Last time:

- finish proof of concept: 3-stage booster to go from 40% error to 35.2% error
- discuss boosting-by-filtering vs boosting over a fixed sample
- start discussion + analysis of AdaBoost (alg. for)

Today:

- analysis of AdaBoost: bound on error rate of its final hyp. over its input sample
- start unit on PAC learning in presence of noise
  - General framework, specific noise models:
    - malicious noise
    - random classification noise

Note: PS 4 due Tues Nov. 11

Questions?

Look at AdaBoost... (with a twist!)

Note: AdaBoost update rule: like WM:

WM	AdaBoost
• expert i	• i^{th} example in S
• t^{th} trial	• t^{th} run of weak learner
• pred. of a_t	• $h_t(x^i)$
• expert i's wt at	• $\vartheta_t(x^i)$
• mult. update	• mult. update

Twist: increase / decrease w_t on mistakes

Interpret. of a/g:

Suppose $h_t(x^i) = y_i$. $\Rightarrow$ $h_t(x^i) \cdot y_i = 1$.

$\Rightarrow \exp(-\alpha_t y_i h_t(x_i)) = \exp(-\alpha_t)$

$= \exp(-\frac{1}{2} \ln \frac{1-\varepsilon_t}{\varepsilon_t})$

$= \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}$.

So in stage t, (ignoring Z_t normaliz.)

any x^i s.t. $h_t(x^i) = y^i$ has its wt

• by $\sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}} < 1$: wt goes down

---

OTOH, Suppose $h_t(x^i) \neq y_i$. $\Rightarrow$ $h_t(x^i) \cdot y_i = -1$.

$\Rightarrow \exp(-\alpha_t y_i h_t(x_i)) = \exp(\alpha_t)$

$= \exp(\frac{1}{2} \ln \frac{1-\varepsilon_t}{\varepsilon_t})$

$= \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} > 1$; wt goes up.

Thm: Suppose run Ada. for T stages. Then its final hyp. H makes errors on $\leq$

the substance $\prod_{t=1}^{T} \sqrt{1-4\gamma_t^2} \leq \exp(-2 \sum_{t=1}^{T} \gamma_t^2)$

frac. of the m points in S, where $\frac{1}{2} - \varepsilon_t = \gamma_t$

γ_t is advantage of t^{th} weak hyp. h_t under D_t

easy: $(1-4\gamma_t^2)^{\frac{1}{2}} \le \exp((-4\gamma_t^2)\cdot\frac{1}{2}) = \exp(-2\gamma_t^2)$ $\quad (1-x \le e^{-x})$

Cor: If each $\gamma_t \ge \gamma$, can run Ada. for $T \ge \frac{1}{2\gamma^2}\cdot\ln(1/\varepsilon)$ stages, & achieve H s.t. wrong on $\le \varepsilon$ frac. of pts in S.

(solve $\varepsilon \ge \exp(-2T\gamma^2)$).

(optimal!)

Thm follows from 3 claims:

$H(x) = \text{sign}(f(x))$, $f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$. (real value)

Claim 1: $\frac{1}{m}\left|\{i \in [m] : H(x^i) \ne y_i\}\right| \le \frac{1}{m}\sum_{i=1}^{m} \exp(-y_i f(x_i))$ (frac. of errors that H makes)

Claim 2: $\frac{1}{m}\sum_{i=1}^{m} \exp(-y_i f(x_i)) = \prod_{t=1}^{T} Z_t$.

Claim 3: For each $t \in T$, $Z_t = \sqrt{1-4\gamma_t^2}$.

Proofs:

Claim 1: $\frac{1}{m}\left|\{i \in [m] : H(x^i) \ne y_i\}\right| \le \frac{1}{m}\sum_{i=1}^{m} \exp(-y_i f(x_i))$

Pf: We'll show: for any i s.t. $H(x^i) \ne y_i$, have $\exp(-y_i f(x_i)) \ge 1$. (This does it b/c other is contrib. 0 on LHS, >0 on RHS.)

Fix an i s.t. $H(x^i) \neq y_i$, i.e. $\operatorname{sign}(f(x^i)) \neq y_i$.
Suppose $y_i = +1$, so $f(x^i) < 0$; this means
$-y_i f(x^i) > 0$, so $\exp(-y_i f(x^i)) > 1$. ▨

Claim 2: $\frac{1}{m}\sum_{i=1}^{m} \exp(-y_i f(x_i)) = \prod_{t=1}^{T} Z_t$.

Pf: Consider $\vartheta_{T+1}(x^i)$:

$$\vartheta_{T+1}(x^i) = \frac{\exp(-\alpha_T y_i h_T(x^i))}{Z_T} \cdot \vartheta_T(x^i)$$

$$= \frac{\exp(-\alpha_T y_i h_T(x^i))}{Z_T} \frac{\exp(-\alpha_{T-1} y_i h_{T-1}(x^i))}{Z_{T-1}} \cdot \vartheta_{T-1}(x^i)$$

$$= \cdots$$

$$= \frac{\exp(-\alpha_T y_i h_T(x^i))}{Z_T} \cdots \frac{\exp(-\alpha_1 y_i h_1(x^i))}{Z_1} \cdot \vartheta_1(x^i) \qquad (= \tfrac{1}{m})$$

$$= \frac{1}{m} \cdot \frac{\exp\left(-\sum_{t=1}^{T} \alpha_t y_i h_t(x^i)\right)}{\prod_{t=1}^{T} Z_t}$$

Sum over $i = 1, \ldots, m$: know $\sum_{i=1}^{m} \vartheta_{T+1}(x^i) = 1$ ", so

$$1 = \sum_{i=1}^{m} \frac{1}{m} \cdot \frac{\exp\left(-\sum_{t=1}^{T} \alpha_t y_i h_t(x^i)\right)}{\prod_{t=1}^{T} Z_t}, \text{ i.e.}$$

$$\prod_{t=1}^{T} Z_t = \sum_{i=1}^{m} \frac{1}{m} \cdot \exp\left(-\sum_{t=1}^{T} \alpha_t y_i h_t(x^i)\right)$$

$$= \sum_{i=1}^{m} \frac{1}{m} \exp\left(-y_i \sum_{t=1}^{T} \alpha_t h_t(x^i)\right) =$$

$$= \frac{1}{m} \cdot \sum_{i=1}^{m} \exp(-y_i f(x^i)) \qquad \blacksquare$$

(Now, in Claim 3, use def of α_t)

Claim 3: For each $t \in T$, $Z_t = \sqrt{1-4\gamma_t^2}$.

Pf: Fix a $t \in [T]$.

a, b, c:
$\frac{a}{a+b+c}, \frac{b}{a+b+c}, \frac{c}{a+b+c}$
sum to 1

$$Z_t = \sum_{i=1}^{m} \vartheta_t(x^i) \exp(-\alpha_t y_i h_t(x^i))$$

$$= \sum_{i : h_t(x^i) \neq y_i} \vartheta_t(x^i) \exp(-\alpha_t y_i h_t(x^i)) = A$$

$$+ \sum_{i : h_t(x^i) = y_i} \vartheta_t(x^i) \exp(-\alpha_t y_i h_t(x^i)) = B$$

We saw: for an i in A, have

$$\exp(-\alpha_t y_i h_t(x^i)) = \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}. \quad \text{So } A =$$

$$\sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} \cdot \sum_{i : h_t(x^i) \neq y} \vartheta_t(x^i) = \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} \cdot \varepsilon_t$$

$$= \sqrt{\varepsilon_t(1-\varepsilon_t)}.$$

We also saw: for an i in the B sum have

$$\exp(-\alpha_t y_i h_t(x^i)) = \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}. \quad \text{So } B =$$

$$\sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}} \cdot \sum_{i : h_t(x^i) = y} \vartheta_t(x^i) = \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}} \cdot (1-\varepsilon_t)$$

$$= \sqrt{\varepsilon_t(1-\varepsilon_t)}\,.$$

So $Z_t = A + B = 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$

$$= \sqrt{(2\varepsilon_t)(2(1-\varepsilon_t))}$$

Since $\gamma_t = \frac{1}{2} - \varepsilon_t$, have $\varepsilon_t = \frac{1}{2} - \gamma_t$

$2\varepsilon_t = 1 - 2\gamma_t$ $\qquad$ $2(1-\varepsilon_t) = 1 + 2\gamma_t$, $\forall$

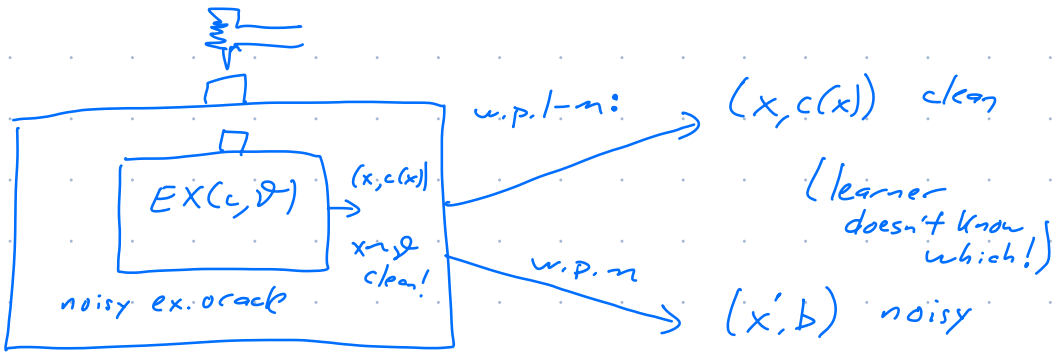
$$= \sqrt{(1-2\gamma_t)(1+2\gamma_t)} = \sqrt{1-4\gamma_t^2}\,.$$

Learning in Presence of Noise

Motivation: obvious: want more robust models.

Our noise framework: still is a target $c \in \mathcal{C}$, an unknown dist $\mathcal{D}$.

We assume learner has access to a noisy example oracle: noise rate $\eta < \frac{1}{2}$



Diff. specific noise models: diff. assump. about (x', b).

Goal: unchanged: still want to, w.p. $1-\delta$, output h s.t. $err_D[h, c] \leq \varepsilon$.
