Lecture 9 Oct 18, 2023

Today's goals:

- VC Dimension

Last lecture :

Recall:

+ 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:

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

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

Suppose we have a set of m points

There are 2^m possible labelings of these m points.

Suppose C is the class of 2^m 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 m points with their label.

$T \leftarrow$ Draw $m/2$ samples from D

(WLOG assume they are unique

How many function in C label T correctly? $2^{m/2}$

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

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

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

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

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

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

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

a random concept in P

c makes $\lceil m.\varepsilon \rceil$ mistakes

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

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

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

Hoeffding bound $\quad \geq 2^{m/2} \cdot 0.99$

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

$\Rightarrow$ 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. $S = \{x_1, \ldots, x_m\}$

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

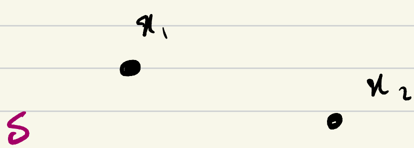
$$C_S : \{(c(x_1), c(x_2), \ldots, c(x_m)) \mid c \in C\}$$

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

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

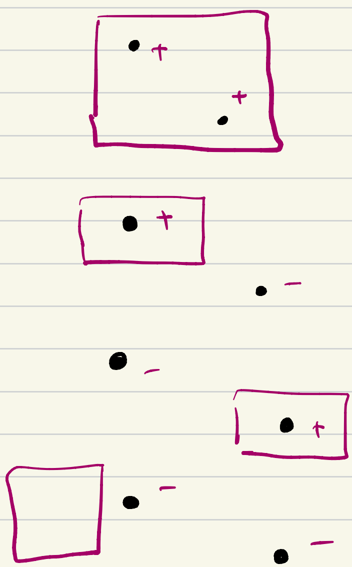


$(+, +)$

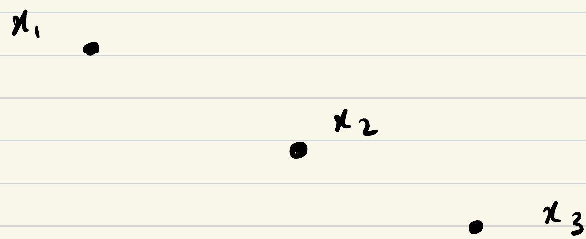
$(+, -)$

$(-, +)$

$(-, -)$



How about 3 points?

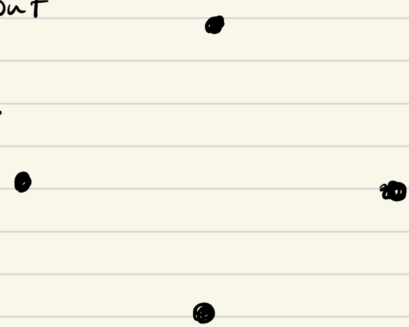


Can you label them with

$(+, -, +)$

C does not shatter this S.

How about 4 points?



---

what we have shown earlier indicates:

if C shatters S, we cannot 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 $VCdim(C) = \infty$

VC dim (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.

finite classes:

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

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

$$\text{VC dim}(|C|) \leq \log |C|$$

The fundamental theorem of PAC learning

For class of concepts $X \to \{0,1\}$ with 0-1 loss function, the following are equivalent:

- C has uniform convergence.

last time

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

$\sim P \Rightarrow \sim q \iff q \Rightarrow P$

Roughly speaking:
what we have shown earlier today says

If ERM works with m sample
$VCdim(C) < 2m$

what have left to show is :

finite VCdim $\Rightarrow$ Uniform convergence.

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

as the number of samples increases the size of the restriction of C to S (the sample set) grows polynomially not exponentially $(2^{|S|})$.

def. growth function

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

the number of functions that we can have by restricting C to S of size m.

$VC\dim(C) = d$

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

Sauer-Shelah-Perles Lemma

If $VC\dim(C) \le d < \infty$, then

$$\forall\, m \quad \tau_C(m) \le \sum_{i=0}^{d} \binom{m}{i}$$

In particular, if $m > d+1$,

$$\tau_C(m) \le \left(\frac{em}{d}\right)^d$$

This is much better than what we naively can imply from the definition

$$\tau_C(m) < 2^m$$