Last time:
- can efficiently improperly learn 3-term DNF; 🙂
- can't " properly " " ", 🙁
  unless NP = RP

Today: Start unit on sample complexity of PAC learning / VCDIM
- lower bound on ●: $\Omega(\mathrm{VCDIM}(\mathcal{C})/\varepsilon)$ $(\delta = 0.1)$
- start upper bound on ●: roughly $O(\mathrm{VCDIM}(\mathcal{C})/\varepsilon)$

Reminder: in-class, closed-book, closed-note midterm on Tues
up thru today

Questions?

Sample Cxity of PAC Learning

Motivation: Let $\mathcal{C}$ be a concept class.

??
- Is $\mathcal{C}$ PAC learnable?
  (is there some $m(\varepsilon,\delta)$ sample cxity that suff. for (ε,δ)-PAC learning?)
- If so, what's $m(\varepsilon,\delta)$?

Ignore computational cxity / runtime

For finite $\mathcal{C}$, half answer: CHF thm ⇒
any CHF, using $\mathcal{C}$ as hyp. class for $\mathcal{C}$,
is $(\overset{0.1}{\varepsilon}, \overset{''}{\delta})$-PAC alg. given

$$m \approx \frac{1}{\varepsilon} \cdot \log|\mathcal{C}| \text{ examples.}$$

?lower bound?	
What if $\lvert\mathcal{C}\rvert = \infty$?	Hope: intervals...

VCDIM($\mathcal{C}$): full answer to all $\mathcal{C}$'s.

Let $\mathcal{C}$ be any concept class over X.
Let $d :=$ VCDIM($\mathcal{C}$). (= size of largest $S \subseteq X$ that's shattered by $\mathcal{C}$.)

This unit's results:

☹: Any PAC alg for $(\overset{0.1}{\varepsilon}, \overset{''}{\delta})$-learning $\mathcal{C}$ must use $\Omega(d/\varepsilon)$ examples. (So if $d = \infty$, can't PAC learn w/ any finite samp. cxity.)

☺: If d finite: a CHF for $\mathcal{C}$ using hyp class $\mathcal{C}$ is an $(\overset{0.1}{\varepsilon}, \overset{''}{\delta})$-PAC learner as long as it's run on

$$m \gtrsim \frac{d}{\varepsilon} \text{ examples.}$$

Analogue of old Occam result, w/

$d = \text{VCDIM}(\mathcal{C})$ in place of $\log|\mathcal{C}|$.

$X = \{1, 2, 3, \ldots\}$

$\mathcal{C} =$ all finite subsets of X

$\text{VCDIM}(\mathcal{C}) = \infty$

☹ Lower Bound of Sample Cxity of PAC Learning

Thm: Let $\mathcal{C}$ = any concept class, let $d = \text{VCDIM}(\mathcal{C})$. Any PAC learner for $\mathcal{C}$ w/ accuracy param. ε, conf. param. $\delta = 0.1$, must use

$$\Omega(d/\varepsilon) \text{ examples.} \qquad \left(\frac{d-1}{32\varepsilon}, \text{ for } d \geq 100\right)$$

Pf idea: labeling of shatt. set completely unpredictable; must see most of it to have high acc. on it.

Pf: First, let's give a $\mathcal{D}$ s.t. need $\geq \frac{d}{2}$ ex. to achieve error $\frac{1}{8} = \varepsilon$, conf. $\frac{1}{8} = \delta$.

Let $S = \{\overbrace{x^1, \ldots, x^d}^{\text{points}}\} \subseteq X$ be shatt. by $\mathcal{C}$.

Let $\mathcal{D}$ put wt $1/d$ on each x^i, no wt on other ex.

Let A be a purported learning alg that only makes $\frac{d}{2}$

calls to $EX(c, \mathcal{D})$.

A "sees" labels of $\le \frac{d}{2}$ $\overset{\text{points}}{\overset{\text{examples}}{\text{elements}}}$ in S.

Consider a $\overset{\text{target}}{c} \in \mathcal{C}$ chosen uniformly from the coll. of 2^d concepts shattering S.

Each of the 2^d lab. of S are =ly likely for such a c; for each $x^i \in S$ that wasn't "seen" by A,

$\frac{1}{2}$ chance $c(x^i)=1$

$\frac{1}{2}$ chance $c(x^i)=0$.

So A's hyp. is very unlikely to do well on $\geq$ half of the wt of $\mathcal{D}$.

More precisely,

$$\mathbb{E}[\text{error of learner}] \geq \overset{\text{for each unseen pt}}{\frac{1}{2}} \cdot \frac{\overset{\text{l.b. on \# of unseen pts}}{d/2}}{\underset{\frac{1}{d} = \text{wt of each pt.}}{d}} = \frac{1}{4}$$

Let acc $= 1 -$ error.

$\mathbb{E}[\text{acc}] \leq 3/4$.

Markov: take $k = 7/6$, &

$$\Pr\left[\text{acc} \geq 7/8 = \underset{\substack{\| \\ k}}{\frac{7}{6}} \cdot \frac{3}{4}\right] \leq \frac{1}{k} = \frac{6}{7}.$$

So $\Pr\{\text{error} \geq \frac{1}{8}\} \geq \frac{1}{7}$.

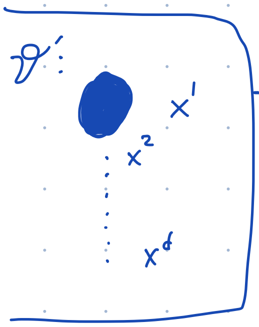
So didn't sat. $(\varepsilon = 1/8, \delta = \frac{1}{8})$ PAC.

Now: give a $\mathcal{D}'$ s.t. need $\Omega(d/\varepsilon)$ ex. to $(\varepsilon, \delta = \frac{1}{10})$-learn.

Let S as before $\{x^1, \dots, x^d\}$ shattered.

Let $\mathcal{D}'$ be

- $\Pr_{\mathcal{D}'}[x^1] = 1 - 8\varepsilon$
- $\Pr_{\mathcal{D}'}[x^i] = \frac{8\varepsilon}{d-1}$ $\quad i = 2, \dots, d.$



(Note: maybe $\{x^1, \dots, x^d\} = S$ is all of X; so we do

Suppose learner A receives only $\frac{d-1}{2}$ ex. from $x^2, \dots, x^d$ from $EX(c, \mathcal{D}')$.

As before, Let c be unif. over 2^{d-1} conc. in $\mathcal{C}$ giving all labelings of $x^2, \dots, x^d$.

Using warmup from above, have that w.p. $\geq \frac{1}{8}$, A has error $\geq \frac{1}{8}$ on the 8ε "amount of $\mathcal{D}'$" that's on $x^2, \dots, x^d$, i.e. overall error $\geq \varepsilon$.

So assuming A receives only $\frac{d-1}{2}$ ex. from $x^2, \dots, x^d$ from $EX(c, \mathcal{D}')$, have that w.p. $\geq \frac{1}{8}$, error $\geq \varepsilon$.

Each call A makes to $EX(c,\mathcal{D}')$ hits $\underline{\{x^2, \dots, x^d\}}$ w.p. $\underbrace{8\varepsilon}_{=p}$.

If A makes only $\overbrace{\frac{1}{32} \cdot \frac{d-1}{\varepsilon}}^{m}$ calls,

$$\mathbb{E}[\#\text{hits}] = \frac{1}{32} \cdot \frac{d-1}{\varepsilon} \cdot 8\varepsilon = \frac{d-1}{4}.$$

Mult. CB $\left(\gamma = 1,\ pm = \frac{d-1}{4}\right)$:

$$\Pr\left[\#\text{hits} \geq \frac{d-1}{2}\right] \leq \underbrace{\exp\left(-\frac{d-1}{12}\right)}_{\leq 0.01}.$$

$\Omega(d/\varepsilon)$

Can assume $d \geq 50$, so

So $\boxed{\text{w. prob.} \geq \frac{99}{100} \cdot \frac{1}{8} > \frac{1}{10}, \text{ error} \geq \varepsilon.}$ ■

Upper Bound on sample cxty of PAC learning

$\boxed{d = \text{VCDIM}(\mathcal{C})}$

Good news: $\underline{\forall \mathcal{D}}$, $\underline{\text{any}}$ hyp $h \in \mathcal{C}$ that's consist. with $m \gtrsim \frac{d}{\varepsilon} + \frac{1}{\varepsilon} \log \frac{1}{\delta}$ rand. ex. from $EX(c,\mathcal{D})$, is ε-acc. w.p. $\geq 1-\delta$. (has $err_D(c,h) \leq \varepsilon$).

Pf a bit complex...

High level outline:

1) **Setup:** think more deeply about $VCDIM(\mathcal{C}) = d$.

Useful to consider a related function (the "growth function of $\mathcal{C}$", $\Pi_{\mathcal{C}}(m)$).

2) **Combinatorics argument:** Amazing Theorem about $\underline{\hspace{4cm}}$, for any concept class $\mathcal{C}$.

3) **Learning-style argument:** Like CHF thm, but more subtle/sophisticated.

---

1) **Setup** $\mathcal{C}$ over X

- Recall $VCDIM(\mathcal{C})$ = size of largest $S \subseteq X$ that's shatt. by $\mathcal{C}$.

- Another P.O.V.: given a set $S \subseteq X$,

Def: let $\boxed{\Pi_{\mathcal{C}}(S)}$ = set of all labelings of S induced by $\mathcal{C}$.

i.e. $\Pi_{\mathcal{C}}(S) = \{ c \cap S : c \text{ ranges over } \mathcal{C} \}$

Ex: $S = \{1,2,3\}$ $\quad X = \mathbb{R}$

$\mathcal{C}$ = all intervals $[a,b]$.

We know $\mathrm{VCDIM}(\mathcal{C}) = 2$.

$S = \{7,9\}$:

$\Pi_{\mathcal{C}}(S) = \{\emptyset, \{7\}, \{9\}, \{7,9\}\}$

Shattered;

$\Pi_{\mathcal{C}}(S) = 2^{|S|}$.

Have

$$\Pi_{\mathcal{C}}(S) = \left\{ \overset{[7,9]}{\emptyset}, \{1\}, \{2\}, \{3\}, \overset{[.5, 2.2]}{\{1,2\}}, \{2,3\}, \overset{[0,98]}{\{1,2,3\}} \right\}$$

$\mathcal{C}$ does <u>not</u> shatter $\{1,2,3\}$; $|\Pi_{\mathcal{C}}(\{1,2,3\})| = 7$

Note: $\mathcal{C}$ shatters $S \iff \Pi_{\mathcal{C}}(S) = 2^{|S|}$.

Define: $\Pi_{\mathcal{C}}(m)$ = max value of $|\Pi_{\mathcal{C}}(S)|$ over all subsets S with $|S| = m$.

"growth function of $\mathcal{C}$".

Ex: as above, $\mathcal{C}$ = intervals

$$\Pi_{\mathcal{C}}(1) = 2$$

$$\Pi_{\mathcal{C}}(2) = 4$$

$$\Pi_{\mathcal{C}}(3) = 7$$

Note: $\mathrm{VCDIM}(\mathcal{C})$ = max value d s.t. there exists a $S \subseteq X$, $|S| = d$, with $\Pi_{\mathcal{C}}(S) = 2^d$, i.e.

max d s.t. $\Pi_e(d) = 2^d$.

End of setup.

Next time: 2) Amazing Theorem