VC classes ⇒ unif concentration from emp. error to generalization error

Q1: $\mathbb{E}_{S\sim D^m,\ S'\sim D^m} \sup_{f\in\mathcal{F}} \mathbb{E}_S f(z) - \mathbb{E}_{S'} f(z)$, (*)

can we directly bound this using Massart's Lemma?

Q2: upper bounding (*) by

$$\mathbb{E}_{S\sim D^m} \sup_{f\in\mathcal{F}} \mathbb{E}_S f(z) + \mathbb{E}_{S'\sim D^m} \sup_{f\in\mathcal{F}} \left(-\mathbb{E}_{S'} f(z)\right)$$

(w/o introducing Rademacher r.v's.)

will the pf still go through?

Lower bounds • for statistical learning

$O(\frac{d}{\varepsilon^2})$ training examples ⇒ ERM has excess error $\hat{\varepsilon} \le \varepsilon$.

what if we only have $O(d)$ examples? $o(d)$

$(A, \mathcal{H})$

discuss learnability as a property of hypothesis class only.

Def: $(\mathcal{H}, P)$ — family of distn

$\mathcal{H}$ is said to be agnostic PAC learnable if there exists an alg A and a sample complexity fn $f: (0,1)^2 \to \mathbb{N}$, s.t. ~~for any distn D~~, $D \in P$ for any $\varepsilon > 0, \delta > 0$, if $m \geq f(\varepsilon, \delta)$, then w.p. $1-\delta$, over the draw of m training examples iid from D,

$$\text{err}(A(S), D) - \min_{h' \in \mathcal{H}} \text{err}(h', D) \leq \varepsilon$$

($A(S)$: h)

Def: $\mathcal{H}$ is said to be (realizable) PAC learnable if there exists an alg A and a sample complexity fn $f: (0$ — realizable by $\mathcal{H}$

$\to \mathbb{N}$, s.t. for any distn D, for any $\varepsilon > 0, \delta > 0$, if $m \geq f(\varepsilon, \delta)$, then w.p. $1-\delta$, over the draw of m training examples iid from D,

$$\text{err}(A(S), D) - \min_{h' \in \mathcal{H}} \text{err}(h', D) \leq \varepsilon$$

($A(S)$: h; $\min$: 0)

Finite VC dim $\Rightarrow$ PAC learnable

PAC learnable $\Rightarrow$ Finite VC dim

($\Rightarrow$) Finite VC dim $\Rightarrow$ uniform convergence

$\Rightarrow$ ERM sample complexity $O(\frac{d}{\varepsilon^2})$

$\Rightarrow$ $\mathcal{H}$ is PAC learnable.

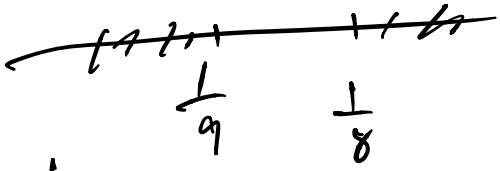
($\Leftarrow$) see next.

Thm: $\mathcal{H}$: $VC(\mathcal{H}) \geq d$. If m = # training example $\leq \frac{d}{2}$.

then, $\forall$ alg A. $\exists$ D. realizable by $\mathcal{H}$,

$$\mathop{\mathbb{E}}_{S\sim D^m} \underbrace{err(A(S), D)}_{X} \geq \frac{1}{4} \quad (*)$$

Remark 1: $(*)$ also implies that



$$P_{S\sim D^m}\left(err(A(S), D) > \frac{1}{8} \right) \geq \frac{1}{8} \quad (\Delta)$$

showing that A does not $(\varepsilon = \frac{1}{8}, \delta = \frac{1}{9})$-PAC learn $\mathcal{H}$ with $m \leq \frac{d}{2}$ examples.

(If A $(\frac{1}{8}, \frac{1}{9})$-learns $\mathcal{H}$ w/ $m \leq \frac{d}{2}$ examples. then

$P_{S\sim D^m}(err(A(S), D) > \frac{1}{8}) \leq \frac{1}{9}$ which contradicts w/ Δ)

$(*) \Rightarrow (\Delta)$

Fact: $X \in [0,1]$, $\mathbb{E}X \geq \frac{1}{4}$. then $P(X > \frac{1}{8}) \geq \frac{1}{8}$

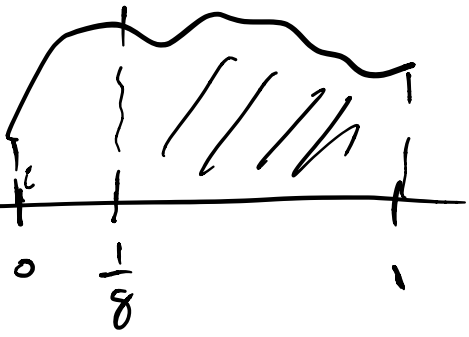
$\geq \frac{1}{4}$ $\quad$ $1-X$

Pf:

$$\mathbb{E}X = \mathbb{E}\left[X \mathbb{I}(X \leq \tfrac{1}{8})\right] + \mathbb{E}\left[X \mathbb{I}(X \in (\tfrac{1}{8}, 1])\right]$$

$\leq \frac{1}{8}$ $\quad$ ≤ 1

$\leq \frac{1}{8}$ $\quad$ $\geq \frac{1}{8}$



$$\leq \frac{1}{8} + P(X > \tfrac{1}{8})$$

$$\Rightarrow P(X > \tfrac{1}{8}) \geq \frac{1}{8}.$$

Remark 2: $VC(H) = \infty \Rightarrow H$ is not PAC learnable (d=2m in the above thm)

b/c, $VC(H) = \infty \Rightarrow \forall m, \forall Alg, \exists$ distn D realizable by H,

$$P_{S \sim D^m}\left(err(A(S), D) > \tfrac{1}{8}\right) > \frac{1}{7}.$$

$f(\frac{1}{8}, \frac{1}{7})$.

$\mathbb{E}_{S \sim D^m} err(A(S), D) \geq \frac{1}{4}$

max
D: D realizable by H

Pf of thm

$\forall$ alg A. $\exists D$. realizable by H,

$$\mathbb{E}_{S \sim D^m} err(A(S), D) \geq \frac{1}{4} \quad (*)$$

$\Updownarrow$

$$\min_{A} \max_{D: \text{realizable by } H} \mathbb{E}_{S \sim D^m} err(A(S), D) \geq \frac{1}{4} \quad (\square)$$

(Minimax lower bound)

define a family of distributions $\mathcal{D} = \{ D_b : b \in \{\pm 1\}^d \}$

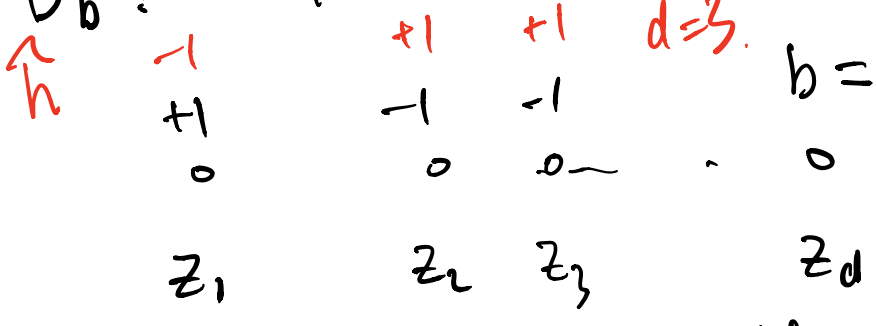
want to show

$$\min_A \mathop{\mathbb{E}}_{b \sim U(\pm 1)^d} \mathop{\mathbb{E}}_{S \sim D^m} \text{err}(A(S), D) \geq \frac{1}{4}$$

(which implies (D))

First: find a set of examples $z_1 \dots z_d$ shattered by $\mathcal{H}$.

define D_b: $P(x = z_i, y = b_i) = \frac{1}{d} \quad \forall i = 1 \dots d$



$b = (+1, -1, -1)$

$b_3 = -1$

all D_h's are realizable by $\mathcal{H}$.

Shorthand: $\hat{h} = A(S)$

$$\forall A. \quad \mathop{\mathbb{E}}_{b,S} \text{err}(\hat{h}, D_b) \geq \frac{1}{4}$$

given h. $\quad \text{err}(h, D_b) = \sum_{i=1}^{d} \frac{1}{d} \mathbb{I}(h(z_i) \neq b_i)$

want to show $\quad \sum_{i=1}^{d} \mathop{\mathbb{E}}_{b,S} \mathbb{I}(\hat{h}(z_i) \neq b_i) \geq \frac{d}{4}$.

we are going to show:

$$\underset{b,S}{P}\left(\hat{h}(z_1) \neq b_1\right) \geq \frac{1}{4} \quad \text{(can show this for other } z_i\text{'s)}$$

(#)

conditioned on $z_1 \notin S_X$ independent of b_1.

$b_1 \cdots b_d$ $\quad S_X = \{z_2, z_3, z_5, z_7\}$

$S_X = \{z_1, z_2, z_5, z_7\}$

$S = \{(x_1, y_1) \cdots (x_m, y_m)\}$ — z_1 b_1

$S_X = \{x_1 \cdots x_m\} \overset{iid}{\sim} \text{unif}(\{z_1 \cdots z_d\})$

the unlabeled training examples.

$\hat{h}$

$I(\hat{h}(z_1) \neq b_1)$

(expectation ?)

$$(\#) \geq \underset{b, S_X}{P}\left(\hat{h}(z_1) \neq b_1,\ z_1 \notin S_X\right)$$

$$= \underset{b, S_X}{P}\left(\hat{h}(z_1) \neq b_1 \mid z_1 \notin S_X\right) \cdot \underset{S_X}{P}\left(z_1 \notin S_X\right)$$

① ②

② $P(z_1 \in S_x)$

$= P(z_1 \in \bigcup_i \{x_i\})$

$\leq \sum_{i=1}^{m} P(z_1 = x_i) = \frac{m}{d} \leq \frac{1}{2}$

② $\geq \frac{1}{2}$.

① conditioned on $z_1 \notin S_x$,

$\hat{h}(z_1)$ independent of b_1

$P(\hat{h}(z_1) \neq b_1 \mid z_1 \notin S_x) = \frac{1}{2}$.

(#) $\geq \frac{1}{2} \cdot \frac{1}{2} \geq \frac{1}{4}$.

<u>Def</u>: $\mathcal{H}$ is said to satisfy the uniform convergence property if there exists a fn $f_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$, such that for any distn D, for any $\varepsilon, \delta > 0$, if $m \geq f_{\mathcal{H}}(\varepsilon, \delta)$, then w.p. $1-\delta$, over the draw of m iid training examples from D,

$$|\,err(h,S) - err(h,D)\,| \le \varepsilon,$$

for all $h \in \mathcal{H}$.

Thm: (the fundamental thm of statistical learning)

The following statements are equivalent:

1. $\mathcal{H}$ satisfies the uniform convergence property
2. $\mathcal{H}$ is agnostic PAC learnable w/ ERM
3. $\mathcal{H}$ is agnostic PAC learnable
4. $\mathcal{H}$ is PAC learnable
5. $\mathcal{H}$ has finite VC dimension

S: observations in nature

$\mathcal{H}$: scientific theory

$\mathcal{H}$ has infinite VC dim "unfalsifiable"

Pf: $1 \Rightarrow 2$. sample size $\geq f_{\mathcal{H}}(\varepsilon/2, \delta)$

$2 \Rightarrow 3$ trivial

$3 \Rightarrow 4$ (seen before)

$4 \Rightarrow 5$ (just proved)

$5 \Rightarrow 1$ (last class)

Occam's Razor