ($|H| < \infty$) are agnostic PAC learnable.

Fix $\varepsilon, \delta \in (0,1)$. We want to show that

$$D^m\left(\{S : \forall h \in H : |L_S(h) - L_D(h)| \leq \varepsilon\}\right) \geq 1 - \delta \quad (*)$$

Rephrasing $(*)$ gives

$$D^m\Big(\underbrace{\{S : \exists h \in H : |L_S(h) - L_D(h)| > \varepsilon\}}_{\bigcup_{h \in H} \{S : |L_S(h) - L_D(h)| > \varepsilon\}}\Big) < \delta$$

By the union bound

$$D^m\left(\bigcup_{h \in H} \{S : |L_S(h) - L_D(h)| > \varepsilon\}\right) \leq \sum_{h \in H} D^m\left(\{S : |L_S(h) - L_D(h)| > \varepsilon\}\right)$$

We now fix $\ell := \ell^{0-1}$ (0/1 loss). We know that $L_D(h) = \mathbb{E}_{z \sim D}[\ell(h, z)]$ and $L_S(h) = \frac{1}{m} \cdot \sum_{i=1}^{m} \ell(h, z_i)$. Hence, we can apply Hoeffding's inequality:

$$D^m\left(\{S : |L_S(h) - L_D(h)| > \varepsilon\}\right) \leq 2 \cdot e^{-2\varepsilon^2 m}$$

$$\Rightarrow D^m\left(\{S : \exists h \in H : |L_S(h) - L_D(h)| > \varepsilon\right) \leq 2 \cdot |H| \cdot e^{-2\varepsilon^2 m} \quad (**)$$

(fix 2023 notes)!

What remains is to let $2|H|e^{-2\varepsilon^2 m}$ be smaller than $\delta \in (0,1)$ and solve for m:

$$2|H|e^{-2\varepsilon^2 m} < \delta$$

$$\Rightarrow \boxed{m > \frac{1}{2\varepsilon^2} \cdot \log\left(\frac{2|H|}{\delta}\right)}$$

This yields the following sample complexity function m_H^{UC} for uniform convergence:

$$m_H^{UC}(\varepsilon, \delta) \leq \left\lceil \log\left(\frac{2|H|}{\delta}\right) \cdot \frac{1}{2\varepsilon^2} \right\rceil$$

(ceiling function)

→ By our previous corollary on the relation between UC and agnostic PAC learnability, we get

$$\boxed{m_H(\varepsilon, \delta) \leq m_H^{UC}\left(\frac{\varepsilon}{2}, \delta\right) \leq \left\lceil \frac{2}{\varepsilon^2} \cdot \log\left(\frac{2|H|}{\delta}\right) \right\rceil}$$

This establishes our claim. ∎

05/07/24

Question: Is there a "universal" learner?

By universal we mean a learner A without any prior knowledge of a learning task, but can be challenged by any task and still achieve low $L_D(A(S))$.

(task defined by D)

Theorem (No-Free-Lunch (NFL)): Let A be a learning algorithm for the task of binary (0/1) classification with respect to the 0-1 loss over a domain X. Also, let m be any number smaller than $|X|/2$, representing the training set size. Then, there exists a distribution D over $X \times \{0,1\}$ such that

1. $\exists f : X \to \{0,1\}$ with $L_D(f) = 0$ // 0 generalization error
2. with probability of at least $\frac{1}{7}$ over the choice of $S \sim D^m$, we have

$$L_D(A(S)) \geq \frac{1}{8}$$

(e.g. ERM_H with $H = \{f\}$)

Here, 1. means that the task can be successfully learned by another learner; 2. means that A fails on that task.

Proof: Let $C \subset X$ of size $2m$, i.e., $|C| = 2m$. The number of possible labelings of C is $2^{2m} = T$. We denote the functions that realize these T labelings as

$$\boxed{f_1, \ldots, f_T}$$

For each f_i, we define

$$\underbrace{D_i(\{(x,y)\})}_{\text{defines the learning task}} = \begin{cases} \frac{1}{|C|}, & \text{if } y = f_i(x) \\ 0, & \text{else} \end{cases}$$

Consequence of this construction: we have $\forall i : L_{D_i}(f_i) = 0$!

We will show now that for every learning algorithm A that receives samples from $X \times \{0,1\}$, there exists a function $f: X \to \{0,1\}$ and a distribution D over $X \times \{0,1\}$, such that

$$L_D(f) = 0 \quad \text{and} \quad \mathop{\mathbb{E}}_{S \sim D^m}\left[L_D(A(S))\right] \geq \frac{1}{4} \qquad (*)$$

In particular, we will show that for every learning algorithm A receiving m samples from $C \times \{0,1\}$ and returning a function $A(S): X \to \{0,1\}$, we have

$$\max_{i \in [T]} \mathop{\mathbb{E}}_{S \sim D_i^m}\left[L_{D_i}(A(S))\right] \geq \frac{1}{4} \qquad (**)$$

$[T] = \{1, \ldots, T\}$

Remark: (*) suffices for $L_D(A(S)) \le \frac{1}{8}$ with prob. of at least $\frac{1}{7}$ over the choice of $S \sim D^m$ (see PS).

Lets continue with (**): we have $(2m)^m = k$ possible training sequences of size m. We are going to call them

$$\boxed{S_1, \ldots, S_k}$$

Lets also call S_j^i the sequence S_j labeled by f_i, i.e.,

$$S_j^i = \big((x_1, f_i(x_1)), \ldots, (x_m, f_i(x_m))\big)$$

So, in case the distribution is D_i, then $S_1^i, \ldots, S_k^i$ are all possible training sequences (of size m) that A can receive. Also, by construction (of the D_i's), all these training sequences have equal probability of being drawn.

$$\Rightarrow \mathop{\mathbb{E}}_{S \sim D_i^m}\left[L_{D_i}(A(S))\right] = \frac{1}{k} \cdot \sum_{j=1}^{k} L_{D_i}\left(A(S_j^i)\right) \quad (***)$$

So, $\max_{i \in [T]} \mathop{\mathbb{E}}_{S \sim D_i^m}\left[L_{D_i}(A(S))\right] = \max_{i \in [T]} \frac{1}{k} \sum_{j=1}^{k} L_{D_i}\left(A(S_j^i)\right)$ // by (***)

// as max $\ge$ avg. $\quad \ge \frac{1}{T} \sum_{i=1}^{T} \frac{1}{k} \sum_{j=1}^{k} L_{D_i}\left(A(S_j^i)\right)$

$$= \frac{1}{k} \cdot \sum_{j=1}^{k} \frac{1}{T} \sum_{i=1}^{T} L_{D_i}\left(A(S_j^i)\right)$$

$$\geq \min_{j \in [k]} \frac{1}{T} \cdot \sum_{i=1}^{T} L_{D_i}\left(A(S_j^i)\right)$$

Lets fix some $j \in [k]$: As S_j is of size m, but C is of size $2m$, there are instances from C which we have not seen. Lets call them

$$v_1, \ldots, v_p$$

We know that $p \geq m$.

Now, for every $h: C \to \{0,1\}$ and every i, it holds that

$$L_{D_i}(h) = \frac{1}{2m} \cdot \sum_{x \in C} \mathbb{1}_{h(x) \neq f_i(x)}$$

$$\geq \frac{1}{2m} \sum_{r=1}^{p} \mathbb{1}_{h(v_r) \neq f_i(v_r)} \quad \text{// as the } v_r\text{'s are less than } 2m$$

$$\geq \frac{1}{2p} \cdot \sum_{r=1}^{p} \mathbb{1}_{h(v_r) \neq f_i(v_r)} \quad \text{// as } p \geq m$$

Combining the results so far:

$$\frac{1}{T} \sum_{i=1}^{T} L_{D_i}\left(A(S_j^i)\right) \geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{2p} \sum_{r=1}^{p} \mathbb{1}_{\underbrace{A(S_j^i)(v_r)}_{h} \neq f_i(v_r)}$$

05/28/25

$$\Rightarrow \frac{1}{T}\sum_{i=1}^{T} L_{D_i}\left(A(S_j^i)\right) \geq \frac{1}{2}\cdot\frac{1}{\rho}\sum_{r=1}^{\rho}\frac{1}{T}\cdot\sum_{i=1}^{T}\mathbb{1}_{A(S_j^i)(v_r)\neq f_i(v_r)}$$

$$\geq \frac{1}{2}\cdot \min_{r\in[\rho]} \frac{1}{T}\sum_{i=1}^{T}\mathbb{1}_{A(S_j^i)(v_r)\neq f_i(v_r)}$$

(again, as $\text{avg} \geq \min$)

Lets try fixing $r \in [\rho]$: we have T labeling functions $f_1, \dots, f_T$ (we have 2^{2^m} of them)

Lets say, we would have $m=1$, so $|C|=2$ (as an example)

	C_1	v_1 (unseen)
f_1	0	0
f_2	0	1
f_3	1	0
f_4	1	1

partition into 2 pairs

(f_1, f_2), (f_3, f_4)

pair 1, pair 2

where (f_1, f_2) only differ on v_1

and (f_3, f_4) only differ on v_1.

In general, we can always partition into $T/2$ disjoint pairs of labeling functions. For every pair $(f_i, f_{i'})$, we have that for every $c \in C$, $f_i(c) \neq f_{i'}(c)$ if and only if $c = v_r$. Further, for such a pair, we also have $S_j^i = S_j^{i'}$.

$$\Rightarrow \mathbb{1}_{A(S_j^i)(v_r) \neq f_i(v_r)} + \mathbb{1}_{A(S_j^{i'})(v_r) \neq f_i(v_r)} = 1 \qquad (\star\star\star\star)$$

Overall, we now have:

$\frac{1}{T} \cdot \sum_{i=1}^{T} \mathbb{1}_{A(S_j^i)(v_r) \neq f_i(v_r)}$ and we re-write this as a sum over all $T/2$ disjoint pairs $(f_i, f_{i'})$, knowing that $(\star\star\star\star)$ holds; we get

$$\frac{1}{T} \cdot \frac{T}{2} = \frac{1}{2}$$

$$\Rightarrow \frac{1}{T} \sum_{i=1}^{T} L_{D_i}\left(A(S_j^i)\right) \geq \frac{1}{4}$$

Remark: Our earlier steps of fixing j and r are irrelevant as we have a constant as a lower bound now.

Combining all partial results, yields:

$$\max_{i \in [T]} \underset{S \sim D_i^m}{\mathbb{E}} \left[L_{D_i}(A(S)) \right] \geq \frac{1}{4}$$

□

Corollary: Let X be an infinitely large domain and let H be the set of all functions from $X \to \{0,1\}$. Then, H is not PAC learnable.

Vapnik – Chervonenkis Dimension (VC Dimension)

(Knowing that in the NFL theorem what mattered most was the behavior of labeling functions on a dataset, it seems intuitive to tie learnability to this behavior.)

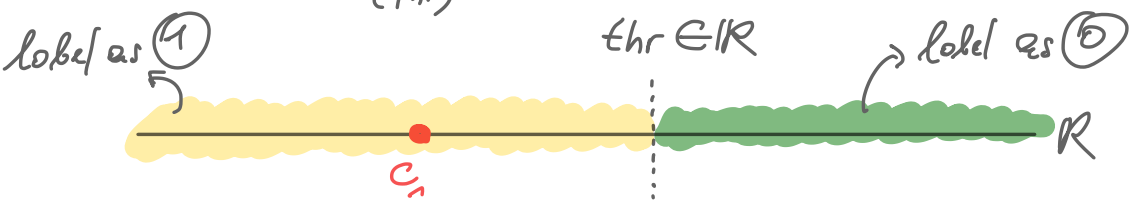
Def. Let H be a class of functions from $X \to \{0,1\}$ and let $C \subset X$, $C = \{c_1, \dots, c_m\}$. We define

$$H_C = \{(h(c_1), \dots, h(c_m)) : h \in H\}$$

as the restriction of H to C.

Example: $C = \{c_1\}, c_1 \in \mathbb{R}$ and lets look at the class of thresholds on the real line.

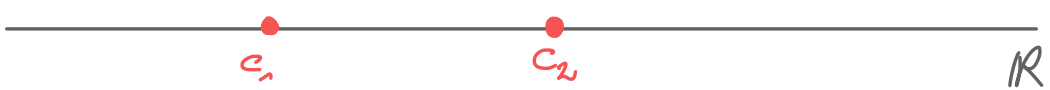
(thr)



Lets call this hyp. class H^{thr}. Then

$$H_C^{thr} = \{(1), (0)\} \Rightarrow |H_C^{thr}| = 2^1 = 2$$

Lets take $C = \{c_1, c_2\}$



$$H_C^{thr} = \{(1,1), (0,0), (1,0)\} \Rightarrow |H_C^{thr}| = 3$$

(all possible labelings would be $2^2 = 4$)

Def. H shatters a finite set C of size m ($|C| = m$), if

$$|H_C| = 2^m$$

Example (continued): The class of thresholds on $\mathbb{R}$ ($= X$) is PAC-learnable.

$$H^{thr} = \{h_a : a \in \mathbb{R}\},$$

$$h_a : \mathbb{R} \to \{0,1\}$$

$$x \mapsto h_a(x) = \mathbb{1}_{x<a} = \begin{cases} 1, & \text{if } x < a \\ 0, & \text{else} \end{cases}$$

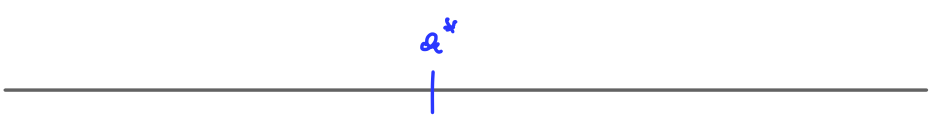
Claim: H^{thr} is PAC learnable with

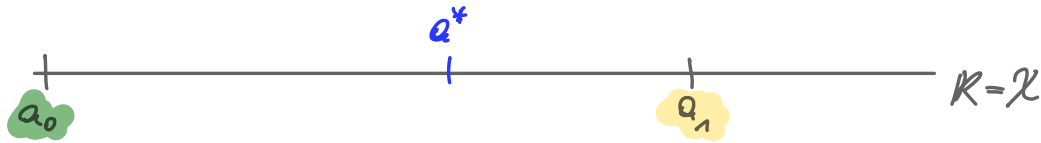
$$m_{H^{thr}}(\varepsilon, \delta) \leq \left\lceil \log\left(\frac{2}{\delta}\right) \cdot \frac{1}{\varepsilon} \right\rceil$$

Proof: We assume realizability, hence $\exists h^* \in H^{thr}$ such that

$$L_{D,f}(h^*) = 0.$$

$\Rightarrow$ we let a^* be the corresponding threshold.





Let a_0 be such that $\mathcal{D}\left(\{x \in \mathbb{R} : x \in (a_0, a^*)\}\right) = \varepsilon$

Let a_1 be $-''-$ $\mathcal{D}\left(\{x \in \mathbb{R} : x \in (a^*, a_1)\}\right) = \varepsilon$

In case $\mathcal{D}\left(\{x \in \mathbb{R} : x \in (a_0, a^*)\}\right) < \varepsilon$, set $a_0 = -\infty$

In case $\mathcal{D}\left(\{x \in \mathbb{R} : x \in (a^*, a_1)\}\right) < \varepsilon$, set $a_1 = +\infty$

} special cases

We have $S = \left((x_1, y_1), \dots, (x_m, y_m)\right)$. An ERM algorithm would be, for instance, to first pick b_0 and b_1 as follows:

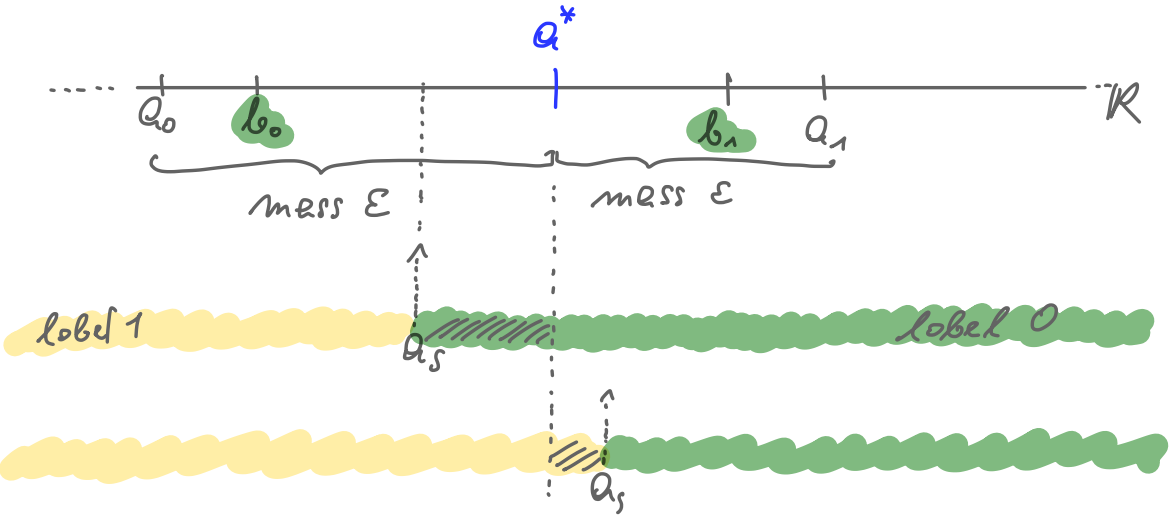
$$b_0 = \max\{x : (x, 1) \in S\}$$
$$b_1 = \min\{x : (x, 0) \in S\}$$

Then, ERM can choose <u>any</u> threshold within (b_0, b_1). We will call the chosen threshold a_S and the corresponding hypothesis h_S.

⟹ By these choices, we get 0 empirical error.

For h_S (with threshold a_S) to have $L_{\mathcal{D},f}(h_S) \leq \varepsilon$, it <u>suffices</u> that

1) $b_1 \leq a_1$ AND
2) $b_0 \geq a_0$



In other words,

$$\mathbb{P}\left[L_{D,f}(h_S) > \varepsilon\right] \leq \mathbb{P}\left[\left(b_0 < a_0\right) \overset{(\vee)}{\text{OR}} \left(b_1 > a_1\right)\right]$$

$$\overset{\text{union bound}}{\leq} \mathbb{P}\left[b_0 < a_0\right] + \mathbb{P}\left[b_1 > a_1\right]$$

Hence, to upper bound $\mathbb{P}\left[L_{D,f}(h_S) > \varepsilon\right]$, we need to check when $b_0 < a_0$ (or $b_1 > a_1$) actually happens. Answer: when there is no data point x in S that is labeled 1 s.t. $x \in (a_0, a^*)$.

We know that $\mathcal{D}(\{x \in \mathbb{R} : x \in (a_0, a^*)\}) = \varepsilon$, hence not seeing a data point in (a_0, a^*) has probability of $1-\varepsilon$. Consequently, not seeing a data point in m i.i.d samples from $\mathcal{D}$ has probability of $(1-\varepsilon)^m$.

Remark: $(1-\varepsilon)^m \leq e^{-\varepsilon m}$

Overall, $\mathbb{P}_{S\sim D^m}\left[h_0 < a_0\right] = (1-\varepsilon)^m \leq e^{-\varepsilon m}$

$\mathbb{P}_{S\sim D^m}\left[h_1 > a_1\right] = (1-\varepsilon)^m \leq e^{-\varepsilon m}$

Combined, we get

$$\mathbb{P}\left[L_{D,f}(h_S) > \varepsilon\right] \leq \underbrace{2 \cdot e^{-\varepsilon m}}_{\text{let this be} < \delta \text{ and solve for } m}$$

↑ ERM hyp.

$$m > \log\left(\frac{2}{\delta}\right)\cdot\frac{1}{\varepsilon}$$

We have seen that "finiteness" of H ($|H|<\infty$) is sufficient, but not necessary (for PAC learnability).

Def. (VC-Dimension): The VC dimension of H (i.e., a class of functions from $X \to \{0,1\}$), written as

$$VC(H),$$

is the maximal size of a set $C \subset X$ that is shattered by H.

Theorem: Let H be a class of functions from $X \to \{0,1\}$. If H has infinite VC dimension, then H is **not** PAC learnable.

(follows immediately from NFL)

Def. (growth function):

Let H be a class of functions from $X \to \{0,1\}$. The growth function of H, $\tau_H : \mathbb{N}_0 \to \mathbb{N}_0$, is defined as

$$\tau_H(m) = \max_{C \subset X, |C|=m} |H_C| \qquad \left(\text{with } \tau_H(0) \overset{\text{def.}}{=} 1\right)$$

Lemma (Sauer, Shelah, Perez; "Sauer's lemma"): Let H be a hyp. class of functions from $X \to \{0,1\}$ with $VC(H) = d$.

Then

$$\tau_H(m) = 2^m \text{ if } m \leq d, \text{ but}$$
$$\tau_H(m) = \left(\frac{em}{d}\right)^d \text{ if } m > d$$

(without proof)

06/11/24

VC-Dimension of finite $\mathcal{H}$:

It's obvious that, if we take any set C (of size m),

$$|\mathcal{H}_C| \leq |\mathcal{H}|$$

So, if $|\mathcal{H}| < 2^m$, then $\mathcal{H}$ cannot shatter C (of size m).

This implies $$VC(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$$

Theorem: Let $\mathcal{H}$ be a hyp. class of functions from $\mathcal{X} \to \{0,1\}$ and $\ell: \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0,c]$, $c>0$ be a loss function. For any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and $\delta \in (0,1)$, we have with probability of at least $1-\delta$ over the choice of $S \sim \mathcal{D}^m$

$$\forall h \in \mathcal{H}: \left| L_{\mathcal{D}}(h) - L_S(h) \right| \leq c \cdot \sqrt{\frac{8 \cdot \log\left(\tau_{\mathcal{H}}(2m) \cdot \frac{4}{\delta}\right)}{m}}$$

$\tau_{\mathcal{H}}(2m)$: growth function, evaluated at $2m$.

(without proof)

Fundamental theorem of statistical learning

Let H be a hyp. class of functions from $X \to \{0, 1\}$, and let l be the 0-1 loss. Then, the following statements are equivalent:

1. H has the uniform convergence property.
2. Any ERM algorithm is a successful agnostic PAC learner for H.
3. H is agnostic PAC learnable
4. Any ERM algorithm is a successful PAC learner for H.
5. H is PAC learnable
6. H has finite VC dimension ($VC(H) < \infty$)

Example: Hyp. class of LINEAR PREDICTORS

We define

$$\{h_{w,b} : w \in \mathbb{R}^d, b \in \mathbb{R}\} =: L_d \text{, with}$$

$$h_{w,b}(x) = \langle w, x \rangle + b \quad ; x \in \mathbb{R}^d$$

as the class of **affine functions**. From L_d, we get different predictors via composition with some

$$\phi : \mathbb{R} \to Y$$

(We can also use

$$w' = (b, w_1, \dots, w_d)^T \in \mathbb{R}^{d+1}$$
$$x' = (1, x_1, \dots, x_d)^T \in \mathbb{R}^{d+1}$$

to write

$$h_{w,b}(x) = \langle w', x' \rangle$$

)

If $\phi : \mathbb{R} \to Y$ is $\phi \equiv \text{sign}$, then we get **halfspace predictors**

$$HS_d = \phi \circ L_d = \{x \mapsto \text{sign}(h_{w,b}(x)) : h_{w,b} \in L_d\}$$

Question: What is the VC dimension of

$$\mathcal{F} = \{x \mapsto \operatorname{sign}(\langle w, x\rangle) : w \in \mathbb{R}^d\},$$

that is $VC(\mathcal{F}) = ?$

Claim: $VC(\mathcal{F}) = d$

Part 1: Show that $VC(\mathcal{F}) \geq d$

$$X = \{c_1, \dots, c_d\} \text{ with } c_i \in \mathbb{R}^d \text{ for all } i \in [d]$$

with

$$c_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad c_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \dots, \quad c_d = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

and

$$w = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{pmatrix} \begin{matrix} \leftarrow \text{label of } c_1 \ (\in \{\pm 1\}) \\ \\ \\ \leftarrow \text{label of } c_d \end{matrix}$$

Remark: c_{ij} refers to the j-th coordinate of c_i

$$\Rightarrow \langle w, c_i \rangle = \sum_{j=1}^{d} w_j \cdot c_{ij} = 1 \cdot w_i = y_i \Rightarrow VC(\mathcal{F}) \geq d$$

Part 2 We need to show $VC(\mathcal{F}) < d+1$. We will show this via contradiction.

↳ no set of size $d+1$ is shattered by $\mathcal{F}$.

Assume that $d+1$ points

$$X = \{c_1, \dots c_{d+1}\}$$

are shattered by $\mathcal{F}$.

This means that $\exists w_1, \dots w_{2^{d+1}}$ weight vectors that will yield all (2^{d+1}) different labelings of our $d+1$ points.

Lets write down all possible inner products:

$$M = \begin{pmatrix} w_1^T c_1 & w_2^T c_1 & \cdots & w_{2^{d+1}}^T c_1 \\ w_1^T c_2 & & & \\ w_1^T c_3 & & & \\ \vdots & & & \\ w_1^T c_{d+1} & w_2^T c_{d+1} & \cdots & w_{2^{d+1}}^T c_{d+1} \end{pmatrix}$$

$d+1$ rows

2^{d+1} columns

We can also write this matrix as

$$XW = H$$

$d+1$ rows
d columns

$$X = \begin{pmatrix} — & c_1 & — \\ & \vdots & \\ — & c_{d+1} & — \end{pmatrix} \qquad W = \begin{pmatrix} | & | & & | \\ w_1 & w_2 & \cdots & w_{2^{d+1}} \\ | & | & & | \end{pmatrix}$$

d rows
2^{d+1} columns

If we would take $\text{sign}(XW)$ we would get all possible labelings. We know that

$$\boxed{\text{rank}(H) \leq \min\{\text{rank}(X), \text{rank}(W)\} = d}$$

06/18/24

↳

$$H = \begin{pmatrix} w_1^T c_1 & \cdots & w_{2^{d+1}}^T c_1 \\ w_1^T c_2 & & \vdots \\ \vdots & & \\ w_1^T c_{d+1} & & w_{2^{d+1}}^T c_{d+1} \end{pmatrix}$$

$$\downarrow$$

$$X w_1 \in \mathbb{R}^{d+1}$$

Remark: Any $v_1, \dots v_n$ (set of vectors) are lin. independent iff

$$a_1 v_1 + a_2 v_2 + \dots + a_n v_n = 0$$

only when $\forall i: a_i = 0$

In our case (looking at the rows of H):

$$a_1 \cdot [w_1^T c_1 \quad \cdots \quad w_{2^{d+1}}^T c_1] +$$
$$a_2 \cdot [w_1^T c_2 \quad \cdots \quad w_{2^{d+1}}^T c_2] +$$
$$\vdots$$
$$a_{d+1} \cdot [w_1^T c_{d+1} \quad \cdots \quad w_{2^{d+1}}^T c_{d+1}] =$$
$$[a^T X w_1 \quad a^T X w_2 \quad \cdots \quad a^T X w_{2^{d+1}}]$$

by our assumption

As $c_1 \dots c_{d+1}$ are shattered by F, there $\exists k \in \{1, \dots, 2^{d+1}\}$ such that

$$\text{sign}(a) = \text{sign}(X w_k)$$

This means that

$$a^T X w_k$$

is a sum of positive numbers $\rightarrow \neq 0$!

$\Rightarrow$ $d+1$ rows of H are lin. independent $\Rightarrow$ $\operatorname{rank}(H) = d+1$

$\Rightarrow$ Contradiction (with $\operatorname{rank}(H) \leq d$).

We conclude that $VC(\mathcal{F}) = d$. (as $VC(\mathcal{F}) \geq d$ and $VC(\mathcal{F}) < d+1$)

RADEMACHER complexity

Unif. conv. (ε-rep.)

$$\forall h \in H: \quad |L_D(h) - L_S(h)| \le \varepsilon$$

We can also say

$$\sup_{h \in H} |L_D(h) - L_S(h)| \le \varepsilon$$

$Z = X \times Y$, class H

loss $\ell: H \times Z \to \mathbb{R}_+$

Def: $\mathcal{F} = \ell \circ H = \{ \underbrace{z \mapsto \ell(h, z)}_{(x,y)} : h \in H \}$

and for $f \in \mathcal{F}$: $\boxed{L_D(f) = \mathbb{E}_z[f(z)]}$ $\quad L_S(f) = \frac{1}{m} \sum_{i=1}^{m} f(z_i)$

Def: $\mathrm{Rep}_D(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} \left(L_D(f) - L_S(h) \right)$ (*)

Take $S = S_1 \cup S_2$ with $|S_1| = |S_2|$ and estimate (*) by

$$\sup_{f \in \mathcal{F}} \left(L_{S_1}(f) - L_{S_2}(f) \right)$$

$$+1 \cdot \frac{2}{m} \cdot \sum_{z \in S_1} f(z) + (-1) \cdot \frac{2}{m} \cdot \sum_{z \in S_2} f(z)$$

$$= \frac{2}{m} \cdot \sum_{z \in S_1} (+1) \cdot f(z) + \frac{2}{m} \cdot \sum_{z \in S_2} (-1) f(z)$$

collect into a vector $\sigma \in \{\pm 1\}^m$

Set $S_1 = \{z_i : \sigma_i = +1\}$, $S_2 = \{z_i : \sigma_i = -1\}$. We rewrite

$$\sup_{f \in \mathcal{F}} \frac{2}{m} \cdot \sum_{i=1}^{m} \sigma_i f(z_i)$$

Def. $\mathcal{F} \circ S = \{(f(z_1), \dots, f(z_m)) : f \in \mathcal{F}\}$

$$RAD(\mathcal{F} \circ S) = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i f(z_i) \right] \qquad \langle \vec{\sigma}, (f(z_1) \dots f(z_m)) \rangle$$

$\sigma \sim \text{Uniform}(\{\pm 1\})^m$

RADEMACHER COMPLEXITY

$$\mathbb{P}[\sigma_i = +1] = \mathbb{P}[\sigma_i = -1] = \frac{1}{2}$$

Lemma $$\mathop{\mathbb{E}}_{S \sim D^m} \left[Rep_D(\mathcal{F}, S) \right] \leq 2 \cdot \mathop{\mathbb{E}}_{S \sim D^m} \left[RAD(\mathcal{F} \circ S) \right]$$

$Rep_D(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} (L_D(f) - L_S(f))$

Theorem (informal)

Assume for all $z \in Z$ and $h \in H$: $|\ell(h,z)| \leq c$.

Then

(1) with prob. of $1-\delta$ over the choice of S, for all $h \in H$:

$$L_D(h) - L_S(h) \leq 2 \cdot \mathop{\mathbb{E}}_{S \sim D^m} \left[RAD(\ell \circ H \circ S) \right] + c \cdot \sqrt{\frac{2 \log(2/\delta)}{m}}$$

(2) —"— $\leq 2 \cdot RAD(\ell \circ H \circ S) + 4c \cdot \sqrt{\frac{2 \log(4/\delta)}{m}}$