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

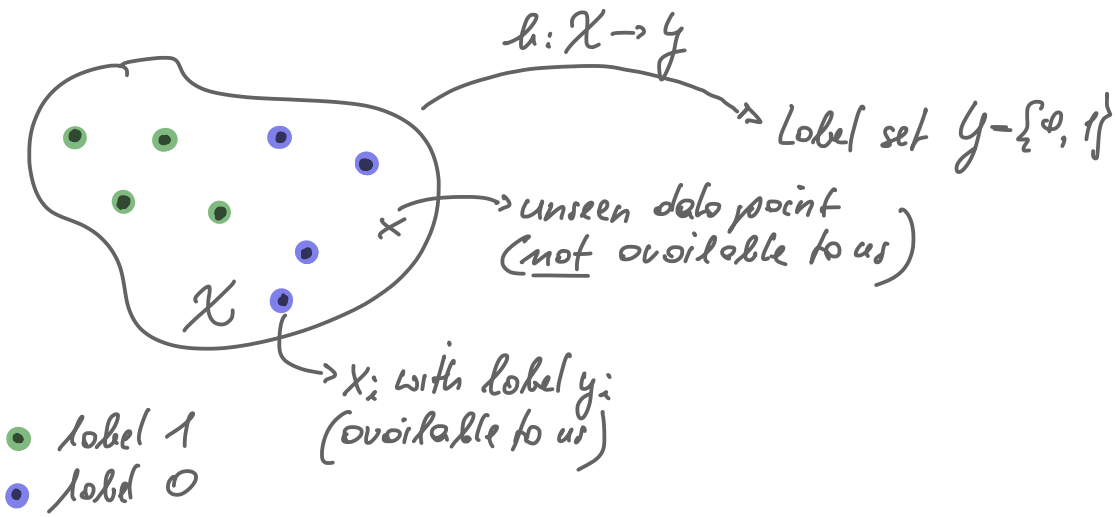
$$((x_1, y_1), \ldots\ldots (x_N, y_N)) = S$$

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

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

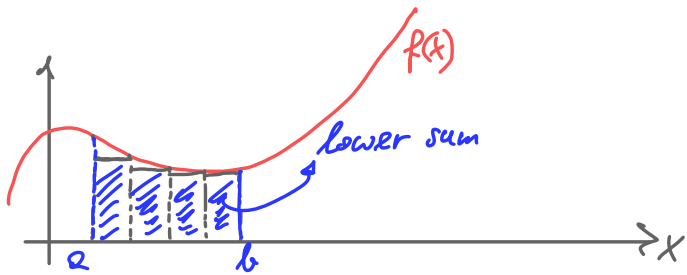
we call $\mathbb{P}_X$ the <u>distribution of X.</u>

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

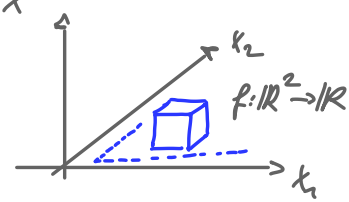
$$\{X \in A\} = \{\omega \in S : X(\omega) \in A\}$$
$$\{X = c\} = \{\omega \in S : X(\omega) = c\}$$
$$\mathbb{P}_X(A) = \mathbb{D}\big(\{\omega \in S : X(\omega) \in A\}\big)$$
$$\mathbb{P}_X(\{c\}) = \mathbb{D}\big(\{\omega \in S : X(\omega) = c\}\big)$$

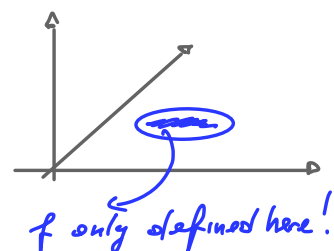
A very short primer on the LEBESGUE INTEGRAL

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



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

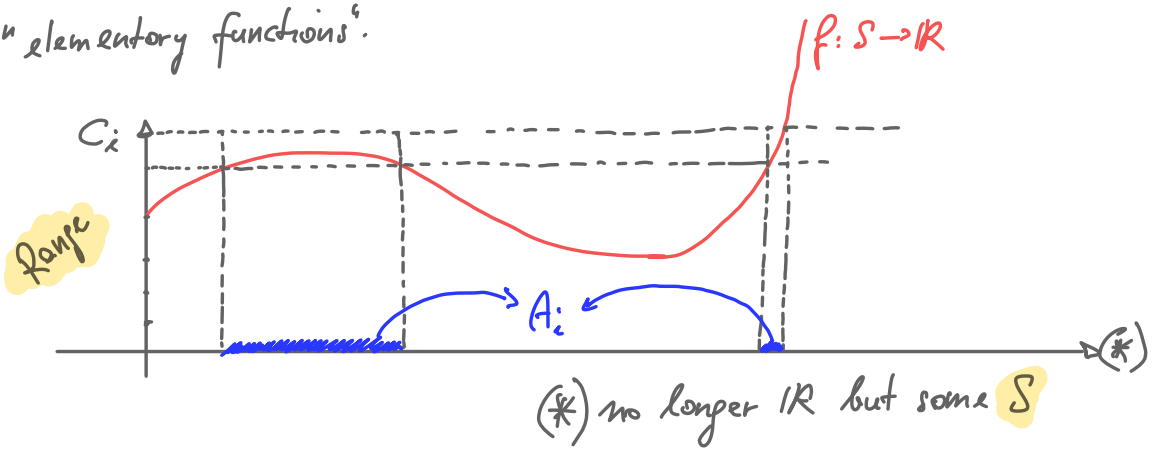




Question: How to partition this domain?

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

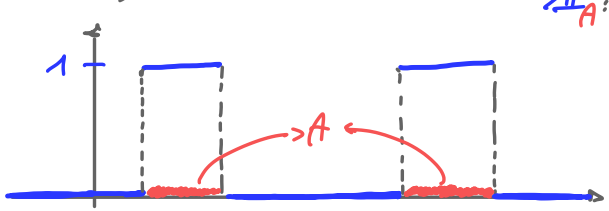
..... this allows us to eventually write

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

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

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

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



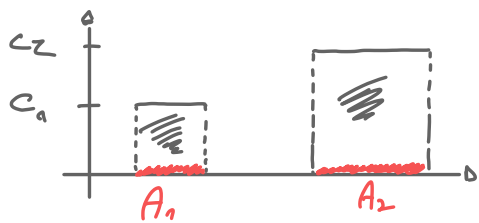
Any reasonable notion of integration of $\mathbb{1}_A$ should return $\mu(A)$:

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

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)

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

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

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

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

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

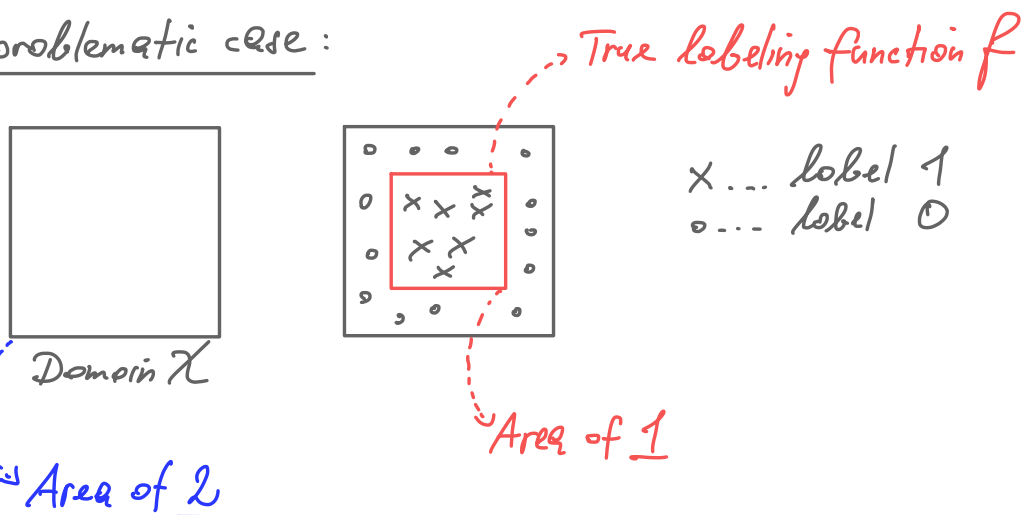
$$= L_{D,f}(h) \quad \text{generalization error}$$

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 function $f: X \to Y$ and any distribution D over X (for which realizability holds), we have that with probability of at least $1-\delta$ over the choice of $S|_x$ (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</u> (PAC).

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.