Learning theory

UML Ch. 3

$$S = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$$

F function class, loss function l

ERM

$$\hat{f} = \underset{f \in F}{\arg\min} \underbrace{\frac{1}{m} \sum_{i=1}^{m} l(f(x_i), y_i)}_{\hat{R}(f)}$$

Memorization prediction:

$$f_{mem}(x) = \begin{cases} y_i & \text{if } \exists (x_i, y_i) \in S \\ & x = x_i \\ 0 & \text{o.w.} \end{cases}$$

This leads to a 0 ERM,

but not a good predictor.

Generalization :

True risk $\quad R(f) = \underset{(x,y)\sim D}{\mathbb{E}} \left[l(f(x), y) \right]$

Loss over D = loss over unseen Examples

$$R(\hat{f}) = \underset{(x,y)\sim D}{\mathbb{E}} \left[l(\hat{f}(x), y) \right]$$

$$R(\hat{f}) = \underbrace{R(\hat{f}) - \hat{R}(\hat{f})}_{\text{generalization gap.}} + \hat{R}(\hat{f})$$

Assume a realizable case, i.e.,

$$\exists f_* \in \mathcal{F} \quad \text{s.t.} \quad R(f_*) = 0$$

we have a perfect ERM

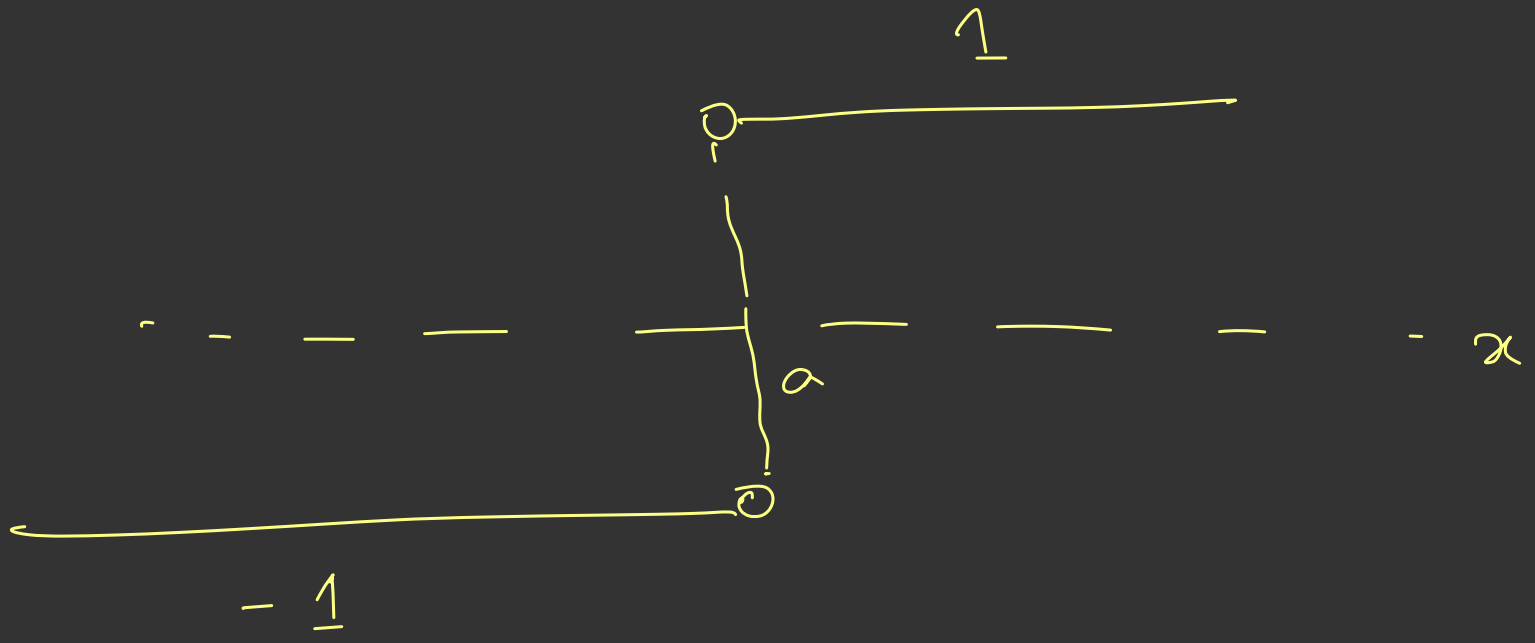
$$\hat{R}(\hat{f}) = 0$$

Q1. Can we find f_* ?

Q2. Can we find a good predictor for all dataset ?

Example - Threshold

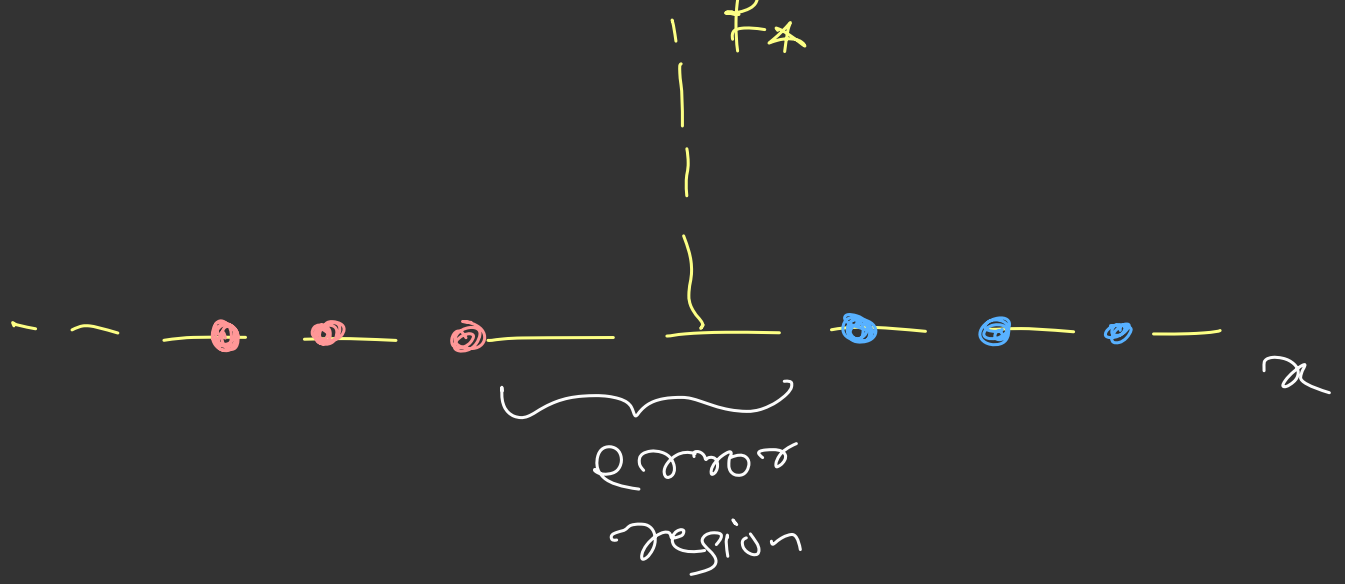
$$f_a(x) = \begin{cases} +1 & \text{if } x \geq a \\ -1 & \text{o.w.} \end{cases}$$



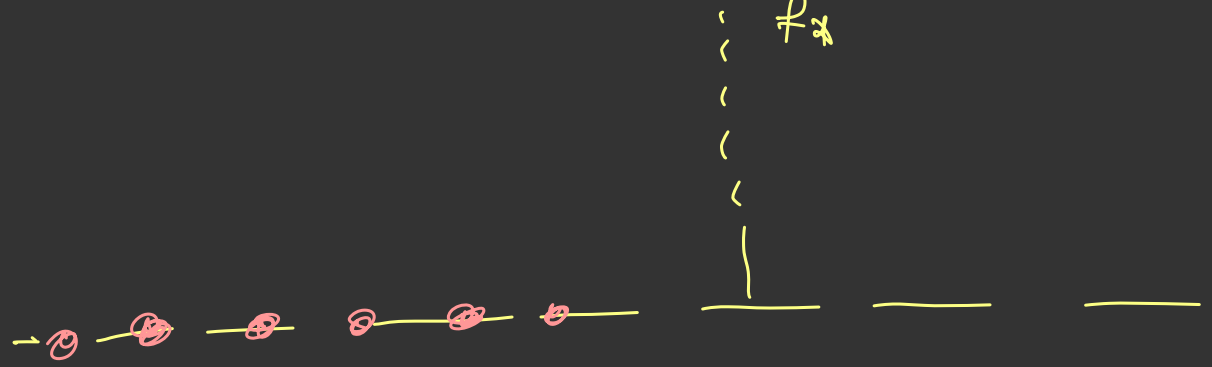
Zero risk ?

Data is uniformly distributed on the line

Draw 1:



Draw 2:



$$\Pr(\text{not seeing } \bullet) = (1-p)^n$$

Probably approximately correct (PAC):

Leslie Valiant (1984)

Error parameter: ε

Confidence parameter: δ

$$Pr\left[err(f) \leq \varepsilon\right] \geq 1-\delta$$

Defn: A function class F is PAC learnable if there exists an algorithm A and a function $m_F : (0,1)^2 \to \mathbb{N}$ with the following property:

For every labelling function $f \in F$, for every distribution D on the instance space X, and for all $\epsilon, \delta \in (0,1)$, if A has access S of size $m \geq m_F(\epsilon, \delta)$, then with probability $1-\delta$ (over the

choices of the training set), A outputs a predictor $\hat{f}$ s.t.

$$\Pr_{x \sim \mathcal{D}} \left[\hat{f}(x) \neq f(x) \right] \leq \varepsilon.$$

Finite function classes are PAC learnable

Consider $\mathcal{F} = \{ f_1, \ldots f_{|\mathcal{F}|} \}$

Binary classification.

Theorem: Every finite function class is PAC learnable with sample complexity

$$m_{\mathcal{F}}(\epsilon, \delta) \leq \left\lceil \frac{\log\left(|\mathcal{F}|/\delta\right)}{\epsilon} \right\rceil$$

Alternatively, $\hat{f}$ is any ERM obtained over a training set of size m, with prob. $1-\delta$

$$R(\hat{f}_S) \leq \frac{\log\left(|F|/\delta\right)}{m}$$

Proof :

Define $B = \{ f \in \mathcal{F} : f \text{ is } \epsilon\text{-bad} \}$

- we say $\hat{f}_S$ is ϵ-bad if $R(\hat{f}_S) > \epsilon$

- we need $P_r \left[\hat{f}_S \text{ is } \epsilon\text{-bad} \right] \leq \delta$

↳ ERM

So

$$P_{S r} \left[\hat{f}_S \text{ is } \epsilon\text{-bad} \right] = P_{S r} \left[\hat{f}_S \text{ is } \epsilon\text{-bad} \wedge \hat{R}(\hat{f}_S) = 0 \right]$$

$$\leq \Pr_S \left[\exists f \in \mathcal{F} : f \text{ is } \epsilon\text{-bad} \wedge \hat{R}(f) = 0 \right]$$

$$= \Pr_S \left[\exists f \in B : \hat{R}(f) = 0 \right]$$

$$\leq \sum_{f \in B} \Pr_S \left[\hat{R}(f) = 0 \right]$$

[Union bound]

$$= \sum_{f \in B} \Pr_S \left[\forall i \in [m] : f(x_i) = f_*(x_i) \right]$$

$$= \sum_{f \in B} \prod_{i=1}^{m} \Pr \left[f(x_i) = f_*(x_i) \right]$$

$$= \sum_{f \in B} \prod_{i=1}^{m} \left[1 - \Pr \left[f(x_i) \neq f_*(x_i) \right] \right]$$

$$= \sum_{f \in B} \prod_{i=1}^{m} \left[1 - R(f) \right]$$

$$\leq \sum_{f \in B} (1 - \epsilon)^m$$

$$= |B| (1 - \epsilon)^m \leq |F| (1 - \epsilon)^m$$

$B \subseteq F$

$$\leq \underbrace{|F| \exp(-m\epsilon)}_{:= \delta}$$

$$\leq \delta$$

$1 - a \leq \exp(-a)$

$$|F| \exp(-m\epsilon) \leq \delta$$

$$\Rightarrow \quad m \geq \frac{\log(|F|/\delta)}{\epsilon}$$

Example:

$$F = \left\{ \underline{w} \mid \sigma(\underline{w}^T \underline{x}), \quad w_i \in \{-10, \ldots 0, \ldots 10\} \right\}$$

$$|F| = 21^d$$

$$m_F(\epsilon, \delta) \leq \frac{d \log 21 + \log(1/\delta)}{\epsilon}$$

Are infinite classes PAC learnable?

Growth function:

$$\Pi_F(S) = \{ f(x_1), \ldots f(x_m) \}$$

All possible labellings that the training points could have according to F

Growth function: (max. possible labellings over all sets of size m)

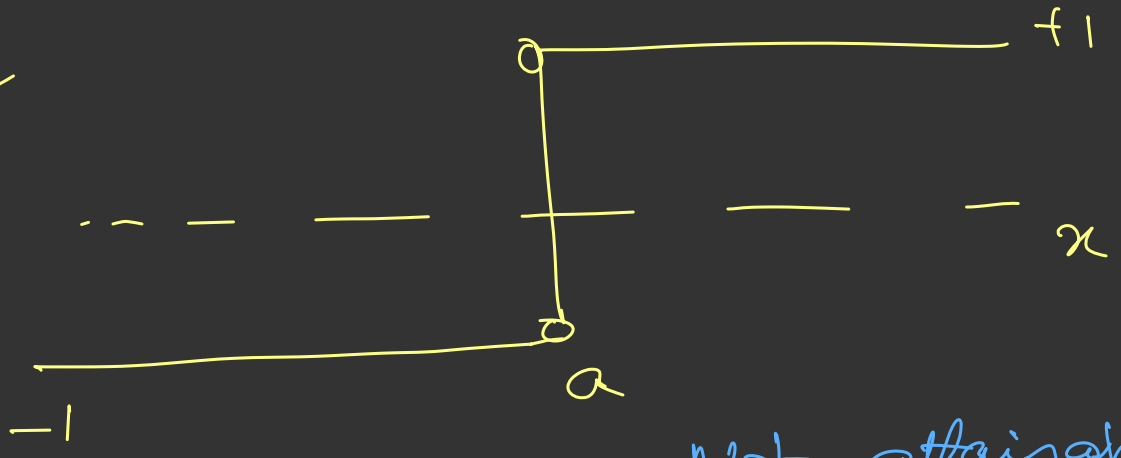
$$\Pi_F(m) := \max_{S : |S| = m} |\Pi_F(S)|$$

Example:

$$f_a(x) = \begin{cases} +1 & \text{if } x \geq a \\ -1 & \text{o.w.} \end{cases}$$

Consider a dataset with $x_1 < x_2 < x_3$

$\Pi_F(3) = 4$



-1, -1, -1

-1, -1, +1

-1, +1, +1

+1, +1, +1

Not attainable

-1, +1, -1

+1, -1, +1

+1, +1, -1

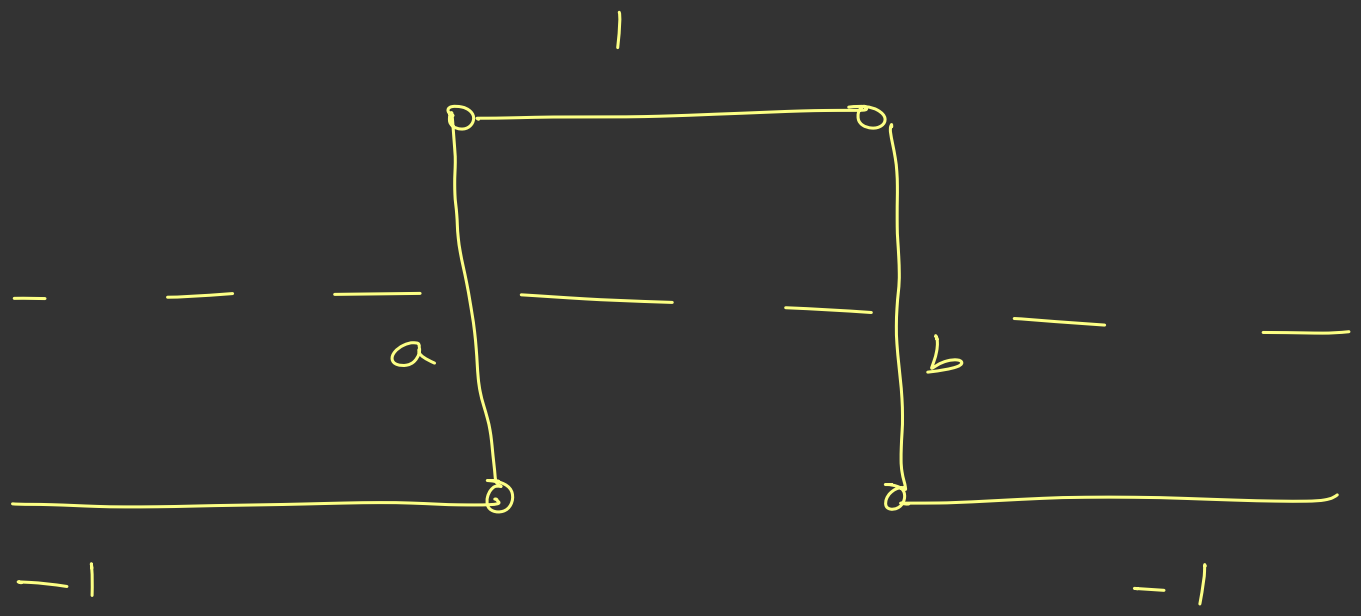
+1, -1, -1

more generally

$\Pi_F(m) = m+1 \ll 2^m$

Example:

$$f_{a,b}(x) = \begin{cases} +1 & x \in [a,b) \\ -1 & \text{o.w.} \end{cases}$$



w.l.o.g. $x_1 < x_2 < x_3$

only $1, -1, 1$ is not attainable

$$\Pi_{\mathcal{H}}(3) = 7$$

$$\Pi_{\mathcal{H}}(m) = \frac{m(m+1)}{2} + 1$$

General bound on $\Pi_F(m)$.

VC dimension of F

/ \

Vapnik – Chervonenkis (1970)

Def (Shattering):

A set S of inputs is said to be shattered by a function class F if $|\Pi_F(S)| = 2^{|S|}$.

Def (VC dimension):

VC dimension of a function class F or $VC(F)$ is the size of the largest set S that can be shattered by F.

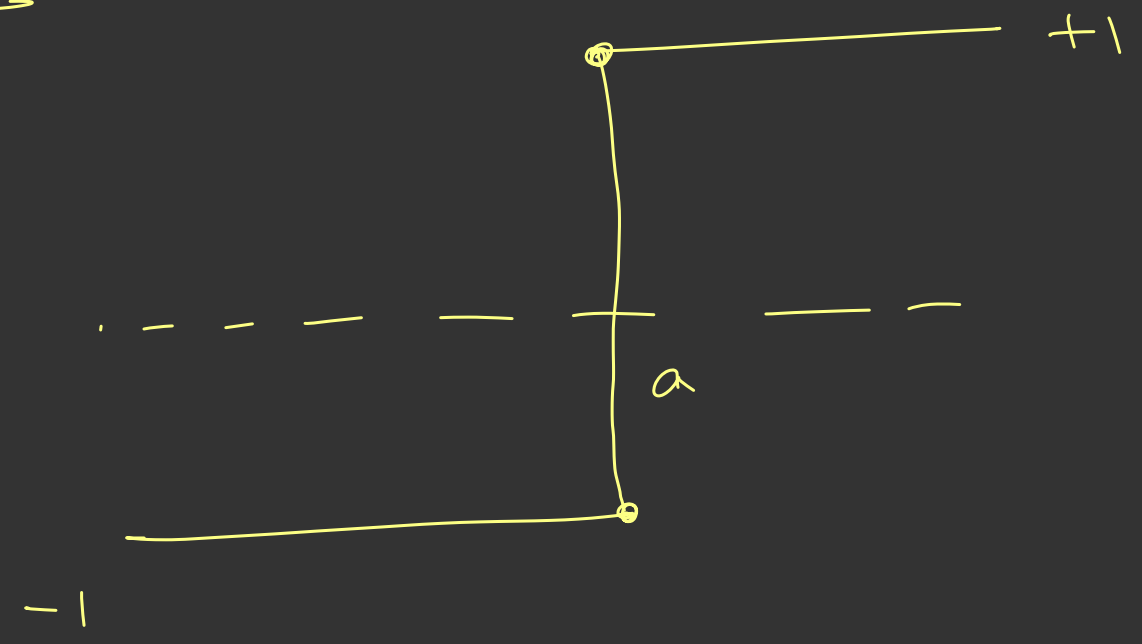
To show that a function class has $VC(F) = d$, we must show that

- ✱ There is a set S of d points that is shattered by F
- ✱ There is not set S of $d+1$ points that is shattered by F.

Examples:

$$f_a(x) = \begin{cases} +1 & \text{if } x \geq a \\ -1 & \text{o.w.} \end{cases}$$

① Threshold:



Any set of 1 can be shattered,
however, no set of size 2 can be shattered.

Thus, $VC(F) = 1$

② Intervals:

$$VC(F) = 2$$

③ Halfspaces: ?