CIS 5200: MACHINE LEARNING

LEARNING THEORY

Surbhi Goel

Content here draws from material by Rob Schapire (Princeton), Hamed Hassani (UPenn) and Michael Kearns (UPenn)

Spring 2023

LEGES SINE MORIBUS VANAE

OUTLINE - TODAY

- Survey Overview
- What about generalization?
- Probably Approximately Correct (PAC) learning
- Finite Function Classes are PAC learnable

SUPERVISED LEARNING - SO FAR

- Training dataset $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$
- Function class $\mathcal{F}$, loss function ℓ
- Empirical Risk Minimizer:

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

We have looked at various methods to find the ERM

Is this good enough for learning?

MEMORIZATION

Memorizer predictor $f_{\text{mem}}(\cdot)$

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

This gets 0 training loss $\hat{R}(f_{\text{mem}}) = 0$, so it is an ERM.

But is it a good predictor?

GENERALIZATION

We want the predictor to perform well not just on the training data but on examples it will see in the future.

Recall how we formalized this:

Training dataset is drawn independently and identically from some unknown but fixed distribution $\mathcal{D}$

loss on future examples = loss over the distribution

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

We ideally want to minimize true risk $\boldsymbol{R}$ and not just empirical risk $\hat{\boldsymbol{R}}$

GENERALIZATION

loss on future examples = loss over the distribution

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

We don't have access to the true risk, we can only see a training set

ERM can guarantee that this is small

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

In this lecture, we will bound this generalization gap

This will depend on the size of the training set and the complexity of the function class $\mathcal{F}$

LET US FORMALIZE THIS!

$\exists f_* \in \mathscr{F}$ *such that* $R(f_*) = 0$

Let us work in the classification setting and assume that there is a perfect classifier (realizable learning model)

$\hat{R}(\hat{f}) = 0$ *since* $\hat{R}(f_*) = 0$

We want to show that $R(\hat{f})$ is small for any empirical risk minimizer $\hat{f}$

Challenge 1: Can we find exactly f_*?

Challenge 2: Can we find a good predictor for all datasets?

PROBABLY APPROXIMATELY CORRECT (PAC) LEARNING

Introduced by Leslie Valiant in 1984, captures the notion of finding approximately good predictors with high probability

Error parameter ϵ

Confidence parameter δ

Definition:

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

for every labelling function $f \in \mathscr{F}$, for every distribution $\mathscr{D}$ on feature space $\mathscr{X}$, and for all $\epsilon, \delta \in (0,1)$, if $\mathscr{A}$ is given access to a training dataset S of size $m \geq m_{\mathscr{F}}(\epsilon, \delta)$ where the features are drawn from $\mathscr{D}$ and labels are according to f, then with probability $1 - \delta$ (over the choice of the training dataset), $\mathscr{A}$ outputs a predictor $\hat{f}$ such that $\Pr_{x \sim \mathscr{D}} \left[\hat{f}(x) \neq f(x) \right] \leq \epsilon$.

PROBABLY APPROXIMATELY CORRECT (PAC) LEARNING

Definition:

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

for every labelling function $f \in \mathscr{F}$, for every distribution $\mathscr{D}$ on feature space $\mathscr{X}$, and for all $\epsilon, \delta \in (0,1)$, if $\mathscr{A}$ is given access to a training dataset S of size $m \geq m_{\mathscr{F}}(\epsilon, \delta)$ where the features are drawn from $\mathscr{D}$ and labels are according to f, then with probability $1 - \delta$ (over the choice of the training dataset), $\mathscr{A}$ outputs a predictor $\hat{f}$ such that $\Pr_{x \sim \mathscr{D}} \left[\hat{f}(x) \neq f(x) \right] \leq \epsilon$.

Function $m_{\mathscr{F}} : (0,1)^2 \to \mathbb{N}$ captures the sample complexity of learning

Depends on complexity of $\mathscr{F}$

EXAMPLE - NOT PAC LEARNABLE

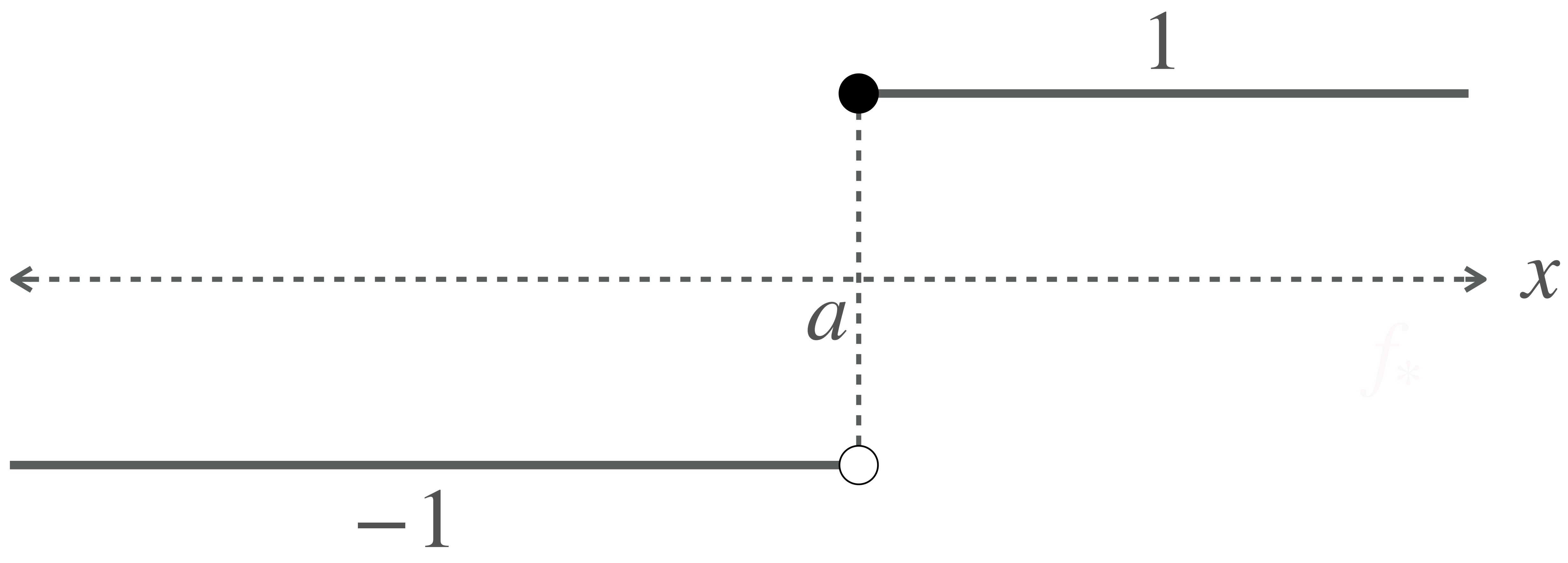
Class of all possible predictors from $\mathscr{X} \to \{-1,1\}$ is not PAC learnable, for any dataset of size m we would have only seen the labels on those m points.

The true function could take any value outside!

We cannot possibly guarantee small generalization error!

EXAMPLE - PAC LEARNABLE

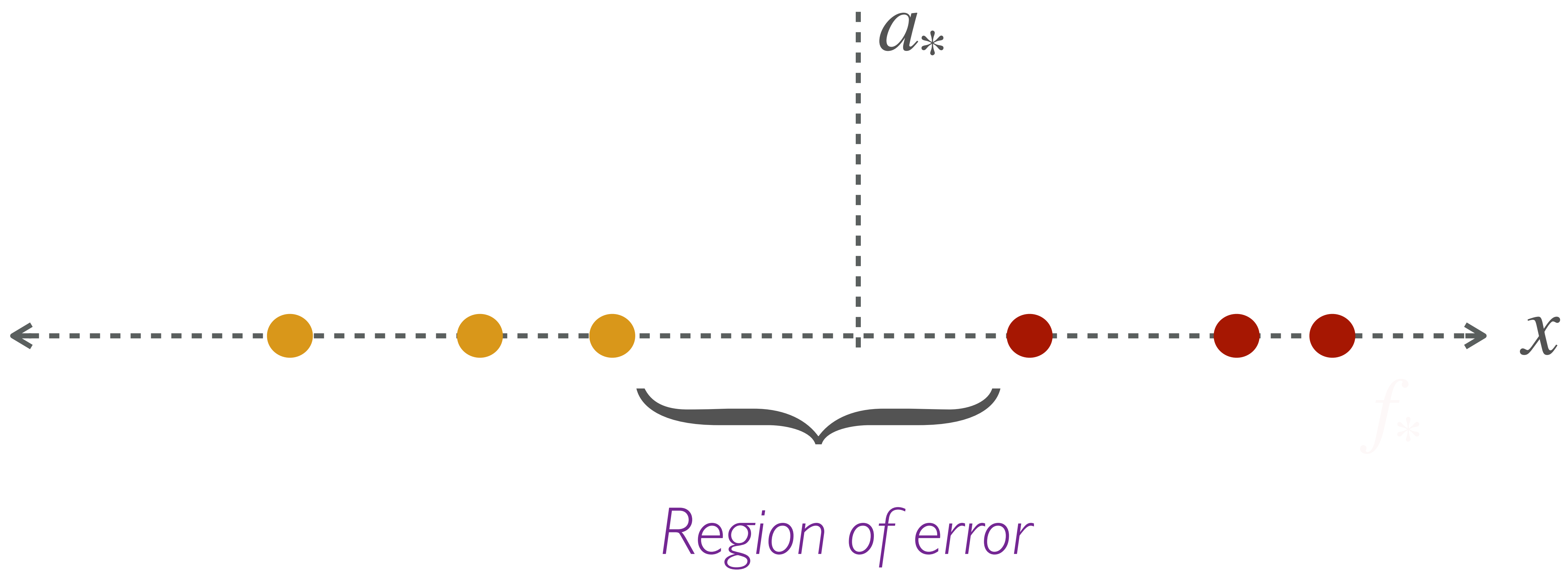
$$f_a(x) = \begin{cases} 1 & \text{if } x \geq a \\ -1 & \text{otherwise.} \end{cases}$$



One dimensional half space or thresholds

EXAMPLE - PAC LEARNABLE

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



As we see more and more samples, this error will shrink

In the next lecture we will quantify how many samples we will need for this

GENERAL - FINITE CLASSES ARE PAC LEARNABLE

Consider a finite function class $\mathcal{F} = \{f_1, \ldots, f_{|\mathcal{F}|}\}$

Theorem:

Every finite function class $\mathcal{F}$ is PAC learnable with sample complexity

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

Observe that it depends on the size of $\mathcal{F}$ which is a natural notion of complexity of $\mathcal{F}$

GENERAL - FINITE CLASSES ARE PAC LEARNABLE BY ERM

Consider a finite function class $\mathscr{F} = \{f_1, \ldots, f_{|\mathscr{F}|}\}$

Theorem:

Every finite function class $\mathscr{F}$ is PAC learnable with sample complexity

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

where the algorithm $\mathscr{A}$ is any empirical risk minimization algorithm.

Proof on the iPad

GENERAL - FINITE CLASSES ARE PAC LEARNABLE BY ERM

Another way to state this is:

Theorem:

For any ERM $\hat{f}$ evaluated over training set of size m, with probability $1 - \delta$,

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

SUMMARY

We studied the notion of PAC learning where we allowed approximately correct learning with high probability

We proved that finite classes are PAC learnable using ERM

Next class: How do we handle infinite classes?