MACHINE LEARNING (911.235)

Summer term 2023

03/07/23

Motivating example

	weight (g)	Color ($\in [0,1]$)	Tasty (0) / Non-Tasty (1)
Papaya 1	100	0.1	yes (1)
—"— 2	⋮	⋮	⋮
⋮			
Papaya N	500	0.8	no (0)
	$\in \mathbb{R}$	$\in \mathbb{R}$	

TRAINING DATA

Our goal would be to find

$$h : \mathbb{R}^2 \to \{0, 1\}$$

that will take, as input, a 2-dimensional vector and output whether a papaya is tasty or not. We call h a *hypothesis*.

What data is available to us?

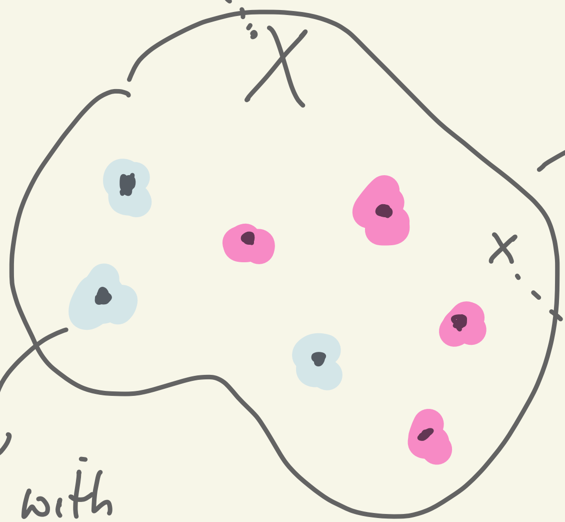
$$S = \left((x_1, y_1), \ldots\ldots, (x_N, y_N) \right),$$ in our case $x_i \in \mathbb{R}^2$, $y_i \in \{0, 1\}$

What do we mean by learning?

Domain

$h: X \to Y$

Label set $Y = \{0, 1\}$



Unseen data point

x_i with label y_i

A **learner** receives S and outputs h.
(some algorithm)

Two assumptions that we will make initially is as follows:

- All the x_i's are drawn independently and identically distributed (iid) from some (unknown) distribution $(\mathcal{D})$ over the domain $\mathcal{X}$.

- The x_i's are labeled by some (unknown) function $f: \mathcal{X} \to \mathcal{Y}$ (true labeling function). This means

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

What do we care about?

$$\underbrace{A = \{x \in X : h(x) \neq f(x)\}}_{\text{all the points where the hypothesis } h \text{ differs from the true labeling function.}} \qquad \text{subset of } X$$

MACHINE LEARNING (911.235)

Summer term 2023

PRELIMINARIES

08/14/23

$\mathcal{P}(X)$ ... power set of X

Remember that we care about $A = \{x \in \overset{\text{domain}}{X} : h(x) \neq f(x)\}$. If $X = \mathbb{R}^n$ and $A \in \mathcal{P}(X)$, we already have a problem.

<u>Measure problem</u>: find $\mu : \mathcal{P}(\mathbb{R}^n) \to [0, \infty]$ with

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 <u>congruent</u>, we want

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

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

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

We will constrain ourselves to sets that are an element 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 $A_i \in \mathcal{F}$, $i \in \mathbb{N}$, it follows that

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$$

Example (smallest σ-Algebra that contains set A)

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

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

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

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

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

we call $\mathcal{E}$ the <u>generator</u> 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 with

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

$$\{U_i\}_{i=1}^{n} \subseteq \tau \Longrightarrow \bigcap_{i=1}^{n} U_i \in \tau$$

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-Algebra): given a topological space $(S, \mathcal{O})$ with $\mathcal{O}$ denoting the system of open sets, then we call

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

the Borel σ-Algebra over S. It's elements are called Borel sets. For $S = \mathbb{R}^n$, we write $\mathcal{B}^n := \mathcal{B}(\mathbb{R}^n)$.

Each of of the following systems of sets are generators for $\mathcal{B}(\mathbb{R}^n)$:

- $\{U \subset \mathbb{R}^n : U \text{ open}\}$
- $\{A \subset \mathbb{R}^n : A \text{ closed}\}$
- $\{]a,b] : a, b \in \mathbb{R}^n \text{ with } a \leq b\}$
- $\{]-\infty, c] : c \in \mathbb{R}^n\}$

$]a,b] =]a_1,b_1] \times \dots \times]a_n,b_n]$ for $a = (a_1 \dots a_n)$, $b = (b_1 \dots b_n)$

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

$$\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$$

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

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

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

a measurable space. Example: $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$

Def. (measurable maps/functions): Given $(S_1, \mathcal{F}_1)$ and $(S_2, \mathcal{F}_2)$ measurable spaces, we call

$$f: S_1 \to S_2$$

a measurable function if $\forall E \in \mathcal{F}_2 : f^{-1}(E) \in \mathcal{F}_1$

Example: $\mathbb{1}_A : S \to \{0,1\}$ indicator function of set $A \subset S$

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

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

- if $\beta \le 0$, then $\{\omega : \mathbb{1}_A(\omega) < \beta\} = \emptyset$
- if $\beta > 1$, then $\{\omega : \mathbb{1}_A(\omega) < \beta\} = S$
- if $0 < \beta \le 1$, then $\{\omega : \mathbb{1}_A(\omega) < \beta\} = S \setminus A$

$\Rightarrow \mathbb{1}_A$ is measurable if $A \in \mathcal{F}$ $\qquad (S, \mathcal{F})$

Def. (Measure): Let $(S, \mathcal{F})$ be a measurable space. A function

$$\mu : \mathcal{F} \to \overline{\mathbb{R}}$$

is called a <u>measure</u>, 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

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

Example: $\mu_{count} : \mathcal{F} \to \overline{\mathbb{R}}$

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

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

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

a <u>measure space</u>.

Some important properties: let $A, B, A_n \in \mathcal{F}, n \in \mathbb{N}$. Then

1. if A and B are disjoint, then $\mu(A \cup B) = \mu(A) + \mu(B)$
2. if $A \subset B$, then $\mu(A) \leq \mu(B)$
3. $$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu(A_i) \quad // \ \sigma\text{-sub-additivity}$$

__Def. (Probability space):__ given $(S, \mathcal{F}, D)$ a measure space, with

$$D(S) = 1,$$

then we call $(S, \mathcal{F}, D)$ a __probability space__.

__Remark:__ A possibility to construct, based on a measure space $(S_1, \mathcal{F}_1, \mu)$ and a measurable function f from S_1 to S_2 (with $(S_2, \mathcal{F}_2)$), another measure space $(S_2, \mathcal{F}_2, \mu_f)$:

$$\begin{aligned} \mu_f : \mathcal{F}_2 &\longrightarrow [0, \infty] \\ B &\longmapsto \mu_f(B) := \mu\left(f^{-1}(B)\right) \end{aligned}$$

// push-forward measure

Def (Random variable): if $(S_1, \mathcal{F}_1, \mathbb{P})$ is a probability space and $(S_2, \mathcal{F}_2)$ a measurable space, then a measurable function

$$X : S_1 \longrightarrow S_2$$

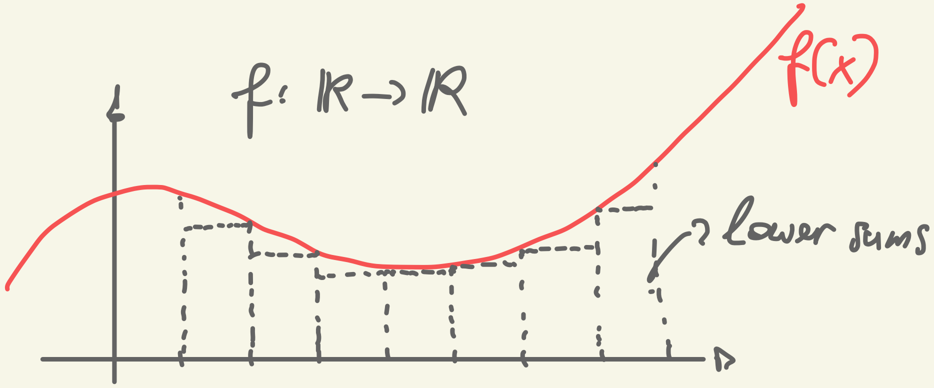
is called a random variable. For $S_2 = \mathbb{R}$, we call X a real random variable.

The push-forward measure $\mathbb{P}_X$ on $(S_2, \mathcal{F}_2)$ is also a prob. measure.

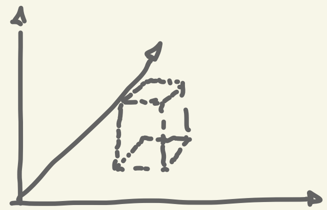
$$\underbrace{\mathbb{P}_X(S_2)}_{\text{"distribution" of } X} = \mathbb{P}(X^{-1}(S_2)) = \mathbb{P}(S_1) = 1$$

Idea of Riemann integral

$f: \mathbb{R} \to \mathbb{R}$



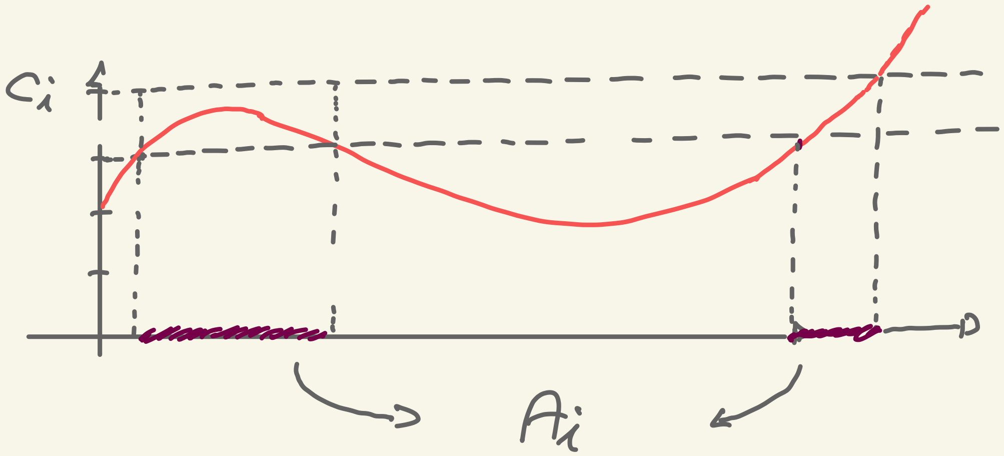
Problems:
- Diff. to extend to higher dimensions
- Reliance on continuity
- ⋮



say f is defined here

Q: How do partition the domain?

Idea of LEBESGUE integration: The key idea is to partition the range of a function $f: S \to \mathbb{R}$ in intervals, to get to an approximation by "elementary" functions.



no longer $\mathbb{R}$, but some domain S

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

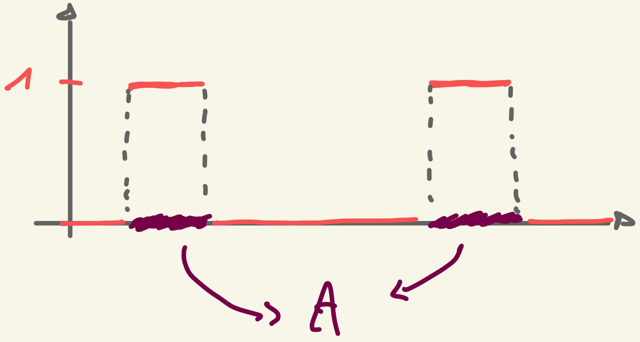
$$c_i \cdot \mu(A_i)$$

measure

.... this will allow us to eventually write

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

In a little bit more detail: In particular, lets look at case of "simple" functions. Say, we have $(S, \mathcal{F}, \mu)$, $\mu : \mathcal{F} \to [0, \infty]$

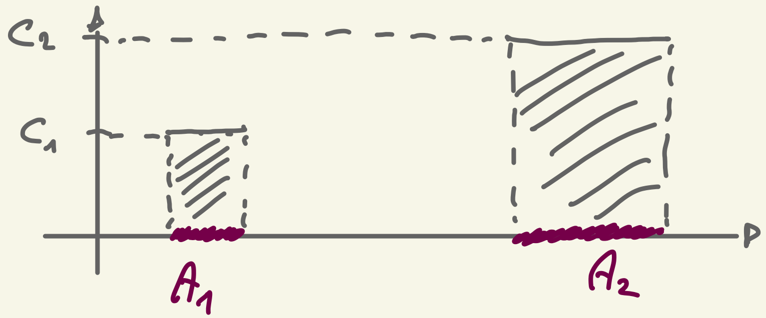


$\mathbb{1}_A : S \to \mathbb{R}$, $A \in \mathcal{F}$
(indicator function of set A)

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

$$\underbrace{I}_{\text{integral}}(\mathbb{1}_A) = \mu(A)$$

What is a "simple" function? $f(x) = \sum_{i=1}^{n} c_i \cdot \mathbb{1}_{A_i}(x)$ $\quad A_i \in \mathcal{F}$, $c_i \in \mathbb{R}$



We can define the integral of f via

$$I(f) = \sum_{i=}^{n} c_i \cdot \mu(A_i)$$

Problem: $\mu: \mathcal{F} \to [0,\infty]$, $c_i \in \mathbb{R} \Rightarrow$ so, we could get something like

$$10 \cdot \infty - 3 \cdot \infty$$

Solution: $\{ f: S \to \mathbb{R} \mid f$ is "simple" and $f \geq 0 \} = \mathcal{T}^+$

For $f \in \mathcal{T}^+$, we have a 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 as

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

Properties: 1. For $f, g \in T^+$ and $\alpha, \beta \geq 0$

$$\int (\alpha f + \beta g)\, d\mu = \alpha \cdot \int f\, d\mu + \beta \int f\, d\mu \qquad \text{LINEARITY}$$

2. If $f, g \in T^+$ and $f \leq g$, it follows that

$$\int f\, d\mu \leq \int g\, d\mu \qquad \text{MONOTONICITY}$$

! This construction extends to all measurable functions $f : S \to \mathbb{R}$.

Def (Expected value of a random variable): given a probability space $(S, \mathcal{F}, P)$ and a (quasi-integrable) real random variable $X : S \to \mathbb{R}$, then

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

is the expected value of X.

Properties (some):

1. if $X \geq 0$, then $\mathbb{E}[X] \geq 0$

2. $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ for random variables X and Y

$\vdots$

Our first inequality: MARKOV inequality

Def (Markov inequality): given a prob. space $(S, \mathcal{F}, \mathbb{P})$ and a non-negative (≥ 0) random variable, we have for $a > 0$

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a} \qquad \left(\mathbb{P}[X \geq a] = \mathbb{P}(\{\omega \in S : X(\omega) \geq a\}) \right)$$

Proof: We define $\varphi : S \to \mathbb{R}$ with

$$\varphi(x) = \begin{cases} a, & \text{if } X \geq a \\ 0, & \text{if } X < a \end{cases}$$

$\Rightarrow 0 \leq \varphi(x) \leq X(x)$. Hence, by <u>monotonicity</u>, we have

$$\int X \, dD \geq \underbrace{\int \varphi \, dD}_{}$$

$$= a \cdot D(\{\omega \in S : X(\omega) \geq a\})$$

Divide by $a > 0$

$$\Rightarrow \frac{1}{a} \cdot \int X \, dD \geq D(\{\omega \in S : X(\omega) \geq a\})$$

$$= \mathbb{P}[X \geq a]$$

Since $\int X \, dD$ is $\mathbb{E}[X]$, we get $\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$ $\blacksquare$

Lets start with actual ML content

(ended here on 08/20)

Recap of our setup:

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 set $S = ((x_1, y_1), \ldots, (x_m, y_m))$ with $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$
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 $\mathcal{D}$ over the domain and labeled by some function $f: \mathcal{X} \to \mathcal{Y}$:

$$x_i \sim \mathcal{D}, \quad y_i = f(x_i)$$

We are interested in

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

The <u>empirical version</u> of this is

$$\frac{1}{m}\left|\{i \in [m] : h(x_i) \neq f(x_i)\}\right| = \underbrace{L_S(h)}_{\text{Empirical error (or empirical risk)}}$$

$$\left([m] = \{1, \ldots, m\}\right)$$

<u>Convention:</u> $S|_X = (x_1, \ldots, x_m)$

Claim: $\mathbb{E}_{S|_X \sim \mathcal{D}^m}\left[L_S(h)\right] = L_{\mathcal{D},f}(h)$

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

$$\mathbb{1}_{h(x_i) \neq f(x_i)} = \begin{cases} 1, & \text{if } h(x_i) \neq f(x_i) \\ 0, & \text{else} \end{cases}$$

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

$$= \frac{1}{m} \cdot \sum_{i=1}^{m} \mathop{\mathbb{E}}_{x \sim D}\left[\mathbb{1}_{h(x) \neq f(x)}\right] \qquad // \text{ as the } x_i\text{'s are iid}$$

$$= \frac{1}{m} \sum_{i=1}^{m} \mathop{\mathbb{P}}_{x \sim D}\left[h(x) \neq f(x)\right] \qquad // \text{ as } \mathbb{1}_{\dots} \text{ only take on values } 0, 1$$

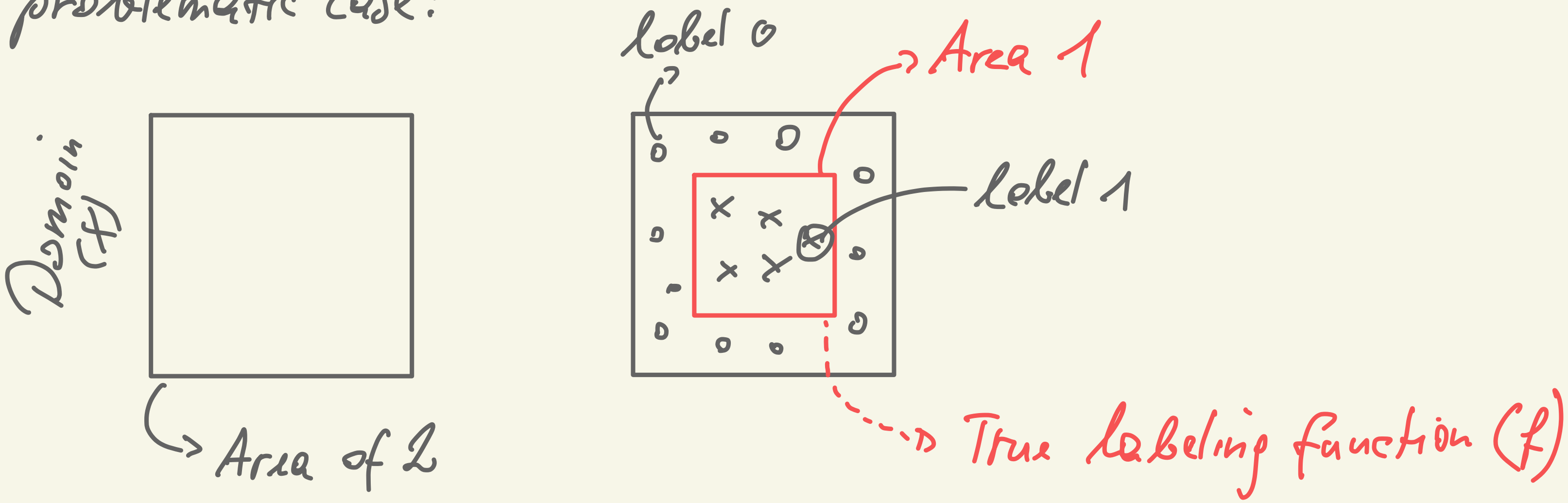
$$= \frac{1}{\cancel{m}} \cdot \cancel{m} \cdot \mathop{\mathbb{P}}_{x \sim D}\left[h(x) \neq f(x)\right]$$

$$= L_{D,f}(h) \qquad // \text{ this establishes the claim.}$$

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 is minimized. We call such an h an empirical risk minimizer (h_S).

A problematic case:



- Say the distribution on $\mathcal{X}$ is uniform

- Say we have an ERM algorithm that returns h_S s.t.

$$h_S(x) = \begin{cases} y_i, & \text{if } \exists i \in [m] : x_i = x \\ 0, & \text{else} \end{cases} \qquad // \text{lookup table}$$

Obviously h_S is correct on our training set $\Rightarrow L_S(h_S) = 0$!

But on unseen instances from $\mathcal{D}$ $(x \sim \mathcal{D})$, h_S is only correct 50% of the time (due to the ratio of areas □ and ▫) $\Rightarrow L_{\mathcal{D},f}(h_S) = \frac{1}{2}$!

This is called overfitting!

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

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

<u>Remark:</u> In our previous example, we did not do this and allowed to memorize the training data!

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

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

Now, any ERM hypothesis h_S will attain 0 empirical error ($L_S(h_S)=0$), as it 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 can write

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

Also, we define

$$M = \{S|_x : \exists h \in \underbrace{H_{BAD}}_{\text{those are the ones with generalization error } \geq \varepsilon},\ L_S(h) = 0\}$$

We observe

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

Emp. risk minimizer

We get (upon measuring with D):

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

$$\leq \sum_{h \in H_{BAD}} D^m\left(\{S|_x : L_S(h) = 0\}\right) \qquad \text{// due to } \sigma\text{-sub-additivity}$$

"Union bound"

<u>Let's fix some $h \in H_{BAD}$:</u>

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

$$= \prod_{i=1}^{m} D\left(\{x_i : h(x_i) = f(x_i)\}\right) \qquad \text{// due to iid assumption}$$

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

$$\leq \prod_{i=1}^{m} (1 - \varepsilon) \qquad \text{// as } h \in H_{BAD}$$

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

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

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

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

If we let this be $< \delta \in (0,1)$ and solve for $\underline{m}$, we get

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

error   confidence

Corollary: 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 domain X (for which realizability holds), we have that with probability of at least $1-\delta$ over the choice of $S|_X$ (of size m) it holds that every ERM hypothesis h_S satisfies

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

Interpretation: For sufficiently large m, ERM_H returns h_S (i.e., a hypothesis) that is PROBABLY APPROXIMATELY CORRECT (PAC).

This leads to:

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

We have already seen that finite hypothesis classes ($|H| < \infty$) are PAC learnable with

$$m_H(\varepsilon, \delta) \leq \left\lceil \frac{1}{\varepsilon} \cdot \log\left(\frac{|H|}{\delta}\right) \right\rceil \to \text{ceiling function}$$

We will now move to a more general setting.

$\boxed{1}$ We will first release the realizability assumption.

(In this setting, the best we can hope for are guarantees relative to the "best" possible hypothesis in the class; $\min_{h \in \mathcal{H}} L_{D,f}(h)$)

<u>Def. (Hoeffding inequality)</u>: Let $X_1, \ldots, X_m$ be iid random variables taking values in $[a_i, b_i]$ for $i \in [m]$. Then, it holds that

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

and

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

Also

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

(as we have $|a| \geq b \iff a \leq -b$ <u>OR</u> $a \geq b$)

Another useful form of this inequality is:

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

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 any single $h : X \to Y$, we have

$$\mathbb{P}_{S|_X \sim D^m}\left[\left|L_S(h) - L_{D,f}(h)\right| > \varepsilon\right] \leq 2e^{-2\varepsilon^2 m} \qquad (a=0,\ b=1)$$

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

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

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

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

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

(by the union bound)

We see that we did need realizability for that!

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

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

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

Both [1] and [2] lead to:

Def. (Agnostic PAC learnability): A hyp. 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 over $X \times Y$, when running A on $m \geq m_H(\varepsilon, \delta)$ iid instances from D, A returns a hypothesis h such that with prob. of at least $1-\delta$ (over the choice of S)

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

(last slide on April 18)