Consistency, Compression, and Learning: The Finite $\mathcal{H}$ Case

Recipe for (PAC) Learning:

1. Design algo L that finds $h \in \mathcal{H}$ consistent with sample S

   $(\hat{\varepsilon}_S(h) = 0)$

2. Analyze how big $m = |S|$ must be s.t. $\varepsilon(h) \le \varepsilon$.

We will show:

- This recipe always works
- Answer to 2. is independent* of L – consistency is all that matters.

Let's warm up with the case of finite $\mathcal{H}$.

Notation

$$\varepsilon(h) \triangleq \Pr_{x \sim D}[h(x) \neq c(x)]$$

true error

$$\hat{\varepsilon}_S(h) \triangleq \frac{1}{m} \sum_i I[h(x_i) \neq y_i]$$

where $S = \{\langle x_1, y_1 \rangle, \dots, \langle x_m, y_m \rangle\}$

training error

Q: When does $\hat{\varepsilon}_S(h) = 0$ imply $\varepsilon(h)$ small?

- Fix $\varepsilon > 0$
- Call $h \in \mathcal{H}$ ε-bad if $\varepsilon(h) \geq \varepsilon$
- $\forall$ fixed ε-bad h:

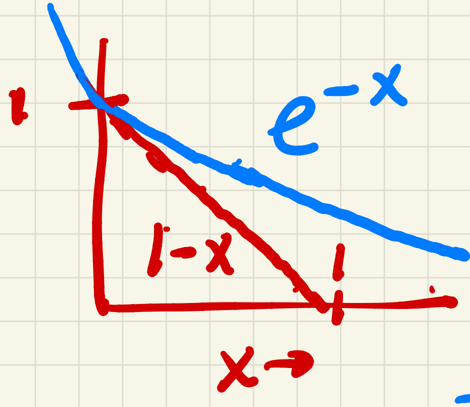
$$\Pr_S[\hat{\varepsilon}_S(h) = 0] \leq (1-\varepsilon)^m$$

$$\Downarrow \quad \text{indep.}$$

$$\Pr_S[\text{any } \varepsilon\text{-bad } h \in \mathcal{H} \text{ has } \hat{\varepsilon}(h) = 0]$$

$$\leq |\mathcal{H}|(1-\varepsilon)^m \quad \text{union bound}$$

Algebra:



$$|\mathcal{H}|(1-\varepsilon)^m \le$$

$$|\mathcal{H}|\, e^{-\varepsilon m} \Rightarrow$$

$\therefore\ 1-x \le e^{-x}$

set $\le \delta$ & solve:

$$\boxed{m \ge \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta}}$$

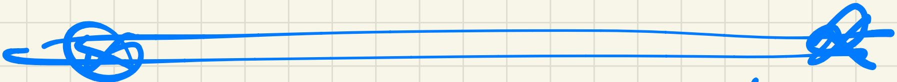
"complexity" of $\mathcal{H}$: $\ln|\mathcal{H}|$

$\approx$

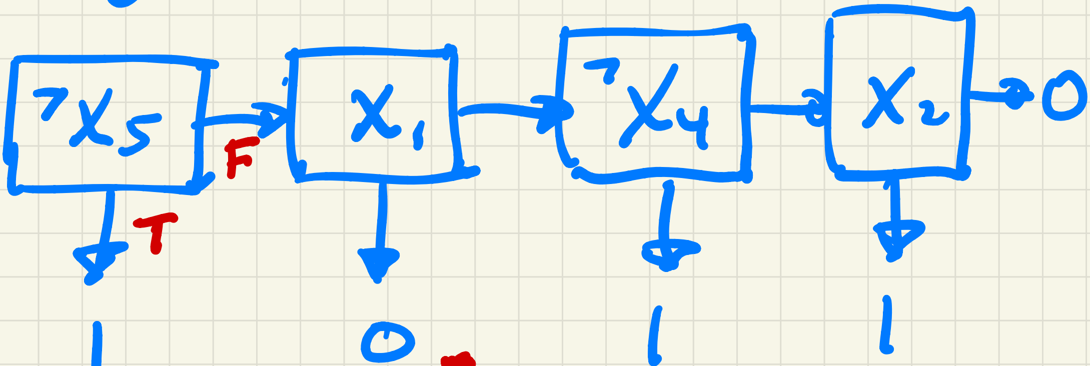
\# bits needed to describe $h \in \mathcal{H}$

- Immediately applies to & simplifies PAC analyses for rectangles, conjunctions, 3CNF
- Any consistent $h \in \mathcal{H}$ suffices

Another application: decision lists over $\{0,1\}^n$.



e.g. $c \in \mathcal{C}$ given by:

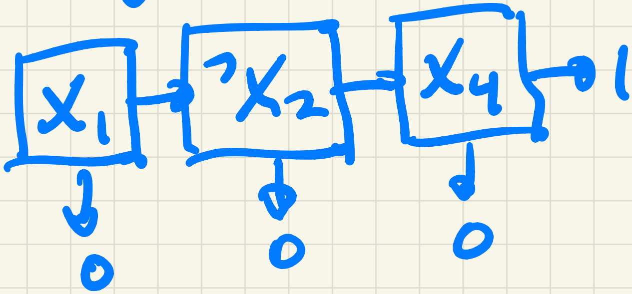


$c(01101) = 1$

$c(10011) = 0$

Contains conjunctions:

$x_1 \neg x_2 x_4 =$



Consistent algo:

- $S \leftarrow$ all examples
- $h \leftarrow$ empty list
- while $(S \neq \emptyset)$:
  - $\forall z, b$:

    $$S_{z,b} \leftarrow \{\langle x, y\rangle \in S : z = 1 \text{ in } x \text{ and } y = b\}$$

  - find max $|S_{z,b}|$ s.t. $S_{z,\neg b} = \emptyset$
  - append $\underset{\downarrow \atop b}{\boxed{z}}$ to h
  - $S \leftarrow S - S_{z,b}$

<u>Theorem</u> Decision lists are PAC learnable (by $\mathcal{H}$ = decision lists).

Learning & Compression

- For $h \in \mathcal{H}$, let $\ell(h)$ be the # bits needed to describe h.

- In general, $\ell(h) \sim \text{poly}(n)$ (dim. of X)

$$\Rightarrow |\mathcal{H}| \sim 2^{\text{poly}(n)}$$

- Suppose we allow

$$\ell(h) \sim \text{poly}(n, m)$$

$|S|$

Good or bad idea?

If we allow

$\ell(h) \sim n \cdot m$ (linear in both)

then we can just

encode/memorize S!

($\mathcal{H}$ = lists of $\langle x, y \rangle$)

So linear dependence on m goes too far.

Let's try $\ell(h) \approx c \cdot m^{\alpha}$ $\qquad \alpha < 1$

includes dep. on n

$$|\mathcal{H}| e^{-\varepsilon m} = 2^{c m^{\alpha}} e^{-\varepsilon m}$$

$$\leq e^{c m^{\alpha} - \varepsilon m}, \text{ set } \leq \delta:$$

$$c m^{\alpha} - \varepsilon m \leq \ln(\delta)$$

$$\varepsilon m \geq \ln(1/\delta) + c m^{\alpha}$$

Satisfied if:

$$m \geq \frac{2}{\varepsilon} \ln(1/\delta) \;\&\; \boxed{m \geq \frac{2 c m^{\alpha}}{\varepsilon}}$$

$$m \geq \frac{2cm^{\alpha}}{\varepsilon}$$

$$m^{1-\alpha} \geq \frac{2c}{\varepsilon}$$

$$m \geq \left(\frac{2c}{\varepsilon}\right)^{\frac{1}{1-\alpha}}$$ ← blows up as $\alpha \to 1$!

$\alpha = 0$: $m \sim 2c/\varepsilon$, original bound

$\alpha = 1/2$: $m \sim \left(\frac{2c}{\varepsilon}\right)^2$

$\alpha = 1$: $m \sim \infty$

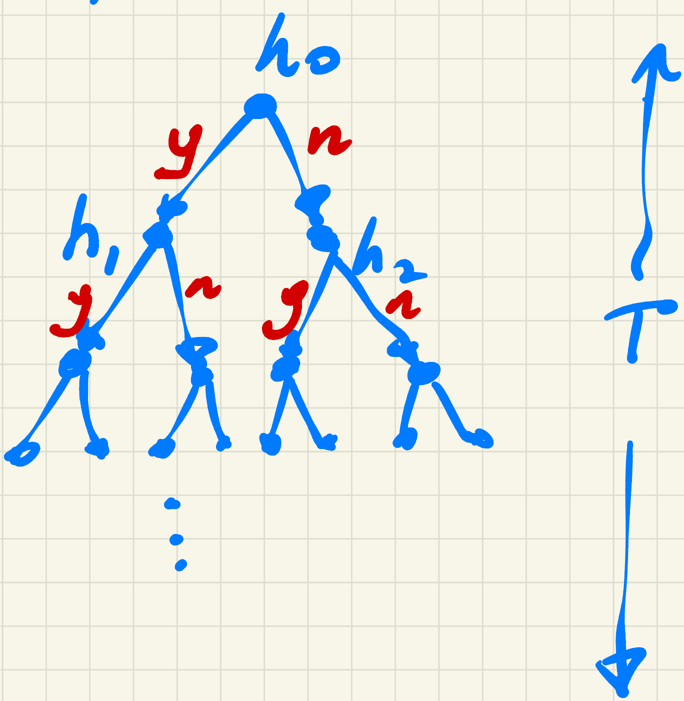
So not just consistency
but even the slightest
compression of S
yields PAC learning
(with larger m).

Application to Adaptive ML/Crowdsourcing

- Imagine we have data S_{train}, S_{test}
- We give S_{train} to crowd
- We keep S_{test} to validate
- Now $\mathcal{H}$ not fixed in advance and not finite!
- How big should $m = |S_{test}|$ be?

A Modified Leaderboard

- When given model h:
  - improvement $< \alpha$ on S_{test}: answer **no**, nothing else
  - improvement $\geq \alpha$ on S_{test}: **yes**, update best error



• Observation: at most $1/\alpha$
y's on any path

$\therefore$ total size of tree

$= |\mathcal{H}| \le \binom{T}{1/\alpha} \le T^{1/\alpha}$

Solve: $T^{1/\alpha} e^{-\varepsilon^2 m} \le \delta$

$$\varepsilon^2 m \ge (1/\alpha) \ln\left(\frac{T}{\delta}\right)$$

$$m \ge \left(\frac{1}{\varepsilon^2 \alpha}\right) \ln\left(\frac{T}{\delta}\right)$$

One more variation.

So far we have shown (for m large)

$$|\hat{\varepsilon}_S(h) - \varepsilon(h)| =$$

$$|0 - \varepsilon(h)| = \varepsilon(h) \leq \varepsilon$$

for all consistent h.

What about all the other $h \in \mathcal{H}$?

Chernoff bounds

- Consider biased coin with $\Pr[\text{heads}] = p$
- Flip m times, let $\hat{p}$ = fraction of heads

Then $\forall \gamma > 0$:

$$\Pr[|\hat{p} - p| \geq \gamma] \leq 2e^{-m\gamma^2/3}$$

$\to 0$ exponentially fast

- A "concentration inequality"

- $\forall$ fixed $h \in \mathcal{H}$:

$\Pr[\,|\hat{\varepsilon}_S(h) - \varepsilon(h)| \geq \varepsilon\,] \leq$ blah

- Prob. any $h \in \mathcal{H}$ has $\leq |\mathcal{H}| \cdot$ blah

$|\cdot| \geq \varepsilon$

- $\leq \delta$ if $\boxed{m \sim \frac{1}{\varepsilon^2} \ln \frac{|\mathcal{H}|}{\delta}}$

"uniform convergence"

WYSIWYG

Fine.

But what if

$\mathcal{H}$ is

<u>infinite</u>???