Peer Eval for PS2 due Tue 10/27 in lecture
PS3 due Thu 10/29
Midterm Thu 11/5 in lecture

Last time:

VC theorem (a.k.a. uniform convergence for $\mathcal{F}$ with finite VC dimension)

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(z_i) - \mathbb{E}[f(Z)] \right| \leq \sqrt{\frac{VC\dim(\mathcal{F}) \log n}{n}}$$

X is v-subgaussian if $K_X(\lambda) \leq \frac{v\lambda^2}{2}$

$\sum_{i=1}^{n} a_i \sigma_i$ where $a_1, \dots, a_n \in \mathbb{R}$, $(\sigma_1, \dots, \sigma_n) \sim \text{Unif}(\{-1, 1\}^n)$

Today:

Sample complexity lower bounds

Upper bounds

VC theorem used to analyze ERM for $\mathcal{C}$ w/ $d = \text{VCdim}(\mathcal{C})$

distribution-free

For any P over $\mathcal{X} \times \{0,1\}$, if training data is iid sample from P of size n, then for any $\delta \in (0,1)$, with probability $\geq 1-\delta$, ERM returns $\hat{c} : \mathcal{X} \to \{0,1\}$ s.t.

$$\text{err}(\hat{c}; P) \leq \underbrace{\inf_{c \in \mathcal{C}} \text{err}(c; P)}_{(*)} + \underbrace{O\left(\sqrt{\frac{d \log(n/d)}{n}} + \sqrt{\frac{\log \frac{1}{\delta}}{n}}\right)}_{(**)}$$

Under $\mathcal{C}$-realizability:

- $(*)$ is 0
- $(**)$ is $\leq \varepsilon$ when $n \geq \frac{d \log \frac{1}{\varepsilon}}{\varepsilon^2} + \frac{\log \frac{1}{\delta}}{\varepsilon^2}$

i.e., ERM (ε, δ)-PAC learns $\mathcal{C}$

0. Can improve dependence on ε to $\frac{d \log \frac{1}{\varepsilon}}{\varepsilon} + \frac{\log \frac{1}{\delta}}{\varepsilon}$
1. Works for any $\mathcal{C}$ (even $|\mathcal{C}| = \infty$) as long as $d < \infty$ ☺
2. "Distribution-free": works for any P (satisfying $\mathcal{C}$-realizability) ☺
3. Sample complexity is polynomial in all relevant quantities ☺

   d, $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$
4. ERM might be computationally intractable ☹

Sample complexity lower bound

Goal: WTS no alg. can get a much better sample complexity and still guarantee distribution-free (ε, δ)-PAC learning of $\mathcal{C}$.

Theorem For any alg. A, there exists a dist. P that satisfies $\mathcal{C}$-realizability such that

if $n < \frac{d-1}{16\varepsilon}$

then with prob. $\geq \frac{1}{4}$ the classifier $\hat{h}: \mathcal{X} \to \{0,1\}$ returned by A has

$$\text{err}(\hat{h}; P) > \varepsilon.$$

1. Shows the VCdim($\mathcal{C}$) is the right property of $\mathcal{C}$ for determining sample complexity. ☺
2. Didn't handle δ 😐 easily fixed.
3. Constants are atrocious ☹
4. Dependence on ε is not matching ☹
   LB: $\frac{1}{\varepsilon}$ UB: $\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}$
   ↑ turns out this can be improved w/ different algo. ☺
5. L.B. is based a contrived P

Proof idea

$d = VC\dim(\mathcal{C})$.

$\rightarrow$ There is a set $T = \{t_1, \ldots, t_d\} \subset \mathcal{X}$ that's shattered by $\mathcal{C}$

$$\mathcal{C}[T] = \{0,1\}^d$$

Suppose $P_X = \text{uniform}(T)$

Suppose $n = d$

$S \sim P^n$

How many of the t_i's do we not see in S? $\frac{d}{e}$

What's the prob we don't see t_1?

$$\left(1 - \frac{1}{d}\right)^d \approx \frac{1}{e}$$

So best you can do is guess randomly on these $\frac{d}{e}$ pts

So your $\hat{h}$ will err on about half of them

$$\text{err}(\hat{h}, P) > \frac{1}{d} \cdot \frac{d}{2e} = \frac{1}{2e}$$

16%

We're happy if the learner just is forced to have error rate ε,

So can afford to put $1 - O(\varepsilon)$ prob. mass on t_d

and uniformly spread out on the rest.

$$P_X(t_1) = \cdots = P_X(t_{d-1}) = \frac{O(\varepsilon)}{d-1}$$

So "useful" data pts only come up every $O\left(\frac{1}{\varepsilon}\right)$ samples

Full proof details

Let $T = \{t_1, \ldots, t_d\}$ shattered by $\mathcal{C}$.
Let $B = (B_1, \ldots, B_d) \sim \text{Uniform}(\{0,1\}^d)$

↳ this defines a random dist. P_B over $\mathcal{X} \times \{0,1\}$
- marginal on $\mathcal{X}$: put $1-4\varepsilon$ mass on t_d, $\frac{4\varepsilon}{d-1}$ mass on t_i, $i<d$.
- label of each t_i is given by B_i

P_B is always $\mathcal{C}$-realizable.

$S \sim P_B^n$ $\quad n < \frac{d-1}{16\varepsilon}$

N = How many of the $\{t_1, \ldots, t_{d-1}\}$ appear in S

$(1-x)^n \geq 1-nx$

$$E[N] = (d-1)\left(1 - \left(1 - \tfrac{4\varepsilon}{d-1}\right)^n\right) \leq (d-1)\cdot\left(1 - \left(1 - \tfrac{4\varepsilon n}{d-1}\right)\right) = 4\varepsilon n$$

$\Pr(N \geq 8\varepsilon n) \leq \frac{1}{2}$.

So w.p. $\geq \frac{1}{2}$, only see less than $8\varepsilon n$ of $\{t_1, \ldots, t_{d-1}\}$ and hence $\leq 8\varepsilon n$ of $\{B_1, \ldots, B_{d-1}\}$

$8\varepsilon n < \frac{d-1}{2}$

So at least $\frac{d-1}{2}$ of the $\{B_1, \ldots, B_{d-1}\}$ are not seen by the learner.

$W = |\{i \in [\frac{d-1}{2}] \mid \hat{h}(t_i) \neq B_i\}|$ even conditional on $\hat{h}$ is $\text{Binomial}(\frac{d-1}{2}, \frac{1}{2})$

$\Pr(W \geq \frac{d-1}{4} \mid \hat{h}) \geq \frac{1}{2}$

So w.p. $\geq \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}$,

$\hat{h}(t_i) \neq B_i$ for $\frac{d-1}{4}$ of the "informative" pts. $\{t_1, \ldots, t_{d-1}\}$

$\text{err}(\hat{h}; P_B) > \frac{4\varepsilon}{d-1} \cdot \frac{d-1}{4} = \varepsilon$

$E[\mathbb{1}\{\text{err}(\hat{h}; P_B) > \varepsilon\}] \geq \frac{1}{4}$

$E[E[\mathbb{1}\{\text{err}(\hat{h}; P_B) > \varepsilon\} \mid B]] \geq \frac{1}{4}$