Last time:
- finish learning $\mathcal{C}$ = intervals of $\mathbb{R}$
- OLMB → PAC conversion (get all our OLMB results "for free" in PAC setting :))
- Revisit PAC learning def:
  - "size" of concepts
  - efficiency of evaluating h

Today:
- "Chernoff bounds"/tail bounds (for hypoth. testing)
- learning by finding consistent hyp. from a fixed $\mathcal{H}$
- "Occam's Razor" ("cardinality" version)

Questions?

Motiv. for :

Suppose you have an $h: X \to \{0,1\}$, you also have $EX(c,\mathcal{D})$.

How good is h ⟵ what's $err_{\mathcal{D}}(h,c) = \Pr_{x \sim \mathcal{D}}[h(x) \neq c(x)]$?

To estimate :

draw m ex from $\mathcal{D}$; eval. $h(x)$ on each,

compute $\hat{\varepsilon} = \frac{1}{m} \cdot (\#$ of the m ex. s.t. $h(x) \neq c(x))$.

How good is this est.?

Exactly ≡ to:

($) $\Pr[H] = p$, $0 \le p \le 1$:
p unknown.

To est. p: $\hat{p} = \frac{1}{m}$ (# of times got H in m tosses)

How good is this est.?

"Chernoff bound": ("relative error") Let $X_1, \ldots, X_m$ be

indep ident. dist. (i.i.d) 0/1 random vars

$\mathbb{E}[X_i] = \Pr[X_i = 1] = p$.

Let $X = X_1 + \cdots + X_m$, so $\mathbb{E}[X] = mp$

Then $\forall\, 0 < \gamma < 1$,

$\Pr[X \le (1-\gamma)mp] \le e^{-\frac{1}{2}\gamma^2 mp}$, $\forall$

$\Pr[X \ge (1+\gamma)mp] \le e^{-\frac{1}{3}\gamma^2 mp}$.

Similar bound ("additive error"):

Let $X_1, \ldots, X_m$ as above;

let $\hat{p} = \frac{1}{m} \cdot X$ (empirical average of m samples.)

Then

$$\Pr[p - \hat{p} \geq \varepsilon] \leq e^{-2m\varepsilon^2}$$

$$+ \quad \Pr[\hat{p} - p \geq \varepsilon] \leq e^{-2m\varepsilon^2}$$

confidence

Ex Suppose team wins each game indep. w.p. $\frac{1}{2}$.

162 game season: what's prob. win ≤ 47 games?

1st bound: $i = 1, \ldots, 162$, $X_i = \begin{cases} 1 & \text{win} \\ 0 & \text{lose} \end{cases}$

$\Pr[X \leq 47]$? $\Pr[X \leq (1-\gamma) m p]$ ($m = 162$, $p = \frac{1}{2}$)

$(1-\gamma) 162 \cdot \frac{1}{2} = 47$

$\gamma = 0.42$

$$\leq e^{-\frac{1}{2} \cdot (0.42)^2 \cdot 162 \cdot \frac{1}{2}} = 0.0008$$

Ex: Survey: what % of people like peppermint pizza, to $\pm 3\%$, with conf. 95%?

How many people do we need to poll?

Addit. bound: $e^{-2m\varepsilon^2} = 2.5\% = \frac{1}{40}$

$\varepsilon = 0.03$:

$$e^{-2m\cdot(0.03)^2} = \frac{1}{40}$$

$$m = \frac{1}{0.0018} \cdot \ln(40)$$

All of above: relied on X = sum of indep rand. vars.

Weaker, but more general, tail bound:

Markov's inequality: If X is any non-neg. RV, then for any $k \geq 1$, have

$$\Pr[X \geq k \cdot \mathbb{E}[X]] \leq \frac{1}{k}.$$

(o/w $\mathbb{E}[X]$ would be $> \mathbb{E}[X]$!)

Ex: X = # children in random U.S. household.

$\mathbb{E}[X] = 1.85$

Then $\Pr[X \geq 10] < \frac{1}{5}$.

Learning by Finding Consistent Hypothes (from some given $\mathcal{H}$)

If $\mathcal{H}$ is fixed, & you can find an $h \in \mathcal{H}$ consistent with "enough data": that's a PAC learner!

Thm: Fix $\mathcal{C}$, $\mathcal{H}$. Fix $c \in \mathcal{C}$.
Fix any dist $\mathcal{D}$ over X.

Let $(x^1, c(x^1)), \ldots, (x^m, c(x^m))$ be i.i.d. draws from $EX(c, \mathcal{D})$, where

$$m = \frac{1}{\varepsilon} \cdot \left(\ln |\mathcal{H}| + \ln \frac{1}{\delta} \right)$$

Say h is bad if $er_{\mathcal{D}}(h, c) > \varepsilon$. Then

$\Pr[$ any bad $h \in \mathcal{H}$ is consistent with $] \leq \delta$.

Pf: • Fix any bad h. $\Pr[h$ consist. with the m ex $]$ is

$< (1-\varepsilon)^m$. (indep.)

- Surely $\mathcal{H}$ has $\leq |\mathcal{H}|$ many bad h's.

(UB): $\Pr[\underline{\text{any}}$ bad $h \in \mathcal{H}$ is consistent with our m ex.$]$

$< \underline{(1-\varepsilon)^m \cdot |\mathcal{H}|}$.

$(1-x)^{\frac{1}{x}} \leq \frac{1}{e}$

Plug in: $|\mathcal{H}| \cdot (1-\varepsilon)^{\frac{1}{\varepsilon}(\ln|\mathcal{H}| + \ln(\frac{1}{\delta}))}$

$\leq |\mathcal{H}| \cdot e^{-(\ln|\mathcal{H}| + \ln(\frac{1}{\delta}))} = e^{-\ln(\frac{|\mathcal{H}|}{\delta})} \cdot |\mathcal{H}|$

$= \frac{\delta}{|\mathcal{H}|} \cdot |\mathcal{H}| = \delta$.

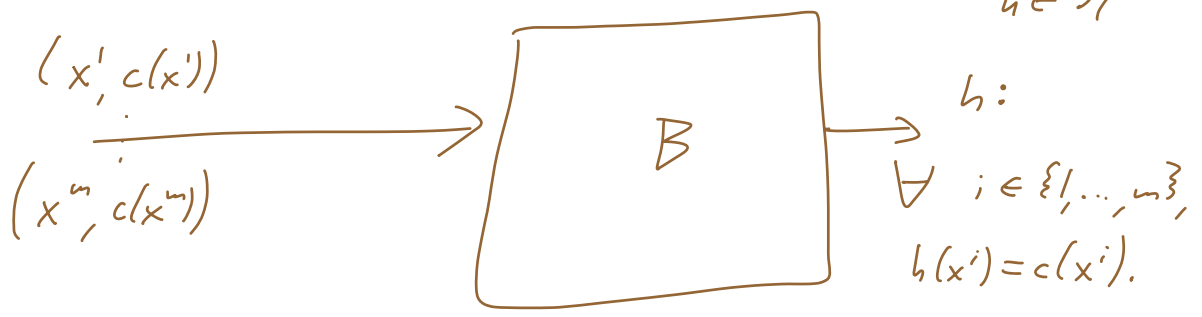
Interpreting/using this: (CHF)

Def: Fix $\mathcal{C}, \mathcal{H}$. A consistent hyp. finder for $\mathcal{C}$ using $\mathcal{H}$ is an alg B w/ following prop.:

if you give B a sample of m ex.

$\left.\begin{matrix}(x^1, c(x^1)) \\ \vdots \\ (x^m, c(x^m))\end{matrix}\right\}$ all lab. by same fixed $c \in \mathcal{C}$, (any $c \in \mathcal{C}$)

then B outputs an $h \in \mathcal{H}$ consist. with them:



Combining above: if have a CHF, gives a PAC learning alg for $\mathcal{C}$ using $\mathcal{H}$:

draw $m = \frac{1}{\varepsilon} \cdot (\ln|\mathcal{H}| + \ln\frac{1}{\delta})$ samples,

feed to CHF, output the h it outputs.

Applic.: let $\mathcal{C}$ = all mon disj over $\{0,1\}^n$.

- Elim alg: OLMB m.b. n.
- OLMB → PAC conv: $m \approx \frac{n}{\varepsilon} \cdot \ln\left(\frac{n+1}{\delta}\right)$ ex.

In fact, elim alg. is a CHF for $\mathcal{C}$ using $\mathcal{H} = \mathcal{C}$.

So... $m = \frac{1}{\varepsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta})$ is enough

$|\mathcal{H}| = |\mathcal{C}| = 2^n$

so $\hookrightarrow \frac{1}{\varepsilon}\left(\ln \frac{2^n}{\delta}\right) = O\left(\frac{1}{\varepsilon}\left(n + \ln \frac{1}{\delta}\right)\right)$

ex is enough.

(Non-mon disj: also ok; $\mathcal{H} = \mathcal{C} =$ all 3^n nonmon. disj

also $O\left(\frac{1}{\varepsilon} \cdot (n + \ln 1/\delta)\right)$.)

Discussion:

So,

- IF you have a CHF for $\mathcal{C}$ using $\mathcal{H}$, get a PAC learner.

We'll see: sometimes ↑ the job of a CHF is comput. hard.

- CRUCIAL that we have a fixed a priori $\mathcal{H}$ to apply this.

Bogus CHF: given sample $S =$ $(x', c(x'))$ ⋮ $(x^m, c(x^m))$

let h be a look-up table

$h(x) =$ " if $x = x'$, output

∀ if $x = x^2$, output

etc "

↳ Bogus b/c didn't have a fixed $\mathcal{H}$.

Occam's Razor: we should prefer short explanations.

Say you have data set of m lab. ex. from $EX(c, \mathcal{D})$.

h "needs to explain" m label bits.

$\mathcal{H}$ is of size $|\mathcal{H}|$: $\log|\mathcal{H}|$ bits

can encode any $h \in \mathcal{H}$.

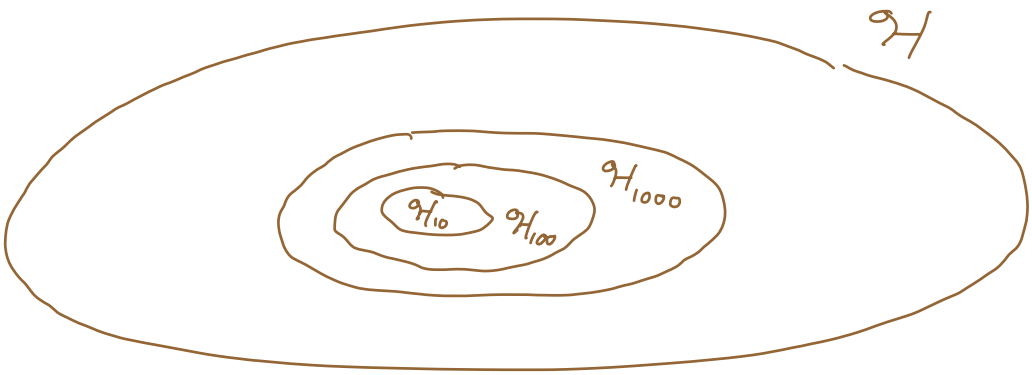
So if $m \gg \log|\mathcal{H}|$, any $h \in \mathcal{H}$ is a succinct explan. of the data

$$m \geq \frac{1}{\varepsilon}\left(\ln|\mathcal{H}| + \ln\left(\tfrac{1}{\delta}\right)\right)$$

Can "boil down" CHF stuff to small-m regime.

Prev thm: $\forall m$ (# samples), CHF gives some $h \in \mathcal{H}$.

Maybe if m small, we can find a consistent h in subclass $\mathcal{H}_m$ of $\mathcal{H}$?



If so, can take advantage of it.

Another version of thm above:

Thm

"Occam's razor, cardinality version":

Fix m, $\mathcal{C}$, $\mathcal{H}$. Suppose there's a subset $\mathcal{H}_m \subseteq \mathcal{H}$ s.t. $m \geq \frac{1}{\varepsilon}\left(\ln|\mathcal{H}_m| + \ln\frac{1}{\delta}\right)$

s.t. given any set S of m lab. ex. acc. to some $c \in \mathcal{C}$, there's an $h \in \mathcal{H}_m$ consistent with the m ex.

If L is any alg. that finds such an $h \in \mathcal{H}_m$ when given a sample of size m, then

L run on m ex from $EX(c, \mathcal{D})$ is an (ε, δ)-PAC learner.

Pf: as before with $\mathcal{H}_m$ in place of $\mathcal{H}$.