LECTURE 25

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: $(X_1, Y_1), \ldots, (X_n, Y_n) \sim \mathcal{P}$ iid. (label) Goal: oracle $h: \mathcal{X} \to \mathcal{Y}$: $h(X) \approx Y$
- 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
- $R_n(h) := \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i)$. $\qquad h_n^* := \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, R_n(h)$. Computable.

- ERM algorithm:

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

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

Lem (Generalization error)

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

test error of ERM

best possible error (with ∞ data) for a given class $\mathcal{H}$

Proof
$$\begin{aligned}
R(h_n^*) &\le R_n(h_n^*) + \varepsilon && (h_n^* \in \mathcal{H}) \\
&\le R_n(h^*) + \varepsilon && (h^* = \text{minimizer of } R_n) \\
&\le R(h^*) + 2\varepsilon && (h^* \in \mathcal{H}) \qquad \square.
\end{aligned}$$

Next time, add a lemma:

if $\|f-g\|_\infty < \varepsilon$ and x^*, y^* are minimizers of f, g, then $|f(x^*) - g(y^*)| < 2\varepsilon$.

Then apply this lemma for $f = R_n$, $g = R$, h_n^*, h^*

$$\varepsilon = \sup_{h\in\mathcal{H}} \Big| \underbrace{\frac{1}{n}\sum_{i=1}^{n} Z_i(h) - \mathbb{E}Z(h)}_{\text{"empirical process"}} \Big| \quad \text{where } Z_i(h) = \ell(h(x_i), y_i) \text{ are iid r.v's } \forall h.$$

$$\| \quad \bar{Z}_i(h)$$

- For binary classification, $\ell(\cdot,\cdot) \in \{0,1\} \implies |Z_i(h)| \le 1$

- $$\mathbb{P}\Big\{ \Big|\frac{1}{n}\sum_{i=1}^{n} \bar{Z}_i(h)\Big| > t \Big\} = \mathbb{P}\Big\{\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \underbrace{\bar{Z}_i(h)}\Big| > t\sqrt{n}\Big\} \le 2\exp\Big(-\frac{(t\sqrt{n})^2}{2}\Big) \quad (*)$$

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

iid mean 0, bdd by 1

General Hoeffding inequality (lec. 5)

- Union bound:

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

$\leftarrow$ assume $\mathcal{H}$ is finite

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

$\implies$ We proved:

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

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

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

$R(h_n^*)$: ERM's test error

$R(h^*)$: best possible error

Hence: the ERM algorithm generalizes well from

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

training data points.

- Good: logarithmic in $|\mathcal{H}|$
- Bad: most hypothesis classes are infinite.

Can $\log|\mathcal{H}|$ be replaced by some "complexity" of $\mathcal{H}$?

Yes: VC dimension.

VC DIMENSION (Вапник-Червоненкис)

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

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

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

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

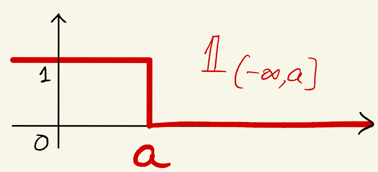
The vc dimension of $\mathcal{H}$, denoted $vc(\mathcal{H})$, is the maximal size d of a subset $\mathcal{H}$ shatters.

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 $h(x_i) = 1$.

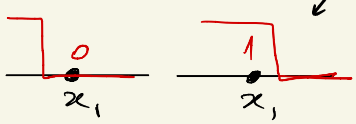
2. Half-lines $\mathcal{H} = \{\mathbb{1}_{(-\infty, a]} : a \in \mathbb{R}\}$

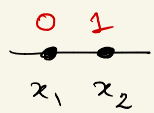


$\boxed{vc(\mathcal{H}) = 1}$

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

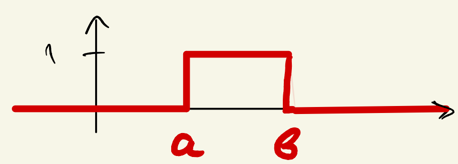
Proof: $\mathcal{H}$ can shatter some 1-point set $\{x_1\}$, but can't shatter any 2-point set $\{x_1, x_2\}$





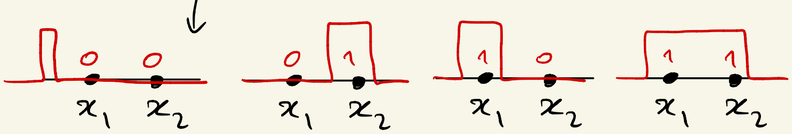
← this label assignment is impossible to realize

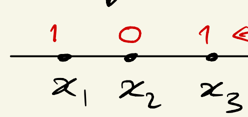
3. Intervals: $\mathcal{H} = \{\mathbb{1}_{[a,b]} : a \le b\}$



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

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

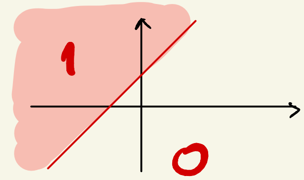




← impossible to realize

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

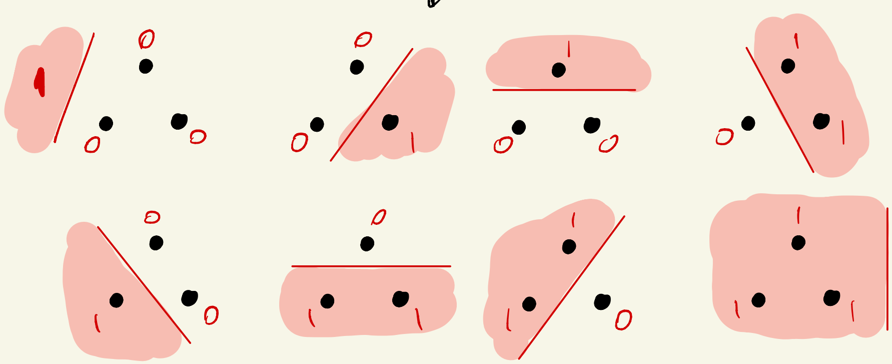
$$\mathcal{H} = \left\{ \mathbb{1}_{\{a_1 x(1) + a_2 x(2) + b \geq 0\}} : a_1, a_2, b \in \mathbb{R} \right\}$$



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

Proof

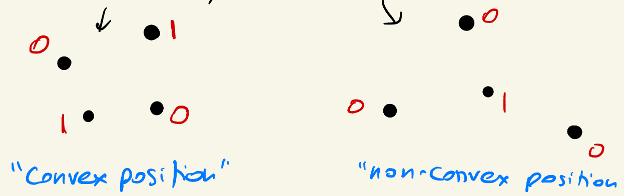
$\mathcal{H}$ can shatter some 3-pt set $\{x_1, x_2, x_3\}$,



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



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