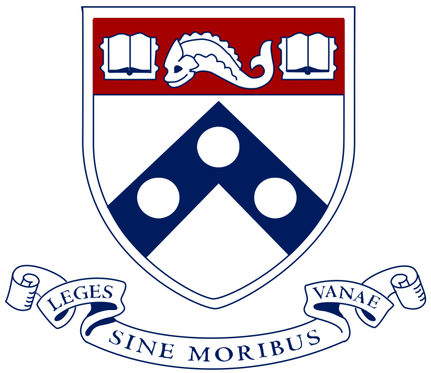
CIS 5200: MACHINE LEARNING

LEARNING THEORY

Surbhi Goel



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

Spring 2023

OUTLINE - TODAY

- Recap:
  - Probably Approximately Correct (PAC) learning
  - Finite Function Classes are PAC learnable
- What about infinite classes?
  - VC Dimension
  - VC Classes are PAC Learnable

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 $\mathscr{D}$

loss on future examples = loss over the distribution

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

We want to minimize true risk R but only have access to a training set

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

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

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

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.

FINITE CLASSES ARE PAC LEARNABLE BY ERM

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

Theorem:

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

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

EXAMPLE - INTEGER WEIGHT LINEAR CLASSIFIER

Consider the class $\mathscr{F}$ of integer weight linear classifiers where the parameter w is such that each $w_i \in \{-10,\ldots,0,\ldots,10\}$ for $i \in [d]$

How many possible functions are there? 21^d

What is the sample complexity for PAC learning?

$$m_{\mathscr{F}}(\epsilon,\delta) \lesssim \frac{d\log 21 + \log(1/\delta)}{\epsilon}$$

WHAT ABOUT INFINITE CLASSES?

If the size of the class is infinite, then the previous bound is useless.

What quantity should replace $\log|\mathcal{F}|$?

Let us think about the behavior of the function on our training dataset:

$$\Pi_{\mathcal{F}}(S) = \{(f(x_1), \ldots, f(x_m)) : f \in \mathcal{F}\}$$

This is all possible labeling that the training points could have according to $\mathcal{F}$

WHAT ABOUT INFINITE CLASSES?

Let us think about the behavior of the function on our training dataset:

$$\Pi_{\mathcal{F}}(S) = \{(f(x_1), \ldots, f(x_m)) : f \in \mathcal{F}\}$$

Define the maximum possible labelings over all training sets of size m

$$\Pi_{\mathcal{F}}(m) = \max_{S;|S|=m} |\Pi_{\mathcal{F}}(S)|$$

Growth function

What is an upper bound on $\Pi_{\mathcal{F}}(m)$? 2^m

But for many $\mathcal{F}$, this is actually much smaller!

(SOME) INFINITE CLASSES ARE PAC LEARNABLE BY ERM

Consider an infinite function class $\mathscr{F}$

Theorem:

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

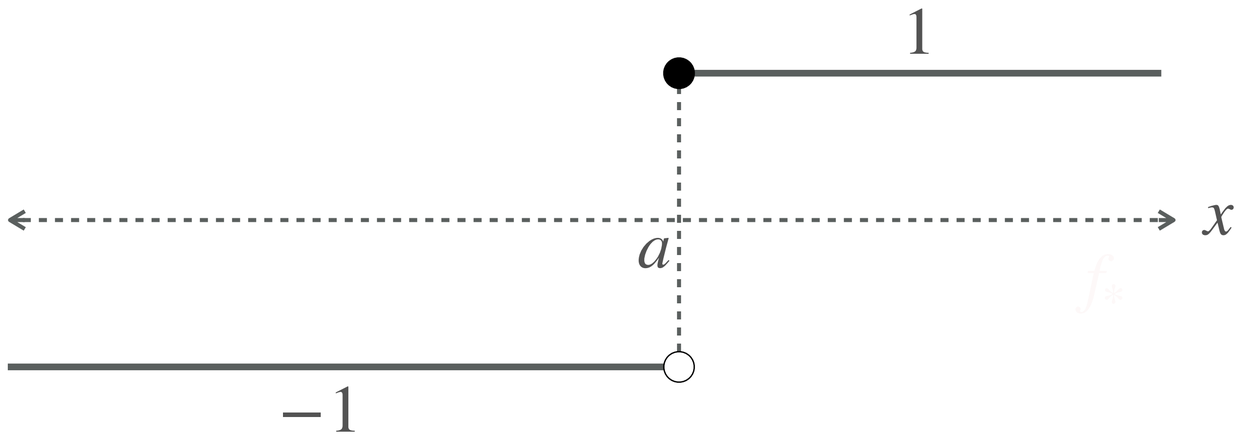
$$R(\hat{f}_S) \leq \left\lceil \frac{\log(|\Pi_{\mathscr{F}}(2m)|/\delta)}{m} \right\rceil .$$

Note that if $\Pi_{\mathscr{F}}(m) = 2^m$ then this is vacuous!

We will not cover the proof in class since it is a bit involved

EXAMPLE - THRESHOLDS

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

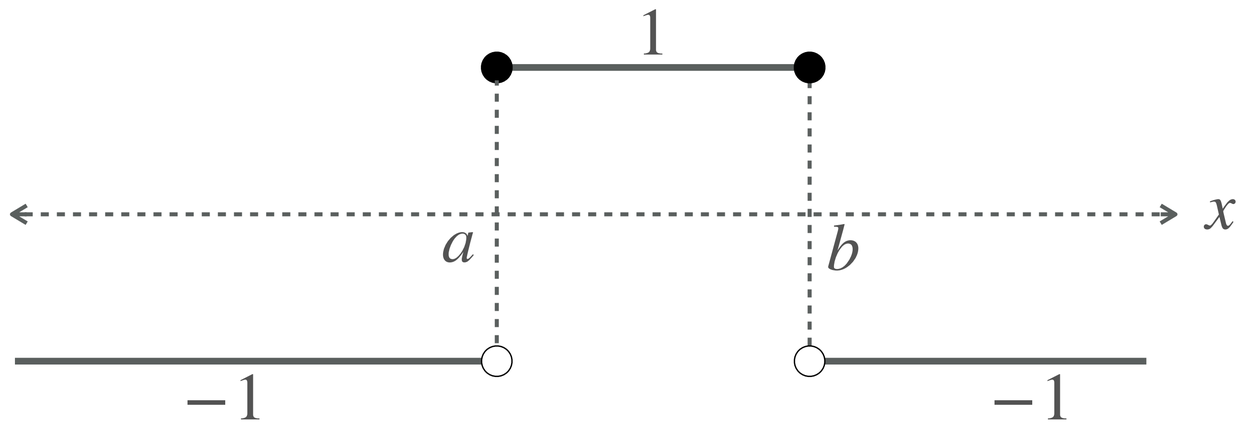


Consider a dataset of three points $x_1 < x_2 < x_3$

How are the possible labelings with the class of thresholds?

EXAMPLE - INTERVALS

$$f_{a,b}(x) = \begin{cases} 1 & \text{if } a \le x \le b \\ -1 & \text{otherwise.} \end{cases}$$



Consider a dataset of three points $x_1 < x_2 < x_3$

How are the possible labelings with the class of intervals?

GENERAL BOUNDS - VC DIMENSION

There is a notion of complexity called Vapnik-Chervonenkis (VC) dimension that bounds $\Pi_{\mathcal{F}}(m)$

Definition (shattering):

A set S of inputs is said to be shattered by function class $\mathcal{F}$ if $|\Pi_{\mathcal{F}}(S)| = 2^{|S|}$, that is, $\mathcal{F}$ can realize all possible labelings for the set of points in S.

Definition (VC dimension):

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

GENERAL BOUNDS - VC DIMENSION

Definition (VC dimension):

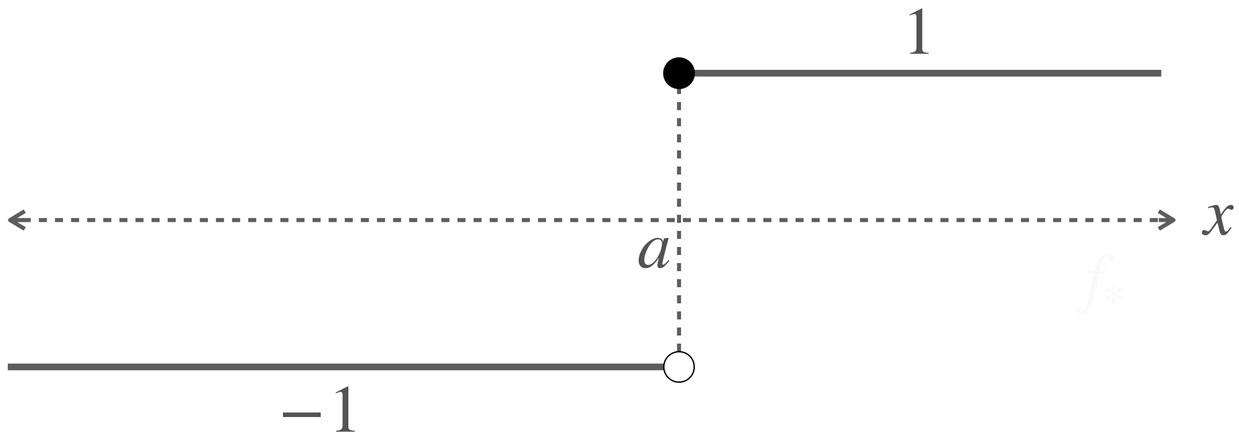
VC dimension of a function class $\mathcal{F}$ ($VC(\mathcal{F})$) is the size of the largest set S that can be shattered by $\mathcal{F}$.

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

- There is a set S of d points that is shattered by $\mathcal{F}$
- There is no set S of $d + 1$ points that is shattered by $\mathcal{F}$

EXAMPLE - THRESHOLDS

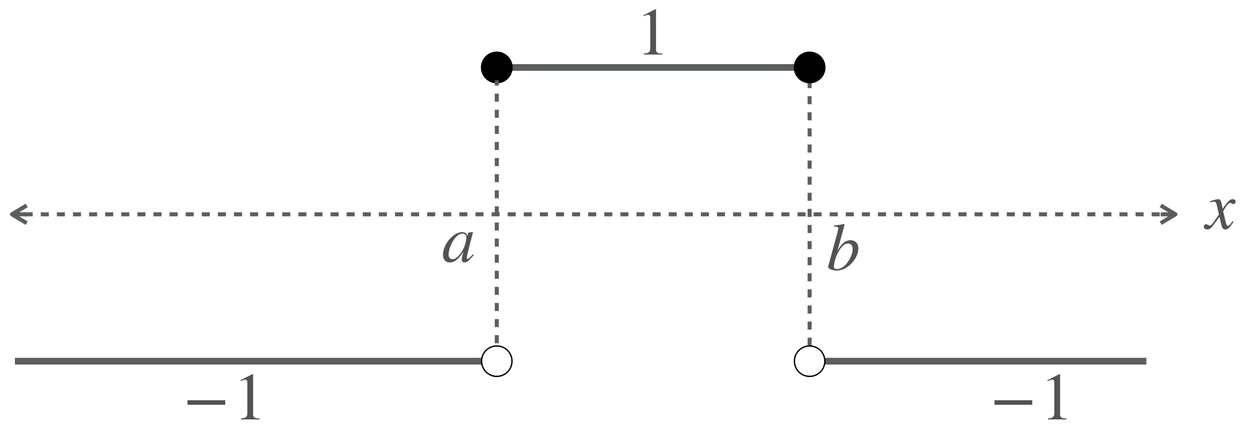
$$f_a(x) = \begin{cases} 1 & \text{if } x \geq a \\ -1 & \text{otherwise.} \end{cases}$$



What is the VC dimension of the class of thresholds?

EXAMPLE - INTERVALS

$$f_{a,b}(x) = \begin{cases} 1 & \text{if } a \le x \le b \\ -1 & \text{otherwise.} \end{cases}$$



What is the VC dimension of the class of intervals?

CONNECTION - VC DIMENSION & GROWTH FUNCTION

Theorem (Sauer's Lemma):

Let $d = VC(\mathcal{F})$, then

- $\Pi_{\mathcal{F}}(m) = 2^m$ for $m \leq d$
- $\Pi_{\mathcal{F}}(m) = O(m^d)$ for $m > d$

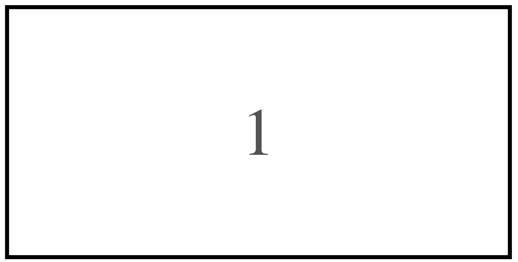
Theorem:

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

$$R(\hat{f}_S) \lesssim \frac{d \cancel{\log m} + \log(1/\delta)}{m}.$$

EXAMPLE - RECTANGLES

$$f(x) = \begin{cases} 1 & \text{if inside the rectangle} \\ -1 & \text{otherwise.} \end{cases}$$

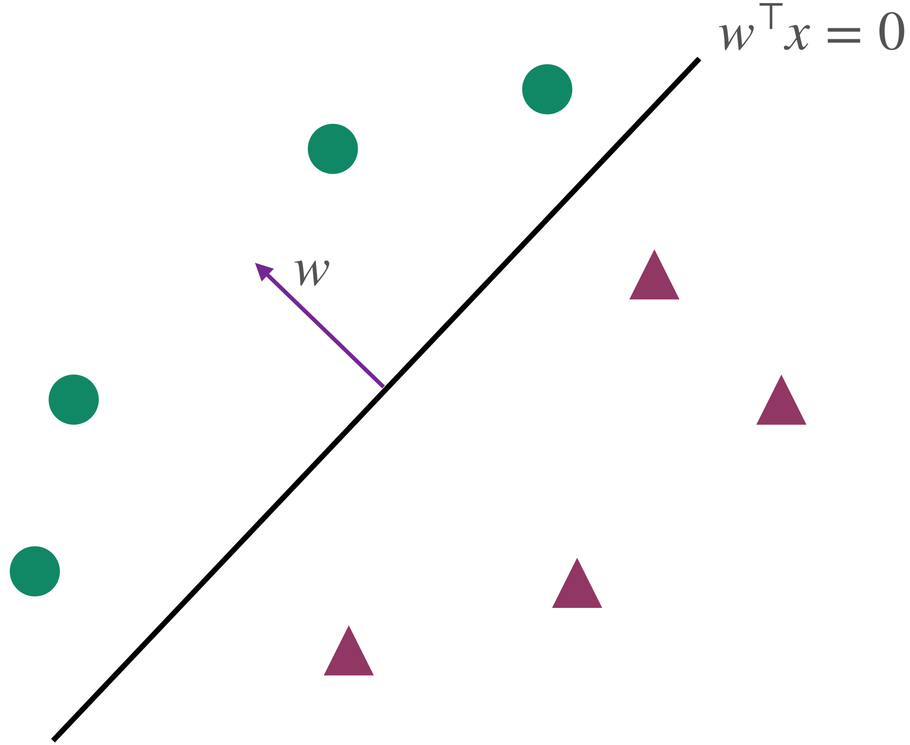


-1

What is the VC dimension of the class of rectangles?

EXAMPLE - LINEAR CLASSIFIERS

$$f_w(x) = \text{sgn}(w^\top x)$$



What is the VC dimension of the class of linear classifiers?

CONCLUSION

Theorem:

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

$$R(\hat{f}_S) \lesssim \frac{d + \log(1/\delta)}{m} .$$

To quantify how many samples we need to learn a particular function class, we can use the VC dimension as a complexity measure