MACHINE LEARNING (911.235)

Summer term 2023

03/07/23

Motivating example

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

Our goal would be to find

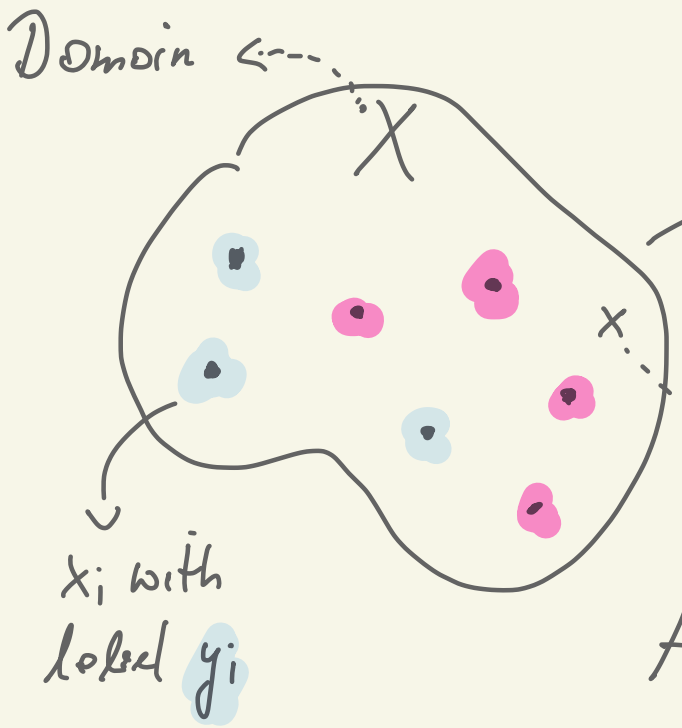
$$h : \mathbb{R}^2 \to \{0, 1\}$$

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?

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

What do we mean by learning?



$h: X \to Y$

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

Unseen data point

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

Two assumptions that we will make initially is as follows:

(iid)

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

- The x_i's are labeled by some (unknown) function $f: X \to Y$ (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?

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

all the points where the hypothesis h differs from the true labeling function.

MACHINE LEARNING (911.235)

Summer term 2023

PRELIMINARIES

08/14/23

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

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

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

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 $\mu([0,1]^n) = 1$

! The measure problem is unsolvable 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)

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

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

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

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 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 with

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

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

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

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

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

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

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

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

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

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

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

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

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

is called a **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

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

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 **measure space**.

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. $$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu(A_i)$$ // σ-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 $(S_2, \mathcal{F}_2, \mu_f)$:

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

$$B \mapsto \mu_f(B) := \mu\left(f^{-1}(B)\right)$$

// 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 measureable space, then a measurable function

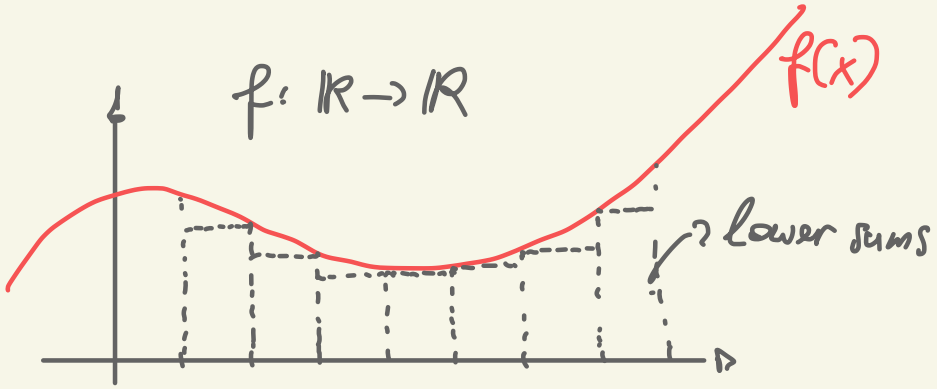
$$X : S_1 \longrightarrow S_2$$

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

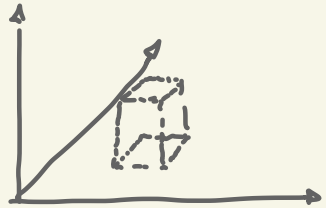
The push-forward measure $\mathbb{P}_X$ on $(S_2, \mathcal{F}_2)$ is also a prob. measure.

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

Idea of Riemann integral



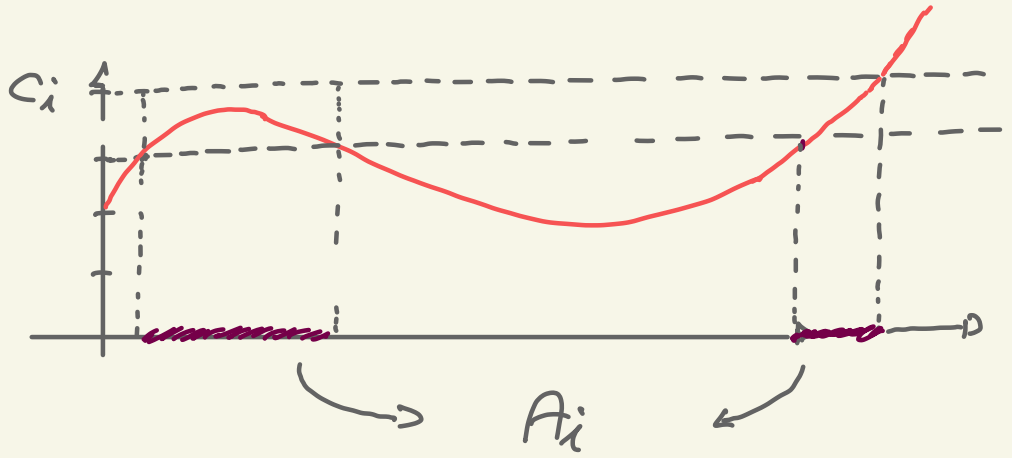
Problems:
- Diff. to extend to higher dimensions
- Reliance on continuity
- ⋮



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

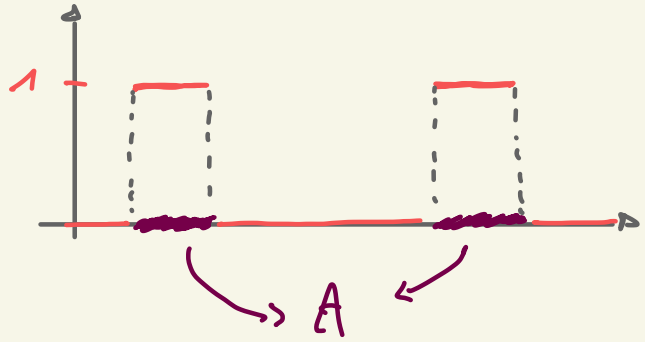
$$c_i \cdot \mu(A_i)$$

μ ⇢ measure

.... this will allow us to eventually write

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

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

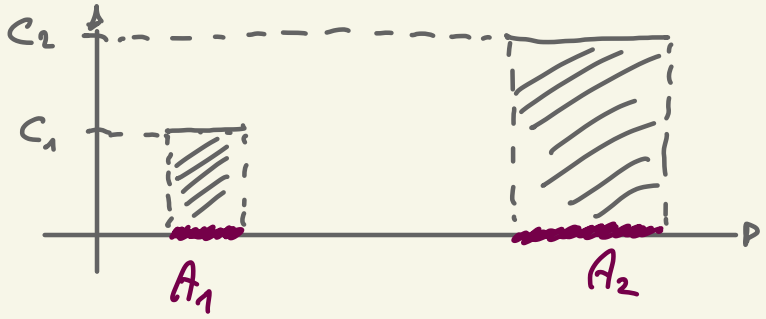


$\mathbb{1}_A : S \to \mathbb{R}$, $A \in \mathcal{F}$
(indicator function of set A)

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? $f(x) = \sum_{i=1}^{n} c_i \cdot \mathbb{1}_{A_i}(x)$ $A_i \in \mathcal{F}$, $c_i \in \mathbb{R}$



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

$$10.\infty - 3.\infty$$

Solution: $\{f: S \to \mathbb{R} \mid f \text{ is 'simple' and } f \geq 0\} = T^+$

For $f \in T^+$, 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

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

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, F, D) and a (quasi-integrable) real random variable $X: S \to \mathbb{R}$, then

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

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

⋮

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 $a > 0$

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$$

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

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 monotonicity, 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 $\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$ ∎

Lets start with actual ML content

(ended here on 03/20)

Recap of our setup:

1. DOMAIN set X; we call $x \in X$ an instance
2. LABEL set Y, e.g. $Y = \{0, 1\}$
3. TRAINING set $S = ((x_1, y_1), \ldots, (x_m, y_m))$ with $x_i \in X$, $y_i \in Y$
4. A LEARNER that receives S and outputs

$$h : X \to Y$$

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 and labeled by some function $f : X \to Y$: $x_i \sim D$, $y_i = f(x_i)$

We are interested in

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

"Generalization error"

The empirical version of this is

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

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

Empirical error (or empirical risk)

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

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

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

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

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

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

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

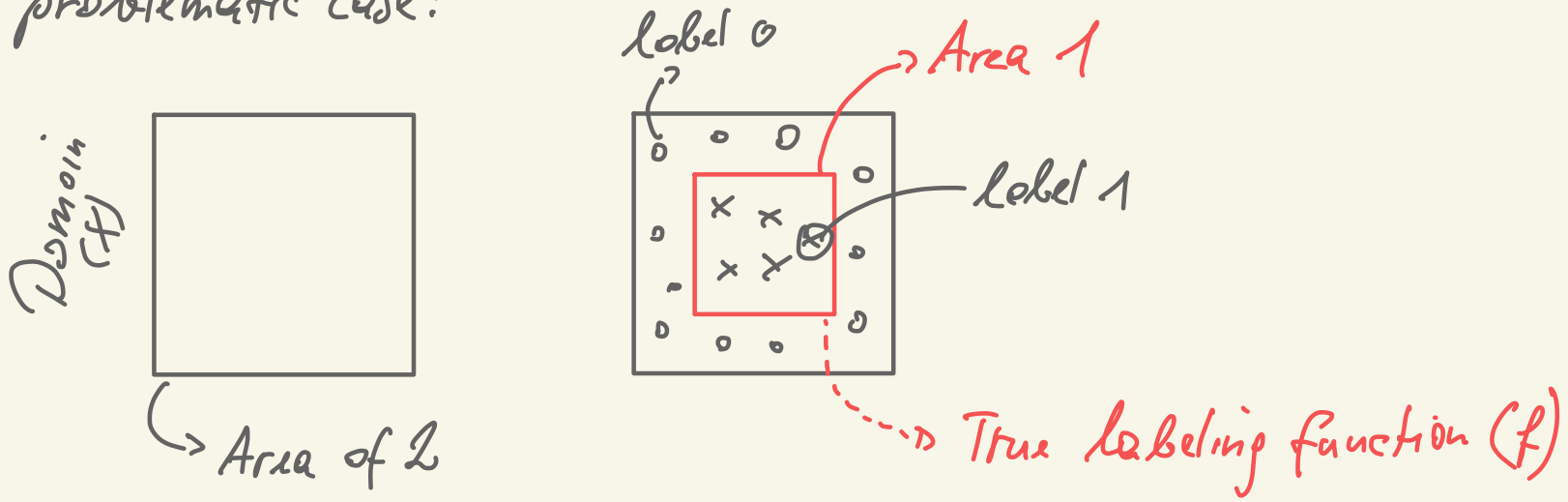
$$= \frac{1}{\cancel{m}}\cdot\cancel{m}\cdot\mathop{\mathbb{P}}_{x\sim D}\left[h(x)\neq f(x)\right]$$

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

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.

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

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 X to Y and write

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

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

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 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

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

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

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

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

We get (upon measuring with D):

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

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

// due to σ-sub-additivity ("Union bound")

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

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

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

// due to iid assumption

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

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

// as $h \in H_{BAD}$

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

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

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

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

if we let this be $< \delta \in (0,1)$ and solve for m, we get

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

ε: error; δ: 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 $m_H : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A with the following properties: (I) for every $\varepsilon, \delta \in (0,1)$ and (II) every distribution D over 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

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

→ ceiling function

We will now move to a more general setting.

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; $\min_{h \in \mathcal{H}} L_{D,f}(h)$)

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

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

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

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

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

(as we have $|a| \geq b \Leftrightarrow a \leq -b$ OR $a \geq b$)

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 $1-\delta$ our choice of δ)

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

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

(by the union bound)

We see that we did need realizability for that!

[2] Next, we release our requirement of a "true" labeling function f.

We do this, by letting D be a distribution over $X \times Y = Z$. We need to adjust our definitions of empirical error and generalization error:

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

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

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)

General loss functions:

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

Example: 0-1 loss

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

square loss

$$\ell^{sq}(h, (x,y)) = (y - h(x))^2$$

↳ r.p. $\in \mathbb{R}$

// r.p.: in regression problems

In case of general loss functions, we adjust our def. of L_D and L_S as follows:

$$L_D(h) = \mathop{\mathbb{E}}_{z}\left[\ell(h, z)\right], \quad L_S(h) = \frac{1}{m} \cdot \sum_{i=1}^{m} \ell(h, z_i)$$

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

$= \mathbb{P}_{z}\left[h(x)\neq y\right]$ // as before!

Uniform convergence

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

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

Lemma: Assume that S is $\varepsilon/2$-representative wrt. $\mathcal{Z}, H, \ell$ 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 $h \in H$:

$$\begin{aligned} L_D(h_S) &\leq L_S(h_S) + \varepsilon/2 && \text{// by def. of } \varepsilon\text{-representative-ness} \\ &\leq L_S(h) + \varepsilon/2 && \text{// as } h_S \text{ is an ERM hypothesis} \\ &\leq L_D(h) + \varepsilon/2 + \varepsilon/2 && \text{// by def. of } \varepsilon\text{-representative-ness} \\ &= L_D(h) + \varepsilon \end{aligned}$$

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

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

This leads us to the definition of uniform convergence.

Def. (uniform convergence): A hypothesis class H has the uniform convergence (UC) property (with respect to $\mathcal{Z}$, and loss function ℓ), if there exists $m_H^{UC}: (0,1)^2 \to \mathbb{N}$, such that for every $\varepsilon, \delta \in (0,1)$ and every distribution D over $\mathcal{Z}$, 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

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

Moreover, ERM is a successful APAC learner for H!

Claim: Finite $\mathcal{H}$ ($|\mathcal{H}| < \infty$) are agnostic PAC learnable.

Given fixed $\varepsilon, \delta \in (0,1)$, we want to show that

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

Equivalently,

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

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

(by union bound)

Lets fix $l \equiv l^{0-1}$. As we know that $L_D(h) = \mathop{\mathbb{E}}_{z}[l(h,z)]$ and $L_S(h) = \frac{1}{m}\cdot\sum_{i=1}^{m} l(h,z_i)$, we can use Hoeffding's inequality (with $b=1, a=0$):

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

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

Let $2\cdot|H|e^{-2\varepsilon^2 m}$ be smaller than δ, i.e.,

$$2\cdot|H|\cdot e^{-2\varepsilon^2 m} < \delta \iff m > \log\left(\frac{2\cdot|H|}{\delta}\right)\cdot\frac{1}{2\varepsilon^2}$$

This gives us the <u>sample complexity function</u> m_H^{uc} for uniform convergence. In other words, for $l : H \times Z \to [0,1]$, we have shown

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

and by our earlier corollary we get for the sample complexity function m_H for agnostic PAC learnability

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

This establishes the claim. □

(ended lec. on April 25)

Question: Is there a "universal" learner?, i.e., a learner A without any prior knowledge (with respect to $\mathcal{H}$ or $\mathcal{D}, \dots$) of a task, but can be challenged by any task (and achieves low $L_{\mathcal{D}}(A(S))$).

↳ specified by $\mathcal{D}$

Theorem (No-Free-Lunch): Let A be a learning algorithm for the task of binary 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 $\mathcal{D}$ over $X \times \{0,1\}$ such that

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

$$L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}.$$

2 means that the learner fails on that task; 1 means that the task can be successfully learned by another learner (e.g. ERR_H for $H = \{f\}$).

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

$$f_1, \ldots, f_T$$

For each f_i, we define

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

Hence, by construction, $L_{D_i}(f_i) = 0$.

With that in mind, we will show that for every learning algorithm A that receives m 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

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

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

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

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

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

Lets show that (xx) holds:

We know that there are $(2m)^m - k$ possible training sequences of size m (remember that $|C| = 2m$) from C. Lets call them

$$S_1, \ldots, S_k$$

Also, lets call S_j^i the sequence S_j labeled by f_i, i.e.,

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

with $S_j = (x_1, \ldots, x_m)$

Now, in case the distribution is D_i, then $S_1^i, \ldots, S_k^i$ are the possible training sequences that A can receive.

Note that, by construction (of the D_i's), all these training sequences have equal probability of being drawn/sampled.

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

↳ # possible training seq. of size m

(Let's remember that we want to show

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

)

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

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

↳

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

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

Lets fix some $j \in [k]$: ($S_j = (x_1, \ldots, x_m)$)

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 also know that $p \geq m$.

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

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

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

$$\geq \frac{1}{2p} \sum_{r=1}^{p} \mathbb{1}_{h(v_r) \neq f_i(v_r)} \qquad (***)$$

(as $p \geq m$)

Combining results gives

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

A run on S_j^i gives some $h : C \to \{0,1\}$ and we know that $(***)$ holds for every $h : C \to \{0,1\}$

$$= \frac{1}{2} \cdot \frac{1}{p} \sum_{k=1}^{p} \frac{1}{T} \cdot \sum_{i=1}^{T} \mathbb{1}_{A(S_j^i)(v_r) \neq f_i(v_r)}$$

$$\geq \frac{1}{2} \cdot \min_{k \in [p]} \frac{1}{T} \cdot \sum_{i=1}^{T} \mathbb{1}_{A(S_j^i)(v_r) \neq f_i(v_k)} \quad \text{// as } \sup \geq \min$$

Lets fix some $k \in [p]$:

Remember that we have $f_1, \ldots, f_T$ (for $T = 2^{2m}$ labelings)

Example: say $m=1$, $|C|=2$

	c_1	v_1	
f_1	0	0	labels
f_2	0	1	
f_3	1	0	
f_4	1	1	

→ partition into 2 pairs (f_1, f_2) and (f_3, f_4) where (f_1, f_2) only differ on v_1 and (f_3, f_4) only differ on v_1.

In general, we can partition into $T/2$ disjoint pairs of functions (f_i, f_i'). 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 $c = v_r$. For such a pair, we have $s_j^i = s_j^{i'}$.

(ended here on Day 2)