Lecture 14. Complexity and generalization (VC dimension)

Consider a simple decision stump.

$$h(\underline{x};\underline{\theta}) = \text{sign}(s(x_k - \theta_0)) \qquad k \in \{1,2,\dots,d\} \quad s \in \{\pm 1\} \quad \theta_0 \in \mathbb{R}$$

$\boxed{(s, k, \theta_0) \to \text{three paras}}$

How simple it is?

$$\mathcal{F} = \{h(\cdot;\underline{\theta}) : \theta = \{s, k, \theta_0\} \in \{\pm1\} \times \{1,\dots,d\} \times \mathbb{R}\}$$

function class

(hypothesis)

Recall: If $|\mathcal{F}| < +\infty$, then with prob $\leq 1-\delta$,

$$\underbrace{R(f)}_{\text{Exp. risk}} \leq \underbrace{R_n(f)}_{\text{Emp. Risk}} + C(n, \mathcal{F}, \delta)$$

$R_n(f) - R(f)$

generalization error

$$\text{where} \quad C(n,\mathcal{F},\delta) = \sqrt{\frac{1}{2n} \log \frac{2|\mathcal{F}|}{\delta}} \to 0 \quad \text{as } n \to \infty$$

Obviously, this bound doesn't apply for decison stumps $\mathcal{F}$.

(Since $|\mathcal{F}| = +\infty$)

Motivating Example 1: $d=1$ (1-dim data) binary labels $y_t \in \{\pm 1\}$

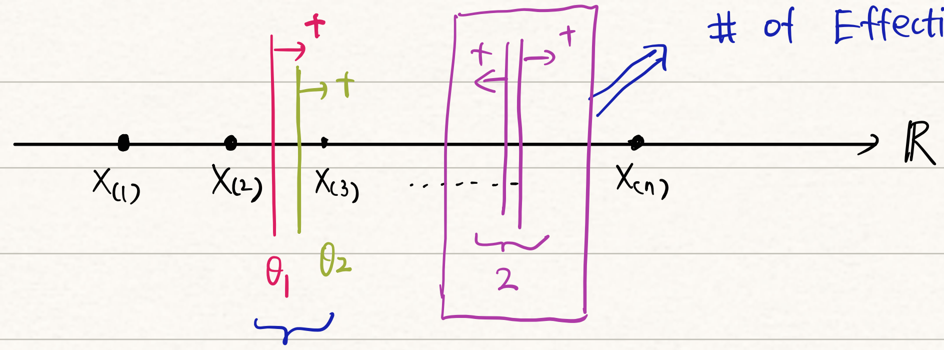
$\mathcal{F} \to$ class of threshold classifiers / decision stump in 1 dim

$\Rightarrow \forall f \in \mathcal{F}, \exists \theta_0 \in \mathbb{R}$ s.t ① $f(x) = 1$ for $x \leq \theta_0$, $f(x) = -1$ for $x > \theta_0$

Or ② $f(x) = -1$ for $x \le \theta_0$, $f(x) = 1$ for $x > \theta_0$

Training sets of points $x_1, \dots, x_n \in \mathbb{R}$

Record $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$



\# of Effective Classifiers :

2n+2
↓
2n label

These two classifiers always give the SAME decision

Thus, for above dataset, there are finitely many Effective classifiers

2n+2

d-dimension decision stump   effective # of labellings

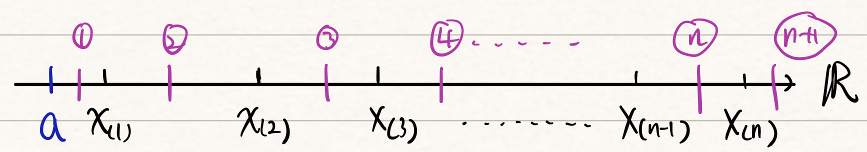
$\le d \cdot 2n$

Motivating Ex. 2 :

$\mathcal{F}$: interval classifier

$\forall f \in \mathcal{F}$, $\exists a, b \in \mathbb{R}$ s.t. $f(x) = +1$ $\forall x \in [a,b]$,

$f(x) = -1$ else



n+1 location for b regarding to a.

$\frac{(n+1)n}{2} + 2$

$\rightarrow$ # of Effective Classifiers $= \frac{(n+2)(n+1)}{2}$. $< \infty$

How can we quantify the "richness" of $\mathcal{F}$?

can be applied to $|\mathcal{F}| < \infty$

Thm: For $|\mathcal{F}| = \infty$, with prob $\geq 1-\delta$, (Proof in lec9.pdf)

$$R(f) \leq \hat{R}_n(f) + \sqrt{\frac{d_{vc} \log \frac{2ne}{d_{vc}} - \log \frac{\delta}{4}}{n}}$$

$$= \hat{R}_n(f) + \tilde{O}\left(\sqrt{\frac{d_{vc}}{n}}\right)$$

And $d_{vc} = d_{vc}(\mathcal{F}) \rightarrow$ is the VC-dimension of $\mathcal{F}$.

[Rmk:] if $|\mathcal{F}| < +\infty$, then $d_{vc}(\mathcal{F}) = \log|\mathcal{F}|$. (tutorial)

measure of "richness"

Defn: $\mathcal{F}$ can shatter points $x_1, x_2, \dots, x_n$

if $\forall$ labellings $(y_1, \dots, y_n) \in \{\pm 1\}^n$, $\exists f \in \mathcal{F}$ that can be achieve 0 training error on $\{(x_t, y_t)\}_{t=1}^n$

$\Downarrow$

$f(x_t) = y_t \quad \forall t = 1, 2, \dots, n$

or $VC(\mathcal{F})$

Defn: The VC-dimension of $\mathcal{F}$, denoted as $d_{vc} = d_{vc}(\mathcal{F})$ is the maximum # of points that can be ARRANGED such that $\mathcal{F}$ can shatter them.

clever way! $\Rightarrow$ choose one way

Calculate d_{vc}

$\Leftrightarrow$ Play this Game

Thought Experiment. Fix $n \in \mathbb{N}$

- Player 1 select points $x_1, \dots, x_n$ 存在!

- Player 2. select labels $(y_1, \dots, y_h) \in \{\pm 1\}^h$ 任意!

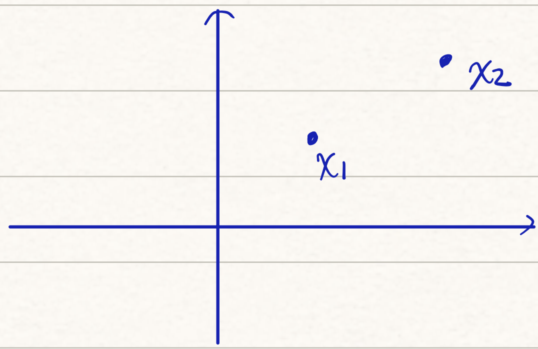
- Player 1 tries to select $f \in \mathcal{F}$ s.t $f(x_t) = y_t \ \forall t$
- If Player 1 succeeds no matter how P2 plays.

then $VC(\mathcal{F}) \geq h$

E.g. 1: [linear classifier in 2D]

$$\mathcal{F} = \{ f(\underline{x}; \underline{\theta}) = \text{sign}(\theta_0 + \theta_1 x_1 + \theta_2 x_2) : \underline{\theta} = (\theta_0, \theta_1, \theta_2) \in \mathbb{R}^3 \}$$

Pf: $d_{VC} \geq 2$.



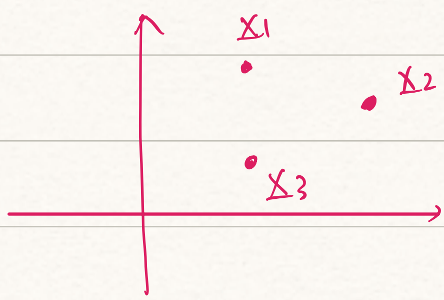
①

- $\boxed{\text{set } h = 2}$
- P1 set x_1, x_2
- P2 set $(y_1, y_2) \in \{\pm 1\}^2$
- P1 can $\boxed{\text{always}}$ find a linear classifier $f(\underline{x}; \underline{\theta})$ to achieve 0 training error

for 4 different labels!

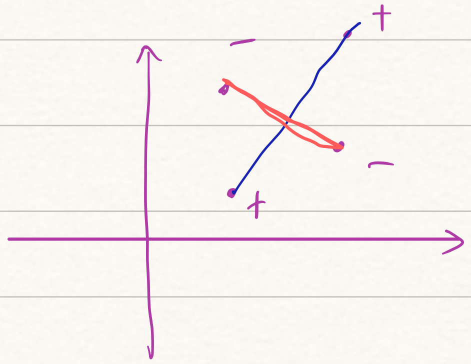
Hence, $VC(\mathcal{F}) \geq 2$

Claim: $d_{VC}(\mathcal{F}) \geq 3$.



$\Rightarrow$ Repeat the procedure!

Claim : $d_{vc}(\mathcal{F}) < 4$

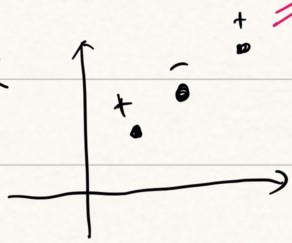


$\Rightarrow$ it is impossible to find a Linear Classifier f to shatter the 4 points !

Conclusion : d_{vc}(Linear Classifier in $\mathbb{R}^2$) $= 3$

Rmk: Important Point

$d_{vc}(\mathcal{F}) \geq 3$



$\Rightarrow$ this case

$\Rightarrow$ It seems that we cannot shatter these 3 Points and P1 fails.

$\downarrow$

But that there exists a layout of 3 points that cannot be shattered by linear Classifiers is IRRELEVANT !

我们只要找一组"存在"的解 $\underline{x}_1, \underline{x}_2, \underline{x}_3$

$\Downarrow$

☆ 证明 $D_{vc} \geq n$

$\Downarrow$

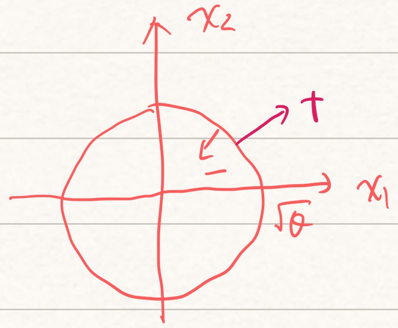
让 P1 clever enough to find one possible selection of $x_1, \dots, x_n$ s.t P1 can win !

Ex 2. $\mathcal{F} = \{ f(\underline{x}; \theta) = \text{sign}(\underline{x}^T \underline{x} - \theta) ; \theta \in \mathbb{R}_+ \}$ $\quad \underline{x} \in \mathbb{R}^2$

i.e. if $\|x\|^2 > \theta$, $f(x;\theta) = 1$

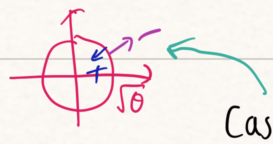
if $\|x\|^2 \le \theta$, $f(x;\theta) = -1$.

Claim: $d_{VC}(\mathcal{F}) = 1$. $\Rightarrow$ obvious ! $\begin{cases} d_{VC}(\mathcal{F}) \ge 1 \\ d_{VC}(\mathcal{F}) < 2. \end{cases}$

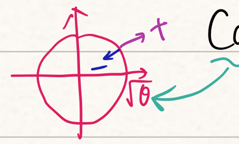


Ex 2' $\mathcal{F} = \{ f(x;\theta) = \text{sign}(s[x^T x - \theta]) : \theta \in \mathbb{R}^+, s \in \{\pm 1\} \}$

$\Downarrow$



Case ① if $\|x\|^2 \le \theta$, $f(x;\theta) = 1$ $(s = -1)$



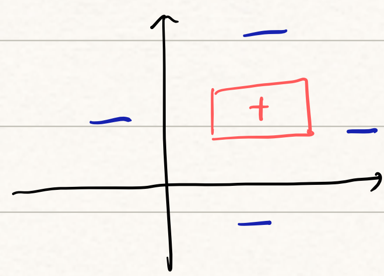
Case ② if $\|x\|^2 \le \theta$, $f(x;\theta) = -1$. $(s = +1)$

Claim: $d_{VC}(\mathcal{F}) = 2$ $\begin{cases} d_{VC}(\mathcal{F}) \ge 2 \\ d_{VC}(\mathcal{F}) < 3 \end{cases}$

Example : [Axis - Aligned Rectangle]

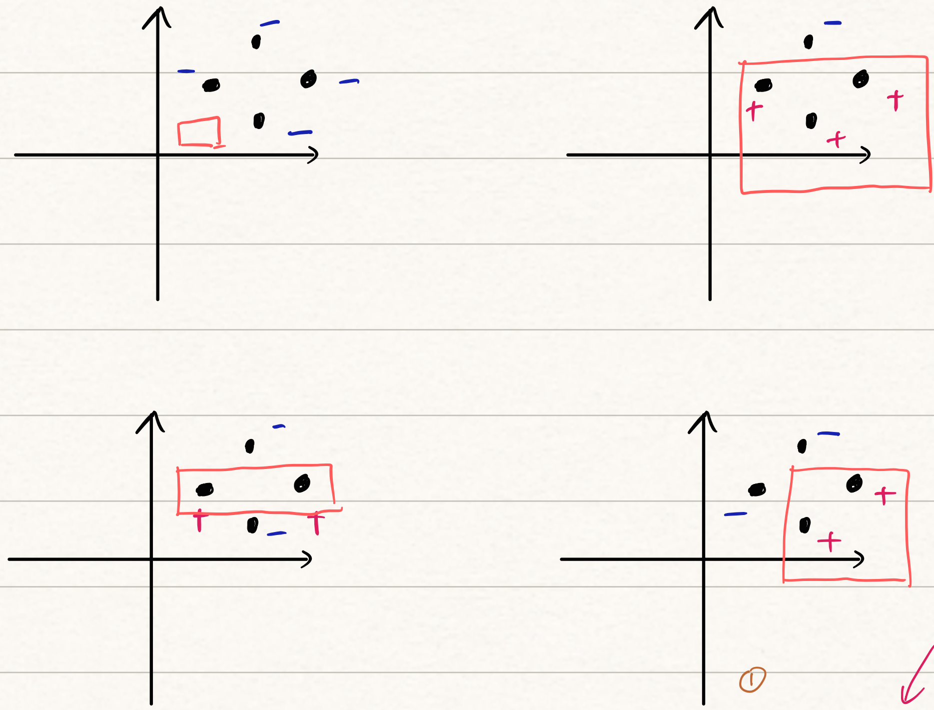
$x \in \mathbb{R}^2$ $\rightarrow$ $\mathcal{F}$ = { class of f's that assign "+" if a point

lies in an axis-aligned rectangle

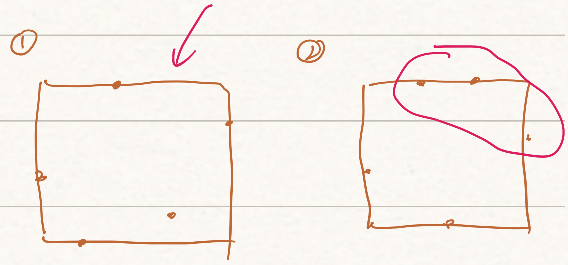


Claim: $d_{vc}(\mathcal{H}) = 4$

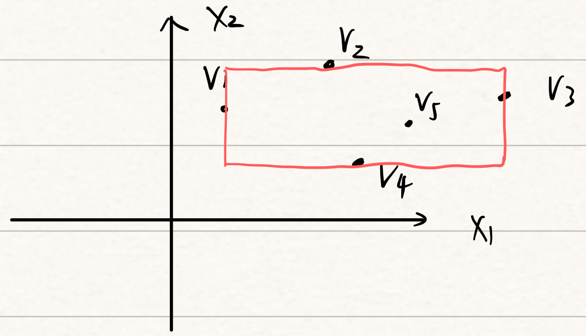
① $d_{vc}(\mathcal{H}) \geq 4$



包括边界 ≤3 case



② $d_{vc}(\mathcal{H}) < 5$ (Difficult)



$T = \{v_1, v_2, v_3, v_4, v_5\}$

$T = v_1 \sim v_5$

Consider ANY 5 distinct Points, named $v_1 \sim v_5$

Consider the rectangle with

① min x_1 coord.

② max x_1 coord.

③ min x_2 coord.

④ max x_2 coord.

Here $S = \{v_1, v_2, v_3, v_4\}$

Call the set of 4 Points are on the Boundary of Reetangle S Here is $\{v_1, v_2, v_3, v_4\}$

Construction [Assign label '+' to Points in S
and label '−' to Points in $T \setminus S$