Last time: Analysis of AdaBoost: proved
\[ \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\epsilon[H(X)] \neq y_i} \leq \sqrt{1 - \frac{2\gamma}{e}}. \]

\[ \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\epsilon[H(X)] \neq y_i} \leq \sqrt{1 - \frac{2\gamma}{e}}. \]

\[ \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\epsilon[H(X)] \neq y_i} \leq \sqrt{1 - \frac{2\gamma}{e}}. \]

• Started learning w/noise. RCN: max. noise. \( S_1, S_2 \)

Today: • Prove this stated last time: \( n > \frac{2\epsilon}{1+\epsilon} \) \( \Rightarrow \) no PAC learning.

• RCN: build intuition, sketch RCN-Yolov2.

For mon. conj. motiv. for new learn model: SQ.

\[ \mathbb{C}: \text{ distinct:} \]

\[ c_1 \cap c_2 \]

Thm: \( \mathbb{C} \) distinct, max. noise rate \( n > \frac{2\epsilon}{1+\epsilon} \) \( \Rightarrow \) can't PAC learn to accuracy conf \( \delta < \frac{1}{2} \).

\[ \text{Admin: Rocco no OH on Fri; will have makeup OH next week.} \]

\[ \text{Pf: Let's consider following } \theta: \]

\[ \theta(u) = \theta(w) = \epsilon, \]

\[ \theta(u) = 1 - 2\epsilon. \]

\[ \text{Two adv. strategies:} \]

\[ A_1: (\text{for } c_1): \frac{m}{2} \text{ prob. on } (u-), \frac{m}{2} \text{ prob. } (w+). \quad 1-m: \text{ no noise.} \]

Learner sees:

\[ (u,+) \text{ w.p. } \frac{m}{2}, \]

\[ (w,+) \text{ w.p. } (1-2\epsilon)(1-m), \]

\[ (w,-) \text{ w.p. } \epsilon(1-m), \]

\[ (w,+) \text{ w.p. } \frac{m}{2}. \]

sum to \( 1-m \)

noise

\[ A_2: (\text{for } c_2): \frac{m}{2} \text{ prob. on } (u+), \frac{m}{2} \text{ prob. } (w-). \quad 1-m: \text{ no noise.} \]

Learner sees:

\[ (u,-) \text{ w.p. } \epsilon(1-m), \]

\[ (u,+) \text{ w.p. } \frac{m}{2}, \]

\[ (w,-) \text{ w.p. } \epsilon(1-m), \]

\[ (w,+) \text{ w.p. } \frac{m}{2}. \]

For \( m > \frac{2\epsilon}{1+2\epsilon} \): same dist: b's.

\[ \frac{m}{2} = \epsilon(1-m), \quad m(1+2\epsilon) = 2\epsilon, \]

\[ m = \frac{2\epsilon}{1+2\epsilon}. \]
No matter #ex, distr. identical, learner’s h either error ≥ ε under c, or ≥ ε under c₂. Can’t have error < ε w.p. > ½.

What can you do in presence of mal. noise? Hope don’t get noisy ex. A → PAC setting.

Best general approach: Sps abs for c in noise-free.

- Run A K times, fresh ex. each time. Get h₁, ..., hₖ.
- Test h₁, ..., hₖ on new ex, & output one making fewest errors.

Idea: if n suff. low, maybe one of K runs got no noisy ex. & w/ p resulting hᵢ, acc.

> acc hᵢ, “shouldn’t look too bad”.

Slightly more detail: say your PAC abs was m ex to (ε/₂, δ/₂) - PAC when no noise.

\[ \Pr[\text{one run has no noisy ex}] = (1-m)^m \approx \frac{e^{-m}}{e} \]

\[ \Pr[\text{one of K runs had no noise}] \approx (1 - \frac{m}{K})^K \approx \frac{e^{-m}}{e} \]

i.e.

\[ \Pr[\text{some run no noise}] \approx 1 - \frac{1}{e} \approx 0.6 \]

etc. etc.
Next subtopic: RCN.

Recall: RCN rate \( \eta < \frac{1}{2} \).

- \( \text{ex} \ (x, c(x)) \sim \text{EX}(c, \theta) \).
- w.p. \( 1-\eta \) learner gets \( (x, c(x)) \).
- w.p. \( \eta \) " " \( (x, \overline{c(x)}) \).

This oracle: \( \text{EX}^\eta(c, \theta) \).

RCN rate \( m \).

Def: Alg A PAC learns \( c \) / RCN if \( \forall c \in C \), \( \forall \text{dist} \)\( D \)

\( \forall \eta < \frac{1}{2}, \forall \varepsilon, \delta > 0 \) Given \( \text{EX}^\eta(c, \theta) \), \( \varepsilon, \delta \), alg A
outputs \( h \) s.t.

\[ \Pr \left[ h(x) \neq c(x) \right] \leq \varepsilon \quad \text{w.p.} \geq 1-\delta. \]

Eff alg in setting: runs in time

\[ \text{poly}(\eta, \text{size}(c), \frac{1}{\delta}, \log \frac{1}{1-\delta}). \]

- 0. Sps \( h \) is s.t. \( \Pr \left[ h(x) \neq c(x) \right] \leq \eta \). (no noise!)

Then "observed error rate in pres. of \( \eta \) noise," \( \Pr \left[ h(x) \neq y \right] \) is

\[ \xi \cdot (1-\eta) + \eta \cdot (1-\xi) = \xi + \eta - 2\eta \xi = \eta + \xi (1-2\eta). \]

\( \uparrow \)

smaller \( \xi \): better hyp appears.

We'll assume \( \eta \) known to learner. If unknown:

- try \( \eta = 0, \delta, 2\delta, 3\delta, \ldots \)
- test hyps.
- testing is possible.
Now: consider \( \text{PAC learning} \) with \( \text{RCN} \).

Recall our PAC (noiseless) alg for mon conj:

1. Start w/ hyp.: \( h(x) = x_1 \wedge x_2 \wedge \ldots \wedge x_n \).
2. Draw data set \( \text{from } \mathcal{EX}(\mathcal{D}) \).
3. Elim: For each pos ex in data set, if \( x_i = 0 \) in that ex, remove \( x_i \) from \( h \).
4. Since data noise-free, resulting \( h \) is cons mon conj; output it.

\( \text{RCN} \): not going to work. Data set may have an ex labeled \( - \), actually \( + \). Causes you to remove \( x_i \)'s that should be kept; incurs error.

Hi-level problem: any indiv ex. can't be trusted.
Hi-level fix: exploit statistics of whole data set.

Our approach: \( i = 1 \ldots n \)

Define \( p_i := \Pr_{x \sim \mathcal{D}}[c(x) = 1 \wedge x_i = 0] \).

(If \( x_i \) in \( c \), then \( p_i = 0 \).

Intuition: each \( x_i \) not in \( c \) "adds" at most \( p_i \) to error of a hyp. \( h \) which erroneously includes \( x_i \).

Can learn successfully if we can identify all \( i \) s.t. \( p_i \geq \varepsilon/n \).
My \( h \) won't include these. If this is \( h \), it misses \( \leq \varepsilon \) of many \( x_i \)'s, each costs \( \leq \varepsilon/n \) error; error of \( h \) c.e.

To learn to acc \( \varepsilon \), suff to be able to est. each \( p_i \) to \( \pm \varepsilon/2n \).
If had $EX(c, \theta)$: easy: draw $(x, c(x))$ & count obs.

Prac. s.t. $c(x) = 1$, $x_i = 0$.

$CB \Rightarrow O\left(\frac{n^2 \cdot \ln\left(\frac{2n}{\delta}\right)}{\varepsilon^2}\right)$ : w.p. $\geq 1 - \frac{\delta}{2}$, all $n$

estimates $i = 1, \ldots, n$ are $\pm \frac{\varepsilon}{2n}$ acc.

We're in $EX^\infty(c, \theta)$.

$(x, y) \sim \theta$

$y = c(x)$ indep.

w.p. $1 - \frac{\delta}{2}$

Want to estim.

$p_i = \frac{Pr[c(x) = 1 | x_i = 0]}{Pr[c(x) = 1 | x_i = 0] \cdot Pr[c(x) = 1 | x_i = 0]} = Pr[C_B].$

$Pr[B] = \frac{Pr[x_i = 0]}{Pr[x_i = 0] \cdot Pr[x_i = 0]}$.

"$x_i = 0$ unaffected by noise"

Can est. $Pr[C_B]$ in.

If $Pr[C_B]$ small ($< \frac{\varepsilon}{4n}$)

then know $p_i \leq \frac{\varepsilon}{4n}$.

don't need to est. $Pr[A \mid B]$.

"0" is $\pm \frac{\varepsilon}{2n}$ acc est. of $p_i$.

So can assume $Pr[C_B]$ not too small: $> \frac{\varepsilon}{4n}$.

This will let me est. $Pr[A \mid B]$.

Recall we have only $EX^\infty(c, \theta)$... it's Noise affects $A$ in predictable way.
Define \( g_i := \Pr[A_i \mid B] = \Pr[c(x_i) = 1 \mid x_i = 0] \).

Consider
\[
\star := \Pr\{y = 1 \mid x_i = 0\} = g_i (1 - m) + (1 - g_i) m
\]
\((x, y) \sim \mathcal{X} \cap \mathcal{D}) = n + g_i (1 - 2m).

Can estimate \( \star \). Sps know \( \star \).

Sps know \( n \).

\[
\begin{aligned}
\star &= n + g_i (1 - 2m) \\
\frac{\star - n}{1 - 2m} &= g_i. \\
\end{aligned}
\]

So can solve \( \star, \) est of \( g_i \).

Est. of \( g_i = \Pr[A \mid B] \)
gets \( \star \) by est of \( \Pr[B] \)
gives est of \( \Pr[A \lor B] = \pi \).

What happened? Got learning abt estimating probabilities.
• Wanted a prob. def'd in terms of noiseless examples, but in noisy world

• Decomposed desired L (noiseless) into A + B. B: unaffected by noise.
  A: totally aff. by noise predictably

Combine to get original desired prob.

Read some Chap. 5 KV book.
Next time: talk abt SQ model.

See: Any c.c. C learnable in SQ model is PAC learnable w/ RCN.