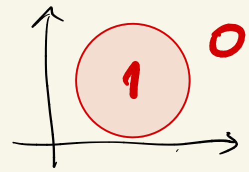
LECTURE 26

Examples continued:

5. Circles in $\mathbb{R}^2$:

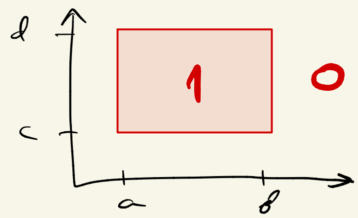
$$\mathcal{H} = \left\{ \mathbb{1}_{\{(x(1)-a_1)^2 + (x(2)-a_2)^2 \leq r^2\}} : a_1, a_2, r \in \mathbb{R} \right\}$$



$\boxed{vc(\mathcal{H}) = 3}$ HW

6. Axis-aligned rectangles in $\mathbb{R}^2$:

$$\mathcal{H} = \left\{ \mathbb{1}_{[a,b] \times [c,d]} : a<b, c<d \right\}$$



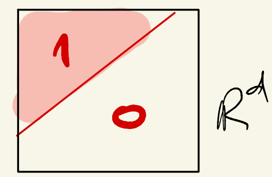
$\boxed{vc(\mathcal{H}) = 4}$ HW

Generalizing Example 4 from last class:

7. Half-spaces in $\mathbb{R}^d$:

$$\mathcal{H} = \left\{ \mathbb{1}_{\{\langle w, x \rangle + b \geq 0\}} : w \in \mathbb{R}^d, b \in \mathbb{R} \right\}$$

"linear classifier" (SVM)



$\boxed{vc(\mathcal{H}) = d+1}$ HW

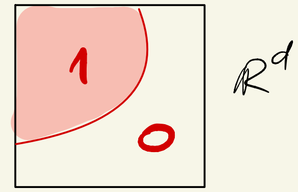
More generally:

8. Polynomial surfaces in $\mathbb{R}^d$: $\mathcal{H} = \left\{ \mathbb{1}_{\{p(x) \geq 0\}} : \deg(p) \leq r \right\}$

$$vc(\mathcal{H}) = \binom{d+r}{r}$$

[Anthony 1995]

"Polynomial classifier"

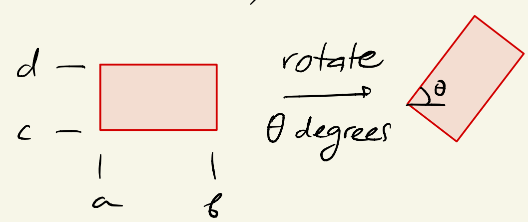


Remark. In all examples above, $vc(\mathcal{H}) = \#$ parameters that describe a function in $\mathcal{H}$.

- This is not true in general. For rectangles in $\mathbb{R}^2$ that are not necessarily axis-aligned as in Ex. 5,

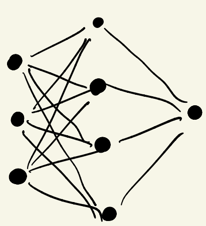
$$vc(\mathcal{H}) = 7$$

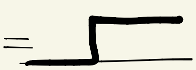
while the # parameters is 5: a, b, c, d, θ.



- But heuristically, and "approximately", this is often true:

8. $\mathcal{H} = \{$functions a given neural network can compute$\}$,



if activation function = ,

then $\boxed{vc(\mathcal{H}) \le C \underbrace{W}_{\#\text{ connections}} \log W}$ [Cover 69; Baum-Haussler 89]

connections (= # weights, parameters)

- ∃ networks for which this bound is tight [Maas 94]
- If activation function is piecewise linear (e.g. ReLU),

then $\boxed{vc(\mathcal{H}) \le C W \underbrace{L}_{\#\text{ layers}} \log W}$ [Bartlett-Harvey-Liaw-Mehrabian 2017]

layers

and ∃ examples showing tightness.

Lem [Pajor 85] $\forall$ finite class of Boolean functions $\mathcal{H}$ on $\mathcal{X}$,

$$|\mathcal{H}| \le \#(\text{subsets of } \mathcal{X} \text{ shattered by } \mathcal{H})$$

Convention: $\emptyset$ is shattered by $\forall$ nonempty $\mathcal{H}$.

Proof WLOG, $\mathcal{X} = \{1, \dots, n\}$. Denote by $sh(\mathcal{H})$ the family of all subsets of $\mathcal{X}$ shattered by $\mathcal{H}$. To prove

$$|\mathcal{H}| \le |sh(\mathcal{H})|,$$

- partition $\mathcal{H}$ according to the value at point n, i.e.

$$\mathcal{H} = \mathcal{H}_0 \sqcup \mathcal{H}_1$$

where $\mathcal{H}_0 = \{h \in \mathcal{H} : h(n) = 0\}$ and $\mathcal{H}_1 = \{h \in \mathcal{H} : h(n) = 1\}$.

- $\forall$ subset $\{i_1, \dots, i_d\} \subset \mathcal{X}$ shattered by $\mathcal{H}_0$ or $\mathcal{H}_1$ is also shattered by $\mathcal{H}$. Thus

(subclasses of $\mathcal{H}$)

$$|sh(\mathcal{H})| \ge |sh(\mathcal{H}_0)| + |sh(\mathcal{H}_1)| \quad \ge \qquad (*)$$

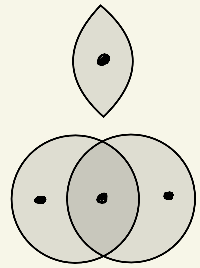
domain $= \{1, \dots, n-1\}$

- Iterate: partition $\mathcal{H}_0$ and $\mathcal{H}_1$ according to the value $h(n-1)$:

$$\ge |sh(\mathcal{H}_{00})| + |sh(\mathcal{H}_{01})| + |sh(\mathcal{H}_{10})| + |sh(\mathcal{H}_{11})| \ge$$

... down to single-point classes, each of which shatters one set $\emptyset$ ... $\ge |\mathcal{H}|$

· MISTAKE: we double counted in (*) the sets that are shattered by both $\mathcal{H}_0$ and $\mathcal{H}_1$

$sh(\mathcal{H}_0)$  $sh(\mathcal{H}_1)$

· FIX: suppose $\{i_1, \dots, i_d\}$ is shattered by both $\mathcal{H}_0$ and $\mathcal{H}_1$ $\Rightarrow$

$\forall$ label assignment $y_1, \dots, y_d \in \{0,1\}$

$\exists h \in \mathcal{H}_0$: $h(i_1) = y_{i_1}, \dots, h(i_d) = y_{i_d}$, $h(n) = 0$

$\exists g \in \mathcal{H}_1$: $g(i_1) = y_{i_1}, \dots, g(i_d) = y_{i_d}$, $g(n) = 1$

$\Rightarrow \underbrace{\{i_1, \dots, i_d, n\}}$ is shattered by $\mathcal{H} = \mathcal{H}_0 \cup \mathcal{H}_1$ (choose either h or g)

This set is NOT shattered by either $\mathcal{H}_0$ or $\mathcal{H}_1$ $\Rightarrow$ it was NOT counted before

· $\Rightarrow$ $\forall$ set that we double counted, we find a set we never counted

$\Rightarrow$ (*) is true. Proceed as before. QED

- By def of vc dimension, $\forall$ subset shattered by $\mathcal{H}$ has cardinality $\le \mathrm{vc}(\mathcal{H}) =: d$. So Pajor's lemma yields

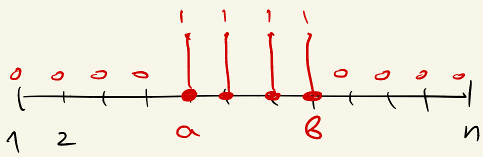
$$|\mathcal{H}| \le \#(\text{subsets of } \{1,\dots,n\} \text{ with card.} \le d) \le \sum_{k=0}^{d} \binom{n}{k}.$$

$$\Downarrow$$

Cor (Sauer-Shelah lemma) Let $\mathcal{H}$ be a class of Boolean functions on an n-point domain. Then

$$|\mathcal{H}| \le \sum_{k=0}^{d} \binom{n}{k} \quad \text{where} \quad d = \mathrm{vc}(\mathcal{H})$$

Examples (skip?)

(a) Integer intervals: $\mathcal{H} = \Big\{$ 

$: 1 \le a \le b \le n \Big\}$

$\mathrm{vc}(\mathcal{H}) = 2$ (as in the previous lecture for real intervals)

$$|\mathcal{H}| = \underset{\uparrow}{1} + n + \binom{n}{2} = \sum_{k=0}^{2} \binom{n}{k} \Rightarrow \text{Pajor lemma is sharp}$$

zero function; $a=b$; # pairs $(a<b)$

(b) $\mathcal{H} = \{$all functions on an n-point domain supported by $\le d$ pts$\}$

$\mathrm{vc}(\mathcal{H}) = d$ **(HW?)** and $|\mathcal{H}| = \sum_{k=0}^{d} \binom{n}{k} \Rightarrow$ sharp again!

Remarks ① $\displaystyle\sum_{k=1}^{d} \binom{n}{k} \le \left(\frac{en}{d}\right)^d$ (HW 3, Problem 3) $\xRightarrow{\text{Pajor}}$

$$d \le \boxed{\log|\mathcal{H}| \le d \log\left(\frac{en}{d}\right), \text{ where } d = \mathrm{vc}(\mathcal{H}).}$$

HW 13

② Heuristically, $\log|\mathcal{H}| = \#$bits to specify a function in $\mathcal{H}$

$d = \mathrm{vc}(\mathcal{F}) \sim \#$parameters that describe functions in $\mathcal{H}$

$\Rightarrow \log|\mathcal{H}| \asymp d$, is expected.