PS 4 out today, due 11/24
Midterm solutions on EdStem, grading in-progress

Last few times...

Boosting

- γ-Weak Learning Assumption
- Connection to linear separability via VNMMT
- Boosting algorithm based on Hedge
- Out-of-sample performance analysis
  - VC dimension of $\text{Majority}_{\mathcal{H},T}$

$$O(VCdim(\mathcal{H}) \cdot T \log T)$$

Today: Margins theory

Boosting algorithm

Input: training data $S = (x_1, y_1), \ldots, (x_n, y_n) \in (\mathcal{X} \times \{-1,1\})^n$, WL

Hyperparameters: $\eta > 0$, $T \in \mathbb{N}$

Init: $p_1 = (\frac{1}{n}, \ldots, \frac{1}{n}) \in \Delta^{n-1}$

For $t = 1, 2, \ldots, T$:

- Call WL w/ training data re-weighting $p_t \in \Delta^{n-1}$, get back $h_t : \mathcal{X} \to \{-1,1\}$
- Define loss vector $\ell_t \in \{0,1\}^n$

$$\ell_{t,i} = \mathbb{1}\{h_t(x_i) = y_i\}$$

- Update: $p_{t+1} \in \Delta^{n-1}$

$$p_{t+1,i} = \frac{p_{t,i} \cdot \exp(-\eta \ell_{t,i})}{Z_{t+1}}$$

Return $\hat{h} = \text{sign} \circ \hat{g}$

where

$$\hat{g} = \frac{1}{T} \sum_{t=1}^{T} h_t$$

Thm Assume P in dist. over $\mathcal{X} \times \{-1,1\}$ s.t. $S \sim P^n$ satisfies γ-W.L.A. — i.e., $\exists$ WL for S with advantage γ

[In particular, $h_t \in \mathcal{H}$ satisfies $\text{acc}_S(h_t, p_t) \geq \frac{1+\gamma}{2}$]

Then boosting alg. w/ WL on input $S \sim P^n$ and $\eta = \gamma/3$ returns $\hat{h}$ s.t. with prob $\geq 1 - \delta$,

$$\text{err}(\hat{h}; P) \leq \underbrace{\exp(-\gamma^2 T/2)}_{\text{(I)}} + \underbrace{O\left(\sqrt{\frac{V \log(n/V)}{n} + \frac{\log 1/\delta}{n}}\right)}_{\text{(II)}}$$

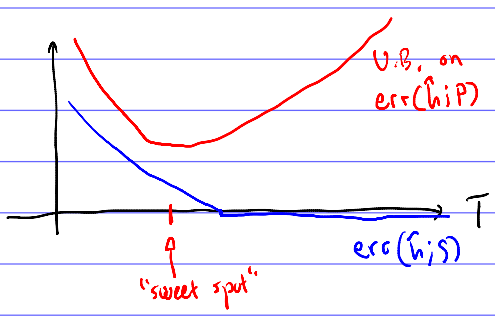
where $V = O(\text{VCdim}(\mathcal{H}) \cdot T \log T)$

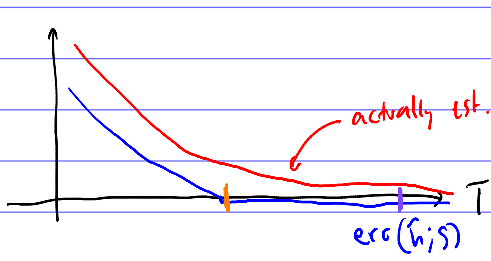
Observation:

(I): decreases w/ T — exponentially

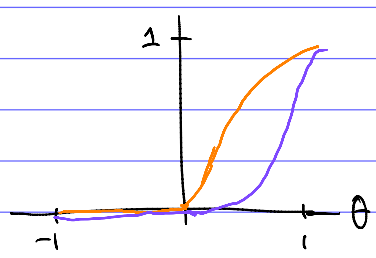
(II): increases w/ T — sqrt-ly

$\therefore \exists$ some "sweet spot" for T.





$\text{err}_\theta(\hat{h}; S)$ on training data



Boosting's output: $\hat{h} = \text{sign} \circ \hat{g}$

$$\hat{g} \in \text{conv}(\mathcal{H})$$

Have $\hat{g}(x) \in [-1,1]$

Define, for $(x,y) \in \mathcal{X} \times \{-1,1\}$, the margin achieved by $g : \mathcal{X} \to [-1,1]$ on (x,y) to be $y \cdot g(x) \in [-1,1]$

For $\theta \in [0,1]$, define for $f : \mathcal{X} \to [-1,1]$

$$\text{err}_\theta(f; S) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{y_i f(x_i) \leq \theta\}$$ "θ-margin error rate on S"

$$\text{err}_\theta(f; P) = \Pr_{(x,y) \sim P}(y f(x) \leq \theta)$$

Conjecture: larger margins $\Longrightarrow$ better generalization

Algorithmically: If S satisfies γ-WLA, then there's some $\eta > 0$, $T \in \mathbb{N}$ s.t.

$$\text{err}_\theta(\hat{h}; S) = 0 \quad \text{for } \theta = \gamma/3$$

Defn $\text{avg}_K(\mathcal{H}) = \left\{ \frac{1}{K}\sum_{t=1}^{K} h_t \;\middle|\; (h_1,\dots,h_K) \in \mathcal{H}^K \right\}$

uniform averages of K hypotheses from $\mathcal{H}$

Thm For any $\mathcal{H} \subseteq \{-1,1\}^{\mathcal{X}}$, any dist P on $\mathcal{X}\times\{-1,1\}$, any $S \in (\mathcal{X}\times\{-1,1\})^n$, any $K \in \mathbb{N}$, any $\theta \in [0,1]$, and any $f \in \text{conv}(\mathcal{H})$:

$$\text{err}(\text{sign}\circ f, P) \le \text{err}_\theta(f;S) + 4\exp(-\theta^2 K/8) + \max_{g \in \text{avg}_K(\mathcal{H})} \left(\text{err}_{\theta/2}(g;P) - \text{err}_{\theta/2}(g;S) \right)$$

← to make this $\le t/n$, $K := \frac{8\log(4n)}{\theta^2}$

$|\text{avg}_K(\mathcal{H})| \le |\mathcal{H}|^K$ ⇒ $\frac{\log|\text{avg}_K(\mathcal{H})|}{n} \le \tilde{O}\left(\frac{t}{\theta^2}\log(|\mathcal{H}|)\right)/n$

This works for every $f \in \text{conv}(\mathcal{H})$

In particular, for every $f \in \text{avg}_T(\mathcal{H})$ even if $T \gg K$

Proof idea:

Every $f \in \text{conv}(\mathcal{H})$ can be approximated by some $g_f \in \text{avg}_K(\mathcal{H})$ in a suitable sense

e.g. $f = 0.1h_1 + 0.5h_2 + 0.39h_3 + \cdots + 0.0001h_{1000}$

Proposition: For any $\mathcal{H} \subseteq \{-1,1\}^{\mathcal{X}}$, any r.v. Z taking values in $\mathcal{X}$, any $\varepsilon > 0$ (will use $\varepsilon = \theta/2$), any $f \in \text{conv}(\mathcal{H})$, $\exists\, g_f \in \text{avg}_K(\mathcal{H})$ s.t.

$$\Pr\left(|f(Z) - g_f(Z)| > \varepsilon\right) \le 2\exp(-\varepsilon^2 K/2).$$

Proof is based on probabilistic method.

Fix $f \in \text{conv}(\mathcal{H})$.

1. Define Random function G taking values in $\text{avg}_K(\mathcal{H})$.
   - Write $f = \sum_{i=1}^{d} \alpha_i h_i$ where $d \in \mathbb{N}$, $\alpha = (\alpha_1,\dots,\alpha_d) \in \Delta^{d-1}$, $(h_1,\dots,h_d) \in \mathcal{H}^d$.
   - Dist. of G is given in terms of this representation for f.

   Let $J_1,\dots,J_K$ be iid random indices
   $\Pr(J_i = j) = \alpha_j \quad \forall j \in [d]$
   $G = \frac{1}{K}\sum_{i=1}^{K} h_{J_i}$

2. For any fixed $x \in \mathcal{X}$, will show $G(x) \approx f(x)$ for most realizations of G

$$\Pr\left(|G(x) - f(x)| > \varepsilon\right) \le 2\exp(-\varepsilon^2 K/2)$$

$G(x) = \frac{1}{K}\sum_{i=1}^{K} h_{J_i}(x)$ average of K iid random variables

$$E[G(x)] = \frac{1}{K}\sum_{i=1}^{K} E[h_{J_i}(x)]$$

$$E[h_{J_i}(x)] = \sum_{j=1}^{d} \Pr(J_i = j)\cdot h_j(x) = \sum_{j=1}^{d} \alpha_j h_j(x) = f(x)$$

3. Replace x w/ r.v. Z & take expectation

$$E\left[\Pr\left(|G(Z) - f(Z)| > \varepsilon \mid Z\right)\right] \le 2\exp(-\varepsilon^2 K/2)$$

4. Switch order of expectations (Fubini's thm)

$$E\left[\Pr\left(|G(Z) - f(Z)| > \varepsilon \mid Z\right)\right]$$
$$= E\left[\mathbb{1}\{|G(Z) - f(Z)| > \varepsilon\}\right]$$
$$= E\left[E\left[\mathbb{1}\{|G(Z) - f(Z)| > \varepsilon\} \mid G\right]\right] \le 2\exp(-\varepsilon^2 K/2)$$

So there's some realization of G, call it $g_f \in \text{avg}_K(\mathcal{H})$, s.t.

$$E\left[\mathbb{1}\{|G(Z) - f(Z)| > \varepsilon\} \mid G = g_f\right] \le 2\exp(-\varepsilon^2 K/2)$$

$\Pr(|g_f(Z) - f(Z)| > \varepsilon)$