Last time: Amazing Fact (Sauer's Lemma):

\[ E \in \text{VCdim} d : \quad \text{Pr}_e \left( \overline{\Pi}_e(m) \leq \left( \frac{e \ln d}{d} \right)^m \right) \text{ for } m \geq d \]

- "Double sample" arg. \[ \forall E \in \text{VCdim} d, \forall \gamma, \forall \epsilon \leq 1, \text{ given } m \geq \text{ EX}(e, \gamma), \text{ if} \]

\[ m \geq \frac{2}{\epsilon} \left( \ln \overline{\Pi}_e(2m) + \ln \frac{2}{\delta} \right) \]

w.p. \( \geq 1 - \delta \) all bad \( h \in E \) not cons. w/ data.

\[ \Rightarrow \text{Pr} \left[ c(x) \neq c(x) \right] > \epsilon \quad x \sim \gamma \]

Today:

- Combine pieces & get "CHF works for classes w/ finite VCdim" theorem.

  - Applic: "PAC alg. for LTFs over \( \mathbb{R}^m \)"

Start new unit: Boosting. Start KV Chap. 4

(go through analysis of "one stage" of booster; 4.0 - 4.32)

Schapire survey 9.1-3)

To get "CHF \( \Rightarrow \) PAC" for \( E \in \text{VCdim}(E) = d \):

\[ m \geq \frac{2}{\epsilon} \left( \ln \left( \frac{e \ln d}{d} \right)^m + \ln \frac{2}{\delta} \right) \]

is enough. i.e.

\[ m \geq \frac{2}{\epsilon} \left( d \ln \left( \frac{e \ln d}{d} \right) + \ln \frac{2}{\delta} \right) \]

Can verify:

\[ m \geq C \left( \frac{d \ln \frac{1}{\epsilon} + \frac{1}{\epsilon} \ln \frac{1}{\delta}}{\epsilon} \right) \]

suff. for \( \otimes \) to hold.
\[ m > \frac{2}{\varepsilon} \frac{d}{1 - (\frac{em}{d})} : \left( \frac{em}{d} \right) > \frac{2e}{\varepsilon} \frac{1}{\ln \left( \frac{em}{d} \right)} \]

\[ A > k \cdot \ln A ; \quad \frac{A}{\ln A} > k ; \quad A = 2k \ln k \text{ enough.} \]

What's this good for?

Applic: \( E \) = all CTFs over \( X = \mathbb{R}^n \).

1. \( |E| = \infty \).

Fact 1: \( \text{VCdim}(E) = n + 1 \). \( (n=2: \ 3) \)

Fact 2: There is an efficient CHF for \( E \) using \( c = \) as hyp's.

(poly-time linear programming)

\[ (x, c(x)) = (0, 3, 7, 4, \ldots, 17) \quad c(x) = +1 \]

\[ w \cdot x > \theta : \quad 0, 3w_1 + 7w_2 + 4w_3 + \ldots + 17w_n > \theta \]

\[ (4, 3, -2, \ldots, 5) \quad c(x) = -1 \]

\[ x \quad 4w_1 + 3w_2 - 2w_3 + \ldots + 5w_n < \theta \]

\( \) gives a CHF for \( E \) using \( E \).

To eff. PAC learn CTFs over \( \mathbb{R}^n \):
\[ \text{draw } m \sim \frac{8}{\varepsilon} \left( (n+1) \ln \frac{1}{\varepsilon} + 1 \cdot \frac{3}{\delta} \right) \]

*feed to poly-time CP to get cons. hyp. \((w, \ldots, w_i; \theta) \sim [w \cdot x > \theta] \quad \varepsilon\text{-acc. w.p. } 1-\delta.\)

---

**New Topic:**

**WEAK LEARNING, STRONG "", "", BOOSTING**

---

Recall PAC learning def:

**Def:** Alg. \(A\) PAC learns class \(C\) if: \(\forall c \in C, \forall \text{dist. } \mathcal{D}, \forall \varepsilon > 0, \forall \delta > 0\), given \(\mathcal{E}(c, \mathcal{D})\), w.p. \(1-\delta\) \(A\) outputs \(\varepsilon\)-acc. hyp.

**Strong guarantee!** 

**STRONG PAC Learning**

**Motiv. q:** Suppose only have weaker guar. on alg. can achieve some \(\varepsilon, \delta\). What's this good for?
A lot!

Suppose you have "limited-conf." ab $\approx (A^')$
w. prob. $\geq \frac{3}{4}$, gives $\varepsilon$-acc. hy. $P_{\delta_{\frac{1}{4}}}$. Can easily convert into strong PAC ab $(arb. \delta)$. To achieve conf $1-\delta$, acc $\varepsilon$

run $A'$ \hspace{0.5cm} $10 \ln \frac{2}{\delta}$ \hspace{0.5cm} $O(\log \frac{1}{\delta})$ times, using $\frac{\varepsilon}{2}$ as its acc.

w. prob. $\leq \left(\frac{1}{4}\right)^{10 \ln \frac{2}{\delta}} < \frac{\delta}{2}$, all runs fail to give $\varepsilon/2$ acc hy. I.e.

w. prob. $\gg 1-\frac{\delta}{2}$, get $h, \ldots, h_{10 \ln \frac{2}{\delta}}$ s.t. one $h_i$ is $\varepsilon/2$ acc.

Draw some more ex, eval. each $h_i$ on them, output the one that does best.

\hspace{1cm} $O(\log \frac{1}{\delta})$ dep.

\hspace{1cm} $\approx \frac{1-\frac{3}{2}}{1}$ estim. acc. of each $h_i$ is $\pm \frac{\varepsilon}{10}$ of true acc.

\hspace{1cm} True-error $\leq \varepsilon_2$ → appear to have error $\leq \frac{6\varepsilon}{10}$

\hspace{1cm} $\varepsilon \geq \varepsilon_2$ → $\varepsilon \geq \frac{9\varepsilon}{10}$

So outputting $h_i$ appearing best, is $\varepsilon$-acc. $\left(\approx \frac{1-\frac{3}{2}}{1-\delta}\right)$

The above is why $O(\log \frac{1}{\delta})$ ex suff. for PAC learning.
What about $e$? Can we boost accuracy?

**Def**: Alg $A$ is a weak PAC learning alg for $c$ with advantage $\gamma$ if $\forall c \in C$, $\forall$ dist $D$, with prob. $1-\delta$, $A$ outputs hyp $h$ s.t.

$$\Pr_{x \sim D} [h(x) \neq c(x)] \leq \frac{1}{2} - \gamma.$$ 

It seems that "$C$ is weakly PAC learnable" not as good as "$C$ strongly PAC learnable" — not so! Weak PAC learn $\not\implies$ strong PAC learn.

Can autom. "boost" weak PAC to strong PAC:

**Thm**: Let $C$ any c.c. Suppose there is (time $T$) an eff. weak PAC alg $A$ for $C$ with adv. $\gamma$. Then there is a poly $(\frac{1}{\delta}, \frac{1}{\epsilon}, \log \frac{1}{\delta}, T)$ strong PAC alg ($\epsilon$-acc hyp. w/prob. $\geq 1-\delta$) for $C$. 

Thus is proved by giving an explicit, efficient boosting alg $B$. "Meta-alg":

$B$ runs $WL \ poly \left( \log \frac{1}{\varepsilon} \frac{1}{y} \right)$

not mistake! Each run can require $poly(\varepsilon)$ time, ex.

<table>
<thead>
<tr>
<th>Booster $B$</th>
<th>call</th>
<th>$WL$ alg</th>
</tr>
</thead>
<tbody>
<tr>
<td>hyp</td>
<td></td>
<td>final hyp</td>
</tr>
</tbody>
</table>

Hi-level idea of boosting: Given $EX(c, D_i)$ & $WL A$, booster does following:

- Run $A$ repeatedly using $EX(c, D_i), EX(c, D_2), \ldots$ to get hyp's $h_1, h_2, \ldots$
- Combines $h_1, h_2, \ldots$ to get final hyp $h$.

More details:
- What are $D_1, D_2, \ldots$?
- How does booster run $A$ using $\ldots$, etc.
only has \( EX(c, \Theta) \)? How to simulate \( EX(c, \Theta) \) given \( EX(c, \Theta) \)?

- how to combine \( h_1, h_2, \ldots \)?

(maj vote)

\( \square \) Why does it work \( h \approx \text{acc. w.r.t. } \Theta \)?

---

Next: proof of concept that boosting works. One stage of original booster (Schapire). 3 calls to \( wc \).

Fix \( wc A \) with \( \gamma = \text{adv} = \frac{1}{10} \). \( \forall \Theta, \forall c \in C \), \( A \) \text{ always } \text{prob. 1-0} \text{ gives } h \text{ w. error } \gamma \leq \frac{4}{10} \).

\[ \text{Assume } J = 0. \] Assume always get exactly \( \frac{4}{10} \) error of \( h \).

We'll run \( A \) 3 times \& get a hyp. with error 35.2%. Recursively, can get \( \epsilon \) error this way.

First stage: \( \Theta_1 = \Theta \). Run \( A \) on \( EX(c, \Theta_1) \), get

\[ h_1, P_{x \sim \Theta_1} \left[ h_1(x) \neq c(x) \right] = 0.4. \]
Task: Come up with new distr. $\mathcal{D}_2$.

Q: How to use $h_1$, $\text{EX}(c, \mathcal{D}_1)$ to simulate a useful $\text{EX}(c, \mathcal{D}_2)$? Next time...