Last time: Proper vs Improper learning
(improperly learning
3-term DNF is computationally easy, but properly learning
3-term DNF is computationally hard)

Today: Start unit on sample complexity of PAC learning (VC Dimension)
- lower bound on \( D \left( \frac{\text{vc}(e)}{e} \right) \)
- start upper bound on \( \Theta \left( \frac{\text{vc}(e)}{e} \right) \)

Reminder: midterm Thurs (cover stuff thru Thurs 10/18)

Questions?

Sample Cxity of PAC Learning

\( C = \) a conc. class. Basic PAC questions:
1) is \( C \) PAC learnable? (by some alg. w/ fixed finite s.c.?)
2) if so, how many ex. needed?

Suppose \( C \) finite. Get (half) answer from CHF thus:
\( \frac{1}{\varepsilon} (1/\ell 1/\ell + 1/n/\ell) \)

ex. suffice. lub.
- what about a lower bound?
What if $C$ infinite?

$\text{VC}(C)$ completely answers all the above $\varphi$'s.

Quick overview of results: Fix $C$. Let $d = \text{VC}(C)$.

1. Any PAC learner for $C$ must use $\Omega\left(\frac{d}{\varepsilon} + \frac{1}{\varepsilon}\ln \frac{1}{\delta}\right)$ examples.

2. If $d$ finite, any CHF for $C$ using $C$ as hyp. is a PAC learner if run on $\mathcal{H}$

$$m \gtrsim \frac{1}{\varepsilon} \cdot d \cdot \log \left(\frac{1}{\varepsilon}\right) + \frac{1}{\varepsilon} \log \frac{1}{\delta}$$ examples.

Analogue of our CHF thm for $\omega$ concept classes;

$\text{VC}(C)$ plays role of $\ln |\mathcal{H}|$.

Today: 😊, start 😃

Lower bd. on s.c. of PAC learning

Thm: Fix c.c. $C$. Let $d = \text{VC}(C)$.

Any PAC learning alg. for $C$ that learns to acc. $\varepsilon$, conf. $\delta = \frac{1}{10}$, must use $\Omega\left(\frac{d}{\varepsilon}\right)$ ex. for some $\varphi$. 

}\]
Pf: First: warm-up: a first need \( \Omega(d) \) ex.

- to achieve \( \varepsilon = \frac{1}{8}, \delta = \frac{1}{8} \)  \( \Rightarrow \) \( \{x_1, \ldots, x_d\} \)
- let \( S \subseteq X \), \( |S| = d \), \( S \) shatter by \( C \).
- let \( \mathcal{D} \) be uniform on \( S \) (\( \forall \) prob. each \( x \in S \))
- zero prob. wt. off \( S \).

Let \( A \) be a purported learner that makes only \( \frac{d}{2} \) calls to \( E_X(c, \Psi) \). After these calls, \( A \) "knows" labels of \( \leq \frac{d}{2} \) pts in \( S \).

Let target \( c \in C \) be chosen uniform from the \( 2^d \) concepts in \( C \) that shatter \( S \).

(So label of \( x_i \) under \( c \) is indep \( \& \).

So for each \( x_i \) not "seen" in \( \frac{d}{2} \) pt sample,
\( c(x_i) \) is indep \( \& \).

So...
\[
\mathbb{E}[\text{error of learner's } h] \geq \frac{d}{2} \cdot \frac{1}{2} \cdot \frac{1}{d} = \frac{1}{4}.
\]

Let \( \text{acc}(h) = 1 - \text{error}(h) \)

Pr.v.:

\[
\text{acc}(h) \leq 1
\]

\[
\mathbb{E} [\text{acc}(h)] \leq \frac{3}{4} \quad \Rightarrow \quad \frac{3}{4} = \frac{7}{6} \cdot \frac{3}{4}
\]

Markov:

\[
\Pr\{\text{acc}(h) \geq \frac{7}{8}\} \leq \frac{1}{4}.
\]

So

\[
\Pr\{\text{error}(h) \leq \frac{1}{8}\} \geq \frac{1}{4}.
\]

So this is not meeting the PAC crit. for
\( \varepsilon = \delta = \frac{1}{8} \). (can't be true that \( \forall c \), \( \text{acc}(h) \leq \frac{7}{8} \) error(\( h \leq \frac{1}{8} \))

\( \checkmark \) Done w. warm-up.
Real thing: a $\Omega(\frac{d}{\epsilon})$ ex.

Again $S = \{x'_1, \ldots, x'_d\}$ is shattered.

Now let $\mathcal{Y}$ be:

- $1 - 8\epsilon$ w.t. on $x^d$
- $\frac{8\epsilon}{d-1}$ w.t. on $x'^1, \ldots, x'^{d-1}$

Suppose $\ast$ A's calls to $\text{EX}(c, \mathcal{Y})$ yield only $\frac{d-1}{2}$ occasions when we get one of:

Let $c \in C$ (target) be u.i.f. over a set of $2^{d-1}$ concepts labeling $x'_1, \ldots, x'^{d-1}$ in all ways.

Same arg. as before: assuming $\ast$, w.p. $\geq 1/8$

by p $h$ has error rate $\frac{\epsilon}{8}$ on $x'_1, \ldots, x'^{d-1}$. These have $8\epsilon$ mass under $\mathcal{Y} \Rightarrow h$ has overall error $\geq \epsilon$ under $\mathcal{Y}$ w.p. $\geq 1/8$.

Let $m = \frac{1}{32} \cdot \frac{d-1}{\epsilon}$.

Each call to $\text{EX}$ hits one of $x'_1, \ldots, x'^{d-1}$ w.p. $8\epsilon$.

$E[\# \text{hits}] = 8\epsilon \cdot m = \frac{d-1}{4}$.

Mult. CB ($\gamma = 1$, $p = \epsilon$, $p^{1/m} = \frac{d-1}{4}$):

$\Pr[\# \text{hits} \geq 8\epsilon \cdot m] \leq e^{-\frac{d-1}{12}}$.

Can assume $d \geq 100$ (asympt. statement), so

$e^{-\frac{d-1}{12}} \ll \frac{1}{100}$.

So w.p. $\geq \frac{99}{100}$, $\ast$ holds: so

w.p. $\geq \frac{99}{100}$ have $\geq \frac{1}{\epsilon}$ chance $h$ is bad.
So w.p. \( \geq 0.99 \cdot \frac{1}{2} \) \( (> \frac{1}{10}) \),
\( h \) is bad.

It's not hard to show \( \sigma_2 \left( \frac{1 - (\delta)}{\epsilon} \right) \)  \( \delta \)-b.i., for general \( \delta \).

**Upper Bound on S.C. of PAC Learning.**

**Goal:** \( \forall \mathcal{D}, \forall h \in \mathcal{C} \) that's cons. with
\( \approx \frac{1}{\epsilon} \cdot d + \frac{1}{\epsilon} \cdot \log \frac{1}{\delta} \)
random ex. from \( \mathcal{D} \) is \( \epsilon \)-good w.p. \( 1 - \delta \).

**Proof has 3 main conceptual components.**

1. **Setup:** Ponder \( V C(\mathcal{C}) \leq a \). We'll consider a function ("growth function" of \( \mathcal{C} \)) that'll give us more info than just \( V C(\mathcal{C}) \).

2. **Combinatorics result:** Amazing theorem about growth function.

3. **Learning / prob. argument** (like CHF Theorem, but more sophisticated), using amazing theorem.
1. Setup

\( X = \text{domain} \quad \mathcal{C} = \text{cc. over } X \)
\( d = \text{VC}(\mathcal{C}) = \text{size of largest } S \subseteq X \text{ s.t. } S \text{ shatter } \mathcal{C} \)

Fix \( S \).

**Def:** \( \mathbb{T}_e(S) = \text{set of all labelings of } S \)
induced by concepts in \( \mathcal{C} \).
\( \mathbb{T}_e(S) = \{ c \cap S : c \in \mathcal{C} \} \)
both subsets of \( X \)
coll. of sets

Recall \( \text{VC}(\mathcal{C}) = 2 \) for \( a, b, c \).

**Ex:** \( X = \mathbb{R}, \mathcal{C} = \text{all intervals}, S = \{1, 2, 3\} \).
\( \mathbb{T}_e(S) = \{ \emptyset, \{1\}, \{2\}, \{3\}, \{12\}, \{23\}, \{123\} \} \)

Have \( \mathcal{C} \) shatters \( S \) iff \( |\mathbb{T}_e(S)| = 2^{|S|} \).

So
\( \text{VC}(\mathcal{C}) = \text{max value } d \text{ s.t. } \exists S \subseteq X, |S| = d, \)
with \( |\mathbb{T}_e(S)| = 2^d \).

**Def:** The growth function \( \mathbb{T}_e(m) \) is
\( \mathbb{T}_e(m) = \text{max value of } |\mathbb{T}_e(S)| \)
over all \( S \subseteq X, |S| = m \).

"max # of dichotomies \( \mathcal{C} \) can induce on any \( m \)-set"
Back to same ex:
\( C = \text{intervals} \)

\[
\begin{align*}
\Pi_C(0) &= 1; & \Pi_C(1) &= 2; & \Pi_C(2) &= 4; \\
\Pi_C(3) &= 7; & \text{in general,} \\
\Pi_C(m) &= O(m^2) \\
\end{align*}
\]

3 Amazing Theorem:
Fix any \( C \) w/ \( VC(C) = d \).
Then
\[
\begin{align*}
\Pi_C(m) &= 2^m \quad \text{if } m \leq d \\
\Pi_C(m) &\leq \left( \frac{em}{d} \right)^d \quad \text{if } m > d.
\end{align*}
\]

\( e = 2.718 \ldots \)

\( \text{polynomial (deg } d) \)!