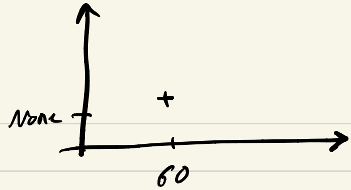
Lecture 2

PAC Learning.

Example 1: Running base on temperature precipitation

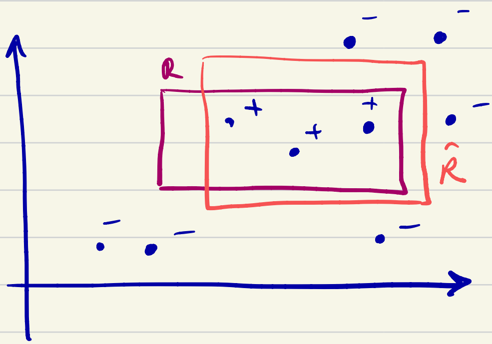


Learning an axis-aligned rectangle R in $\mathbb{R}^2$

Samples: points $p_1, \dots, p_n \sim D$ over $\mathbb{R}^2$

label $y_1, \dots, y_n$

$$y_i = \begin{cases} +1 & \text{if } p_i \in R \\ -1 & \text{otherwise} \end{cases}$$



Goal: output $\hat{R}$ s.t. error of $\hat{R}$ is small (say ϵ) with high probability (say $1-\delta$)

$$\mathrm{err}(\hat{R}) = \Pr_{p \sim D}[\hat{R} \text{ mislabel } p]$$

$$= \Pr_{p \sim D}\left[\begin{array}{lcl} (p \in R & \text{and} & p \notin \hat{R}) \\ & \text{or} & \\ (p \notin R & \text{and} & p \in \hat{R}) \end{array}\right]$$

D is arbitrary but fix.

while D can be potentially unusual / irregular, the notion of error is also defined based on the same D.

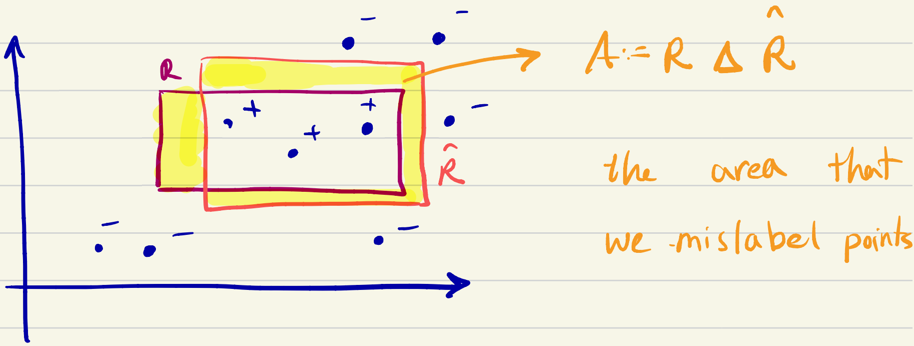
Solution:

Algorithm:

1- Draw m samples (for sufficiently large)

2- Set $\hat{R}$ to be a rectangle that

correctly label all the sample points



$$\text{err}(\hat{R}) = \Pr_{p \sim D}[p \in A] = D(A)$$

by our definition of $\hat{R}$, there is no sample point in $A := R \Delta \hat{R}$

If $\text{err}(\hat{R}) > \varepsilon \Rightarrow D(A) > \varepsilon$

How likely it is to not see any sample from A?

Ideally, we want:

$$\Pr_D[\# \text{ samples in } A = 0] \overset{?}{\leq} \delta$$

$$= (1 - D(A))^m \leq (1-\varepsilon)^m \quad \left(\begin{matrix}\text{independent} \\ \text{samples}\end{matrix}\right)$$

$$\leq e^{-\varepsilon m}$$

set $m = \frac{\log 1/\delta}{\varepsilon}$

$$\leq \delta$$

$\Rightarrow$ Hence, with probability at least $1-\delta$

$\mathrm{err}(\hat{R}) \leq \varepsilon$.

efficient $\begin{cases} \# \text{ samples } = O\left(\frac{\log 1/\delta}{\varepsilon}\right) \\ \text{time } O(m) \end{cases}$

Well behaved target class

Probably Approximately Correct (PAC)

X instance space (input space / domain) — set of all instances

$c : X \to \{+1, -1\}$ concept — a function to label elements

C concept class — a collection of labeling functions

c^* target concept — $c^* \in C$ and label all instances correctly

D target distribution — distribution over instances

sample / training data set

$$\left\{ \begin{array}{l} \langle x_1 , c^*(x_1) \rangle \\ \langle x_2 , c^*(x_2) \rangle \\ \quad \vdots \\ \langle x_n , c^*(x_n) \rangle \end{array} \right.$$

+ "distribution free" setting

samples drawn from an arbitrary distribution.
but error is measured according to the same distribution.

Some papers focus on specific class of distributions such as Gaussians.

+ We say we are in the realizable case if there exists a concept $c^* \in C$ that label all the instances in the domain perfectly

+ The goal is to find an unknown target concept c in a known concept class using labeled samples.

- find $\hat{c}$ in C with small error w.h. prob.
- Efficiency : # samples & time

PAC learning (Probably Approximately Correct)

Suppose that we have a concept class C over X. We say that C is **PAC learnable** if there exists an algorithm A s.t:

$\forall\ c \in C$, $\forall\ D$ over X, $\forall\ \varepsilon, \delta \in (0, 0.5]$

A receives ε, δ, and samples $\langle x_1, c(x_1)\rangle, \ldots, \langle x_n, c(x_n)\rangle$ where x_i's are iid samples from D.

proper $[\in C]$

Then, w. p. $\geq 1-\delta$, A outputs $\hat{c}$ s.t.

$$\mathrm{err}(\hat{c}) \leq \varepsilon.$$

The probability is taken over the randomness in the samples and any internal coin flips of A.

+ Usually efficiency means :

sample complexity & time complexity

$= O(\text{ploy}(1/\varepsilon, 1/\delta))$

+ ε = error parameter

δ = confidence parameter

These two parameters capture two kinds of error :

ε: small discrepancy between concepts is not detectable.

δ: with some small probability, the sample set is not representative of reality.

other notation

true error:

$$\mathrm{err}(c) = \Pr_{(x,y)\sim D}\left[c(x) \neq y \right]$$

training error:

$$\widehat{\mathrm{err}}(c) = \frac{\#\text{ samples in } T \text{ s.t. } c(x_i) \neq y_i}{|T|}$$

fraction of samples in the training set that c is mis-labeled.

ERM

In both example we picked concepts $\hat{R}$ and $\hat{h}$ that were consistent with the samples in the training set

What we did is called :

ERM : Empirical Risk Minimization

comes from samples   error

ERM algorithm: it finds a concept $\hat{h}$ such that $\hat{err}(\hat{h}) = 0$

+ Uniform convergence. (UC)

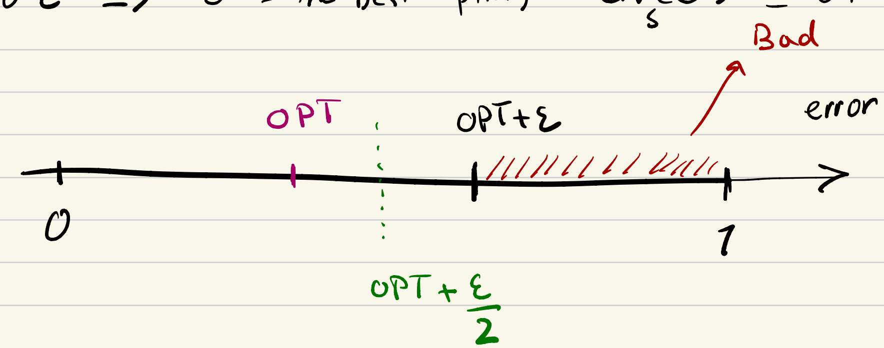
Class C has the uniform convergence property if $\forall\ \varepsilon, \delta \in (0,1)$, dist D $\exists\ m$ (as a function of $\varepsilon, \delta, \mathcal{H}$, but not D since we don't know D). s.t. for a training set of size m:

$$\Pr_{T\sim D^m}\left[\forall c \in C : \left|\widehat{err}_T(c) - err(c)\right| \le \varepsilon\right] \ge 1-\delta$$

Uniform convergence implies agnostic PAC learnability via EMR.

$UC \Rightarrow \forall c \in C_B \quad \widehat{err}_S(c) > OPT + \varepsilon/2$

$UC \Rightarrow c^* =$ the best option) $\quad \widehat{err}_S(c^*) \le OPT + \varepsilon$



There are two types of error in the agnostic setting:

$$\mathrm{err}(\hat{c}) < \underbrace{\min_{c \in C} \mathrm{err}(c)}_{\varepsilon_{app} := \text{approximation error}} + \underbrace{\varepsilon}_{\varepsilon_{est} := \text{estimation error}}$$

$\downarrow$

depends only to the choice of the class C

- Is C rich enough to capture how data is labeled?

C larger more complex $\uparrow$ | ε_{app} $\downarrow$ | ε_{est} $\uparrow$

$\boxed{\text{* ERM works for a finite class } C \text{ if we have enough samples.}}$

- Problem setup:

samples $(x_1, y_1), \ldots, (x_m, y_m) \sim D$

$c \in C : \text{err}(c) := \Pr_{(x,y) \sim D} [c(x) \neq y]$

Realizable case

Assume $\exists\, c^* \in C$ s.t. $\text{err}(c^*) = 0$

- Goal

find $\hat{c} \in C$ s.t. with probability $1-\delta$, $\text{err}(\hat{c}) \leq \varepsilon$.

- Proof

Bad hypotheses $C_B := \{c \in C \mid \text{err}(c) > \varepsilon\}$

training set

$$\hat{err}_T(c) := \frac{|\{(x,y) \in T \mid c(x) \neq y\}|}{|T|}$$

Misleading training samples

$$M := \{ T \mid \exists c \in C_B \text{ s.t. } \hat{err}_T(c) = 0 \}$$

upon observing T, we may pick c that is a bad choice, but it "looked" good from ERM perspective, since $\hat{err}_T(c) = 0$.

Our goal is to show observing a dataset $T \in M$ happens only with probability δ.

This is sufficient to prove $\star$.

fix $c \in C_B$.

what is the probability of

$$\hat{err}_T(c) = 0$$

$$\Pr_{T \sim D^m}\left[\hat{err}_T(c) = 0 \right]$$

$$= \Pr_{T \sim D^m}\left[\forall (x,y) \in T \,.\, c(x) = y \right]$$

iid samples $\rightarrow$ $$= \left(\Pr_{(x,y) \sim D}\left[c(x) = y \right] \right)^m$$

$err(C) > \varepsilon$ $\rightarrow$ $$< (1-\varepsilon)^m \leq e^{-\varepsilon m}$$

Now, we are ready to bound

$$\Pr_{T \sim D^m} [T \in \mathcal{M}]$$

$$= \Pr_{T \sim D^m} [\exists c \in C_B \text{ s.t. } \hat{err}_T(c) = 0]$$

$$= \sum_{c \in C_B} \Pr_{T \sim D^m} [\hat{err}_T(c) = 0]$$

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

set $m = \dfrac{\log(|C|/\delta)}{\varepsilon}$

$\Rightarrow$ Pr[outputting a misleading c]

$\leq \delta$

$\square$

The agnostic case :

What if there is no perfect $c \in C$?

$$\forall c \in C \quad err(c) > 0$$

Goal

Find $\hat{c} \in C$ s.t.

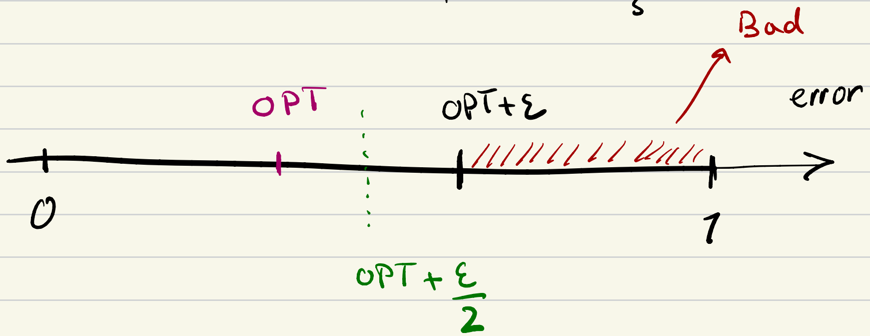
$$err(\hat{c}) < \underbrace{\min_{c \in C} err(c)}_{= OPT} + \varepsilon$$

the best possible option

$\boxed{\text{Uniform convergence implies agnostic PAC learnability via EMR.}}$

$UC \Rightarrow \forall c \in C_B \quad \hat{err}_S(c) > OPT + \varepsilon/2$

$UC \Rightarrow c^* = \text{the best option)} \quad \hat{err}_S(c^*) \leq OPT + \varepsilon$



Exercise!

Suppose we have a finite class C,

and $m = O\left(\frac{(\log |C| / \delta)}{\varepsilon^2}\right)$. then w.p. at least

$1-\delta$, for all $c \in C$, we have:

$$|\hat{err}_S(c) - err(c)| < \varepsilon/2$$

No free lunch theorem says if there is no universal learner :

for a complex C even when ε_{app} is 0, $\varepsilon_{est} \gg$ constant with some constant probability

[unless we have $\Omega(|X|)$ samples]

Suppose we have a set of $2m$ points

There are 2^{2m} possible labelings of these $2m$ points.

Suppose C is the class of 2^{2m} func. that assigns these labelings to these points.

Assume this is the true labeling.

Fix a labeling of the points

Now assume D is the uniform distribution on the 2m points with their label.

T ← Draw m samples from D

(WLOG assume they are unique)

How many function in C label T correctly? 2^m

$$P := \{ c \in C \mid \hat{err}_T(c) = 0 \}$$

↳ promising hypothese. $|P| = 2^{m/2}$

How many of them has error $< \varepsilon$?

c is misleading if $\begin{cases} err(c) > \varepsilon \\ \text{and } \hat{err}_T(c) = 0 \end{cases}$

$$\mathcal{M} := \{ c \in \mathcal{C} \mid err(c) > \varepsilon \ \& \ \hat{err}_T(c) = 0 \}$$

$$|\mathcal{M}| = \frac{|\mathcal{M}|}{|P|} \cdot |P|$$

$$= 2^m \cdot \Pr_{c \sim_U P}[c \in \mathcal{M}]$$

a random concept in P

c makes $\geq m.\varepsilon$ mistakes in expectation

$$= 2^m \cdot \Pr\left[\frac{\#\text{mistake}}{m} < \varepsilon\right]$$

$$= 2^m \left(1 - \Pr\left[\frac{\#\text{mistakes}}{m} < \frac{1}{2} - (\tfrac{1}{2} - \varepsilon)\right]\right)$$

$$> 2^m \left(1 - e^{(-2m(\frac{1}{2}-\varepsilon)^2)}\right)$$

Hoeffding bound

$$\geq 2^{m/2} \cdot 0.99$$

$\varepsilon \leq \frac{1}{4}$ $\quad m \geq 40$

$\Rightarrow$ 0.99 % of the promising concept are bad!

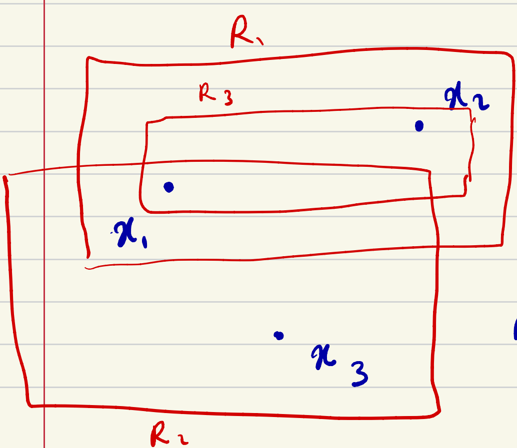
Def. Restriction of C to S

Let S be a set of m points in domain X. $S = \{x_1, \dots, x_m\}$

The restriction of C to S is the set of functions from S to $\{0,1\}$ that can be derived from C.

$$C_S : \{ (c(x_1), c(x_2), \dots, c(x_m)) \mid c \in C \}$$

where we represent each function from S to $\{0,1\}$ as a vector in $\{0,1\}^{|S|}$ or $\{0,1\}^m$



$C = \{R_1, R_2, R_3\}$

assign positive label to points inside the rectangle

$$\text{Restrictions}: \left\{ \begin{array}{ccc} x_1 & x_2 & x_3 \\ (+, & +, & -) \\ (+, & -, & +) \end{array} \right.$$

while C might have infinitely many hypotheses, its "effective size" is small

def. growth function

Let C be a concept class. Then, the growth function of C, denoted $\tau_C : \mathbb{N} \to \mathbb{N}$, is defined as:

$$\tau_C(m) = \max_{S \subset X : |S| = m} |C_S|$$

$\tau_C(m) \approx$ number of functions from S to $\{0,1\}^m$ that can be obtained by $c \in C$.

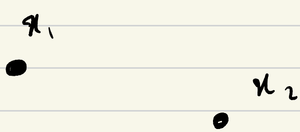
- With no assumption, we know $|C_S|$ is bounded by $2^{|S|} = 2^m$

def. shattering

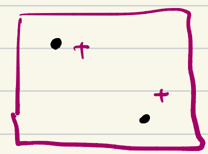
A class C shatters a finite set S if the restriction of C to S is the set of all functions from C to $\{0, 1\}$. That is $|C_S| = 2^{|S|} = 2^m$

Example $\quad C$ = axis-aligned rectangles

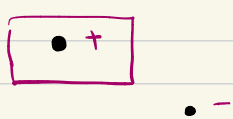
S



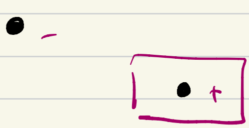
$(+, +)$



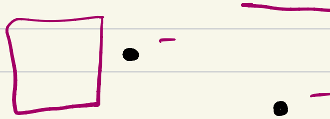
$(+, -)$



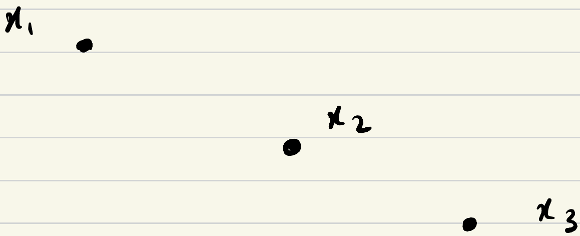
$(-, +)$



$(-, -)$



How about 3 points?



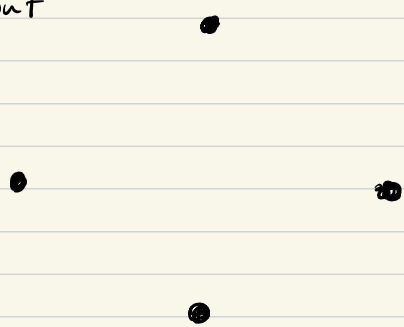
Can you label them with

$(+, -, +)$

C does not shatter this S.

How about

4 points?



---

what we have shown earlier indicates:

if C shatters S, we <u>cannot</u>

learn with $|S|/2 = m/2$ samples.

Def. VC Dimension

The **VC dimension** of a concept class C, denoted by $VCdim(C)$, is the maximal size of a set S that can be shattered by C.

If C can shatter sets of arbitrary large size, we say $VCdim(C) = \infty$

Example 1:

$VCdim$ (Axis-aligned rectangle) $= 4$

We need to show:

- there is a set of size 4 that is shattered.
- No set of size 5 is shattered.

Example 2: finite classes:

$$|C_S| \leq |C| = 2^{\log |C|}$$

C cannot shatter any set of size larger than $\log |C|$

$$VCdim(|C|) \leq \log |C|$$

If $VCdim(C) = d$

$$\forall\, m \leq d \implies \tau_C(m) \leq 2^m$$

$$\forall\, m > d \implies \tau_C(m) < 2^m$$

VC dimension

- infinite classes can still be PAC-learnable.

$\Rightarrow$ size is <u>not</u> determinant of learnability.

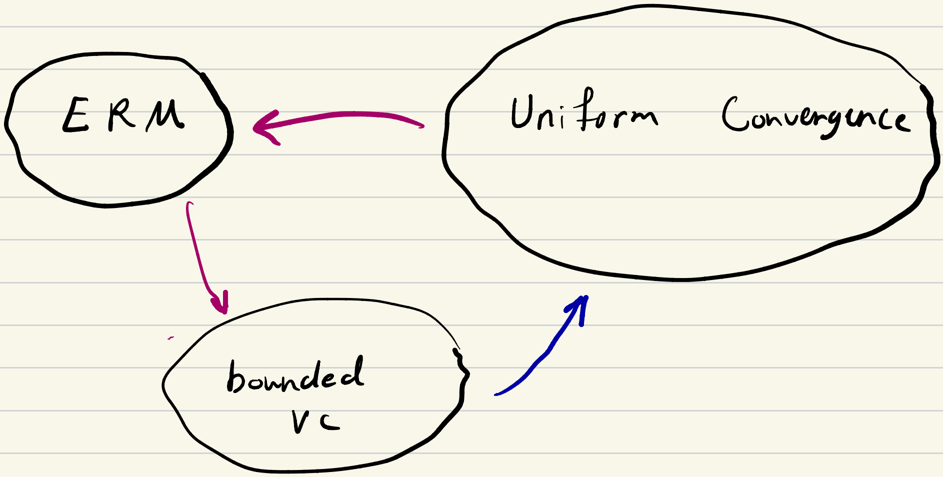
So, what is then?

VC-dim of C characterizes its learnability!

The fundamental theorem of PAC learning

For a concept class C of $c: X \to \{-1, +1\}$ with 0-1 loss function, the following are equivalent:

- C has uniform convergence.
- Any ERM is a successful agnostic PAC learner
- H has a finite VC dim.


ERM
Uniform Convergence
bounded
VC