LECTURE 35

SUMMARY of math framework of ML (Last class)

- $\mathcal{P}$: an unknown distribution on $\mathcal{X} \times \mathcal{Y}$; ($\forall$ sets)
- We see training data: $(X_1, Y_1), \dots, (X_n, Y_n) \sim \mathcal{P}$ iid. (label) Goal: oracle $\mathcal{X} \to \mathcal{Y}$
- Choose a hypothesis class $\mathcal{H}$ (functions $\mathcal{X} \to \mathcal{Y}$)
- $\forall h \in \mathcal{H}$, Risk, a.k.a. "test error":

loss function, e.g. $\mathbb{E}(h(X)-Y)^2$ (quadratic loss)

$$R(h) := \mathbb{E}\, \ell(h(X), Y) \qquad h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, R(h).$$

Not computable

- Empirical risk a.k.a. training error
- $R_n(h) := \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i)$. $h_n^* := \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, R_n(h)$. Computable.

- ERM algorithm:

① Training: for input data $(X_1, Y_1), \dots, (X_n, Y_n)$; compute h_n^*.

② Prediction: on query X, output $h_n^*(X)$ "oracle"

Lem (Generalization error)

$$\boxed{R(h_n^*) \le R(h^*) + 2 \sup_{h \in \mathcal{H}} |R_n(h) - R(h)|}$$

test error of ERM; best possible error (with ∞ data)

$$= 2 \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \underbrace{\left(\ell(h(X_i), Y_i) - \mathbb{E}\,\ell(h(X_i), Y_i) \right)}_{= Z_i(h) \text{ iid mean 0 r.v.'s}} \right| = 2 \sup_{h \in \mathcal{H}} \underbrace{\left| \frac{1}{n} \sum_{i=1}^{n} Z_i(h) \right|}$$

= stochastic process, called "empirical process"

• For <u>binary</u> classification, $\ell(\cdot,\cdot)\in\{0,1\} \implies |Z_i(h)|\le 1$

• $$P\Big\{ \Big|\frac{1}{n}\sum_{i=1}^{n} Z_i(h)\Big| > t \Big\} = P\Big\{ \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \underline{Z_i(h)}\Big| > t\sqrt{n} \Big\} \le 2\exp\Big(-\frac{(t\sqrt{n})^2}{2}\Big) \quad (*)$$

multiply both sides by $\sqrt{n}$ to scale like in CLT

iid mean 0, bdd by 1

Hoeffding's inequality (Lec. 7, Sep. 16)

• Union Bound:

$$P\Big\{ \underbrace{\sup_{h\in\mathcal{H}}}_{\text{"}\exists h\in\mathcal{H}\text{" = union}} \Big|\frac{1}{n}\sum_{i=1}^{n} Z_i(h)\Big| > t \Big\} \le \sum_{h\in\mathcal{H}} P\Big\{ \Big|\frac{1}{n}\sum_{i=1}^{n} Z_i(h)\Big| > t \Big\}$$

← assume $\mathcal{H}$ is finite

$$\overset{(*)}{\le} |\mathcal{H}|\cdot 2\exp(-t^2 n/2) = 2\exp\big(\log|\mathcal{H}| - t^2 n/2\big)$$

$$\le 0.01 \quad \text{if } t = C\sqrt{\frac{\log|\mathcal{H}|}{n}}.$$

⇒ We proved:

THM (Generalization Bound) If the hypothesis class $\mathcal{H}$ is finite,

$$\boxed{R(h_n^*) \le R(h^*) + \underbrace{C\sqrt{\frac{\log|\mathcal{H}|}{n}}}_{\le\, 0.01 \text{ if } n \ge C'\log|\mathcal{H}|} \qquad \text{with prob. } \ge 0.99.}$$

ERM's test error

best possible error

Hence the ERM algorithm generalizes well from

$$n \sim \log|\mathcal{H}|$$

training data points.

• Good: logarithmic in $|\mathcal{H}|$

• Bad: most hypothesis classes are <u>infinite</u>.

Can $\log|\mathcal{H}|$ be replaced by some "<u>complexity</u>" of an infinite $\mathcal{H}$?

Yes: VC dimension

VC DIMENSION (Вапник–Червоненкис)

- Heuristically: vc($\mathcal{H}$) = largest #(data $\mathcal{H}$ overfits)

i.e. functions $h : X \to \{0,1\}$

Def Let $\mathcal{H}$ be any collection of Boolean functions on a set X.
We say that $\mathcal{H}$ overfits, or "shatters" a subset $\{x_1, \dots, x_d\} \subset X$ if $\forall$ labels $y_1, \dots, y_d \in \{0,1\}$ $\exists\, h \in \mathcal{H}$ such that

$$h(x_i) = y_i \qquad \forall\, i = 1, \dots, d.$$

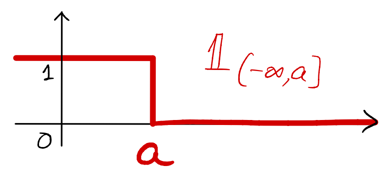
The vc dimension of $\mathcal{H}$, denoted vc($\mathcal{H}$), is the maximal size d of a subset $\mathcal{H}$ shatters.

Examples

$h(x) = 1 \ \forall x$

1. $\mathcal{H} = \{\mathbb{1}\}$ has $\boxed{vc(\mathcal{H}) = 0}$: it can't shatter even one point x_i since $h(x_i) = 1$.

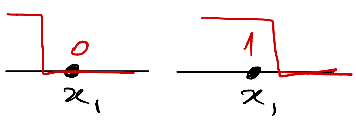
2. Half-lines $\mathcal{H} = \{\mathbb{1}_{(-\infty, a]} : a \in \mathbb{R}\}$

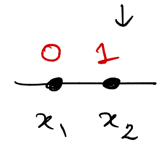


$\boxed{vc(\mathcal{H}) = 1}$ Proof:

HW: $\{\mathbb{1}_{(-\infty,a]} ; \mathbb{1}_{[b,+\infty)}\}$

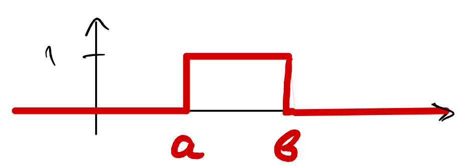
$\mathcal{H}$ can shatter some 1-point set $\{x_1\}$, but can't shatter any 2-point set $\{x_1, x_2\}$





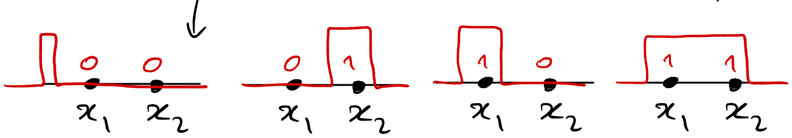
← this label assignment is impossible to realize

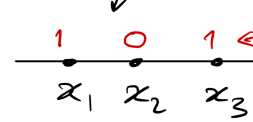
3. Intervals: $\mathcal{H} = \{\mathbb{1}_{[a,b]} : a \le b\}$



$\boxed{vc(\mathcal{H}) = 2}$ Proof:

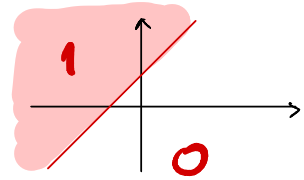
$\mathcal{H}$ can shatter some 2-pt set $\{x_1, x_2\}$, but can't shatter any 3-point set $\{x_1, x_2, x_3\}$





← impossible to realize

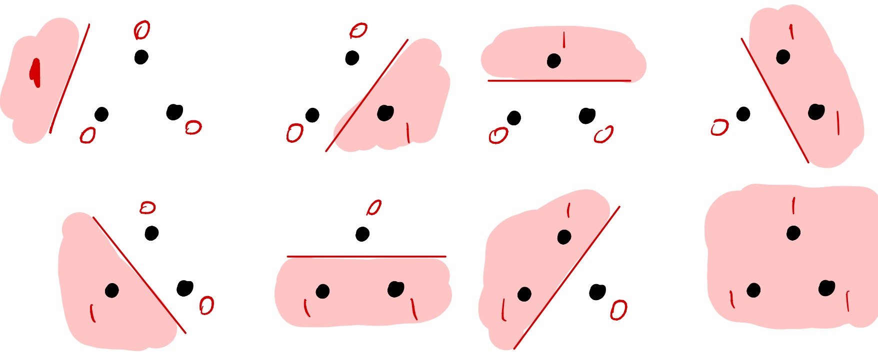
4. Half-planes in $\mathbb{R}^2$:



$$\mathcal{H} = \left\{ \mathbb{1}_{\{a_1 x(1) + a_2 x(2) + b \geq 0\}} : a_1, a_2, b \in \mathbb{R} \right\}$$

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

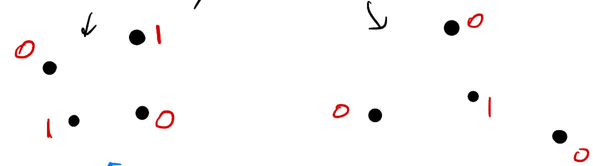
$\mathcal{H}$ can shatter some 3-pt set $\{x_1, x_2, x_3\}$,



but can't shatter any 4-point set $\{x_1, x_2, x_3, x_4\}$:

$\forall$ 4-point set is

like this, or like this

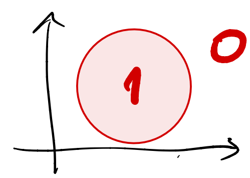


"convex position"

In either case, $\exists$ label assignment that is impossible to realize

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

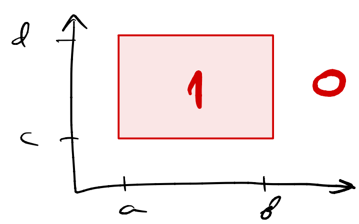
$$\mathcal{H} = \left\{ \mathbb{1}_{\{(x(1)-a_1)^2 + (x(2)-a_2)^2 \leq r^2\}} : a_1, a_2, r \in \mathbb{R} \right\}$$



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

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

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



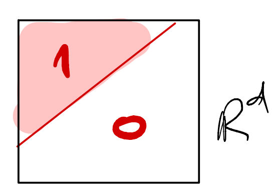
$\boxed{vc(\mathcal{H}) = 4}$

Generalizing Example 4:

6. 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)



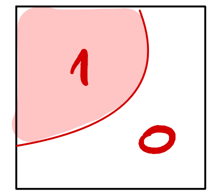
$\boxed{vc(\mathcal{H}) = d+1}$

More generally:

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

$$vc(\mathcal{H}) = \binom{d+r}{r}$$ [Anthony 1995]

"Polynomial classifier"



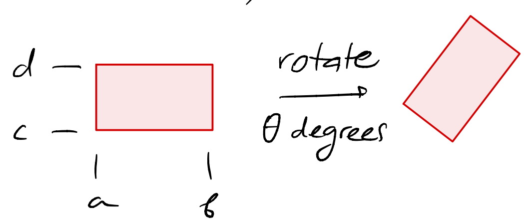
$\mathbb{R}^d$

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 $\mathbb{R}^2$, not necessarily axis-aligned as in Ex. 5,

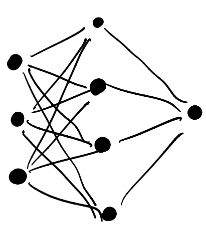
  $$vc(\mathcal{H}) = 7$$

  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 = 

  network computes a composition of linear classifiers.

  Here $\boxed{vc(\mathcal{H}) \leq C\, \underbrace{W}_{\#\text{ connections } (= \#\text{ weights, parameters})} \log W}$ [Cover 69; Baum–Haussler 89]

- $\exists$ networks for which this bound is tight [Maas 94]
- If activation function is piecewise linear (e.g. ReLU) $\Rightarrow$

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

  and $\exists$ examples showing tightness.