Midterm grades on gradescope now

- Regrade requests: if your answer is 100% correct but we deducted points, can submit regrade request on gradescope before 3PM tomorrow.

PS4 due in 2 weeks − 2 days

Last time: - Started Margins theory
- proved approximation result (Kind of like ε-cover)

Today: - Finish Margins theory
- Boosting as an optimization alg.
- Linear learning paradigm

Theorem For any $H \subseteq \{-1,1\}^X$
any dist. P over $X \times \{-1,1\}$
any data set $S \in (X \times \{-1,1\})^n$
any $K \in \mathbb{N}$
any $\theta \in (0,1)$
any $f \in \text{conv}(H)$

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

to make this $\frac{1}{n}$, choose $K = \frac{8 \log 4n}{\theta^2}$

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

If $|H| < \infty$
then $|\text{avg}_K(H)| = |H|^K$
So by Hoeffding + union bound, if $S \sim P^n$
this term is $\leq \sqrt{\frac{K \ln|H| + \log\frac{1}{\delta}}{2n}}$ w.p. $\geq 1-\delta$.

$$= \sqrt{\frac{8(\log(4n)\cdot\log|H|}{2\theta^2 n} + \frac{\log 1/\delta}{2n}}$$

Proposition For any $H \subseteq \{-1,1\}^X$
any r.v. Z (taking values in X)
any $\varepsilon > 0$

For each $f \in \text{conv}(H)$, there exists $g_f \in \text{avg}_K(H)$ s.t.

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

quality of approximation

fraction of Z's does approx the fail

Key idea: $\theta/2$-approx is good enough b/c
we're comparing 0-margin errors
to $\theta/2$-margin errors

LHS: $\text{err}_0(f; P)$

$$\leq \text{err}_{\theta/2}(g_f; P) + 2\exp(-\theta^2 K/8)$$

RHS: $\text{err}_\theta(f; S)$

$$\geq \text{err}_{\theta/2}(g_f; S) - 2\exp(-\theta^2 K/8)$$

If $y \cdot f(x) \leq 0$ & $|f(x) - g_f(x)| \neq \theta/2$
then $y \cdot g_f(x) \leq \theta/2$

$$\mathbb{1}\{y f(x) \leq 0\} \leq \mathbb{1}\{|f(x) - g_f(x)| > \theta/2\} + \mathbb{1}\{y g_f(x) \leq \theta/2\}$$

Boosting as an optimization algorithm

Think of the hypothesis class $\mathcal{H} = \{h^{(1)}, \ldots, h^{(d)}\}$ used by WL as feature map $\phi : \mathcal{X} \to \mathbb{R}^d$

$$\phi(x) = (h^{(1)}(x), \ldots, h^{(d)}(x))$$

Boosting finds LTF $\text{sign}(\langle \phi(x), \lambda \rangle)$ for some $\lambda = (\lambda^{(1)}, \ldots, \lambda^{(d)}) \in \mathbb{R}^d$

$$L(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \exp\left(-y_i \langle \phi(x_i), \lambda \rangle\right)$$

Coordinate descent: repeatedly pick some $j \in [d]$ and update $\lambda^{(j)}$, leaving all other $\lambda^{(j')}$ alone

Init: $\lambda_1 = (0, \ldots, 0) \in \mathbb{R}^d$

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

$$\frac{\partial L}{\partial \lambda^{(j)}}(\lambda_t) = -\frac{1}{n} \sum_{i=1}^{n} \exp\left(-y_i \langle \phi(x_i), \lambda_t \rangle\right) y_i h^{(j)}(x_i)$$

$$= -\frac{1}{n} Z_t \sum_{i=1}^{n} \underbrace{\frac{\exp\left(-y_i \langle \phi(x_i), \lambda_t \rangle\right)}{Z_t}}_{p_{t,i}} y_i h^{(j)}(x_i)$$

p_t - correlation of $h^{(j)}(x_i)$ w/ y_i

γ-WLA says WL can find some $h^{(j)}$ s.t. this is at least γ

Pick $j_t \in [d]$ s.t. $\frac{\partial L}{\partial \lambda^{(j_t)}}(\lambda_t)$ is noticeably negative

Update $\lambda_{t+1}^{(j_t)} = \lambda_t^{(j_t)} + \eta$

Update of λ_t to λ_{t+1} corresponds to update of p_t to p_{t+1}

$$p_{t+1,i} \propto \exp\left(-y_i \langle \phi(x_i), \lambda_{t+1} \rangle\right)$$

$$= \exp\left(-y_i \langle \phi(x_i), \lambda_t \rangle - y_i h^{(j_t)}(x_i) \cdot \eta\right)$$

$$\propto p_{t,i} \exp\left(-\eta y_i h^{(j_t)}(x_i)\right)$$

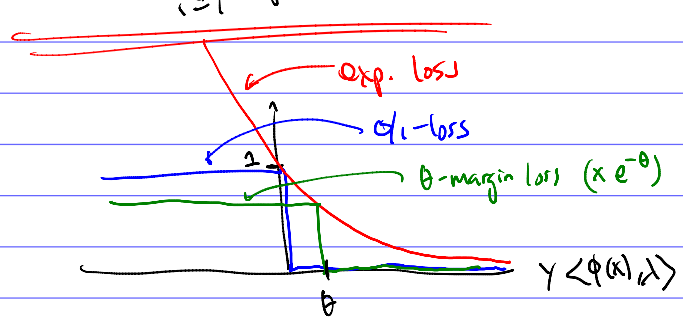
↳ some update to example weights in boosting.

Interpreting $L(\lambda)$

$$L(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i \langle \phi(x_i), \lambda\rangle\right)$$

vs $\mathrm{err}_0(x \mapsto \langle \phi(x), \lambda\rangle)$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{ y_i \langle \phi(x_i), \lambda\rangle \leq 0 \}$$



Boosting "naturally" is looking for λ achieving small err_θ

General paradigm for ML: linear learning

① Design $\phi : \mathcal{X} \to \mathbb{R}^d$ (perhaps implicitly like in boosting)

② Training: use training data & opt algo. to find $\lambda \in \mathbb{R}^d$ that (approx) minimizes

$$\frac{1}{n}\sum_{i=1}^{n} \ell(y_i \langle \phi(x_i), \lambda\rangle)$$

③ Return LTF $\quad x \longmapsto \mathrm{sign}(\langle \phi(x), \lambda\rangle)$

Can we do this w/ ℓ = 0/1 loss fn?

Thm The following problem is NP-hard (i.e. probably intractable)

Given: $S \subset \{0,1\}^{2d} \times \{-1,1\}$ of n examples, $k \in \mathbb{N}$
Decide: $\exists\, \hat{h} \in \mathrm{LTF}_d$ s.t. $\mathrm{err}(\hat{h}, S) \leq k/n$?

In worst case, w/ 0/1-loss, linear learning is computationally hard.

Coping w/ intractability

Change the problem

① Settle for $\hat{h}$ s.t.

$$\mathrm{err}(\hat{h}; P) \leq 2 \times \min_{h \in \mathrm{LTF}_d} \mathrm{err}(h, P) + \varepsilon$$

② Settle for $\hat{h}$ s.t.

$$\mathrm{err}(\hat{h}; P) \leq \min_{g \in \mathcal{C}} \mathrm{err}_0(g, P) + \varepsilon$$

$\mathcal{C} = \mathrm{sign} \circ \mathcal{Y}$

③ Make assumptions on P

- P is $\mathcal{C}$-realizable
- P_x is Uniform or Gaussian

Strategies: ① Change 0/1 loss to a nicer loss fn

② Use improper learning, enlarge the hypothesis class