Sample compression

Concept class $C \subseteq \{0,1\}^X$

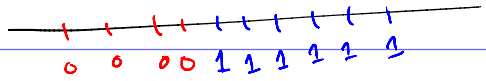
$d := VCdim(C)$

Sally Floyd
Manfred Warmuth

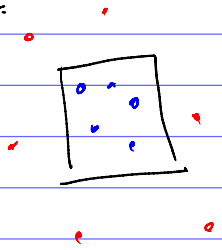
$THR_{\mathbb{R}} = \{c_\theta \mid \theta \in \mathbb{R}\}$

$c_\theta(x) = 1\{x > \theta\}$

$VCdim(THR_{\mathbb{R}}) = 1$



$RECTANGLES_{\mathbb{R}^2}$:



$VCdim(R_{\mathbb{R}^2}) = 4$

Conjecture: There's always a sample compression scheme that chooses $d = VCdim(C)$ pts from a sample

$O(d)$

$poly(d)$ ← saturday $O(d^2)$

$f(d)$ ← 2016 Shay Moran
Amir Yehudayoff

$VCdim(C^*) \leq 2^{(VCdim(C))} = 2^{f(d)}$

PS4 due tonight

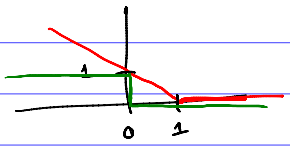
PS4 peer evaluation (last one)
& PS5 (last one)
available next week

Last time:
Linear learning of Lipschitz ridge functions
via polynomial approximation

Today: Limitations of linear learning

Thm Assume $\|\phi(x)\| \le G \ \forall x \in X$ and we use $\eta = W/G\sqrt{T}$

$\ell = \ell_{\text{hinge}}$



then

$$E[R_\ell(\bar{w})] \le \min_{w \in \mathbb{R}^d : \|w\| \le W} R_\ell(w) + \frac{WG}{\sqrt{T}}$$

To get this to ϵ, sample size $T = \left(\frac{WG}{\epsilon}\right)^2$ sufficies.

If $\exists\ w \in \mathbb{R}^d$ s.t.

$y = 1 \Rightarrow \langle \phi(x), w \rangle \ge 1$

$y = -1 \Rightarrow \langle \phi(x), w \rangle \le -1$

then $R_\ell(w) = 0 = \text{err}(\text{sign}(\langle \phi(\cdot), w \rangle))$

Today: restrict attention to C-realizable dist. P for concept class C.

Q: For which concept classes C can we PAC-learn via linear learning? with SGD?

Linear learning: Need $\phi : X \to \mathbb{R}^d$ ← used by learner

$\psi : C \to \mathbb{R}^d$ ← only for the analysis of learner.

s.t.

$$\forall x \in X, c \in C \begin{cases} c(x) = 1 \Rightarrow \text{sign}(\langle \phi(x), \psi(c) \rangle) \ge 1 \\ c(x) = -1 \Rightarrow \text{sign}(\langle \phi(x), \psi(c) \rangle) \le -1 \end{cases}$$

$$c(x) \langle \phi(x), \psi(c) \rangle \ge 1$$

So we can PAC learn by learning $\text{HLTF}_{\mathbb{R}^d}$ but sample complexity bound is $\times d$.

Linear learning via SGD also want $\|\phi(x)\|$ & $\|\psi(c)\|$ to be not too big.

We'll say C has a (W, G)-rep. (ϕ, ψ) in $\mathbb{R}^d$ if $c(x) \langle \phi(x), \psi(c) \rangle \ge 1 \ \forall x \in X, c \in C$.

& $\|\phi(x)\| \le G$

$\|\psi(c)\| \le W$

$D := |\mathcal{X}|$, $N := |\mathcal{C}|$

domain num of concepts

Trivial Representation

$\mathcal{X} = \{x^{(1)}, \ldots, x^{(D)}\}$

$\phi : \mathcal{X} \to \mathbb{R}^D$, $\phi(x^{(i)}) = e_i$ $\quad \|\phi(x)\| = 1$

$\psi : \mathcal{C} \to \mathbb{R}^D$, $\psi(c) = \sum_{i=1}^{D} c(x^{(i)})\, e_i$ $\quad \|\psi(c)\| = \sqrt{D}$

$\langle \phi(x^{(i)}), \psi(c) \rangle = c(x^{(i)})$

Monotone Conjunctions $\mathcal{C} = MCONJ_n$ $\quad (\mathcal{X} = \{0,1\}^n$, $D = 2^n$, $N = 2^n)$

$\mathcal{C} = \{ c_S \mid S \subseteq [n] \}$ $\quad c_S(x) = \bigwedge_{i \in S} x_i$

Claim: $\mathcal{C}$ has a $(n, \sqrt{n+1})$-rep in $\mathbb{R}^{n+1}$.

Pf $\quad \phi(x) = (2x_1 - 1, 2x_2 - 1, \ldots, 2x_n - 1, 1) \in \mathbb{R}^{n+1}$

$$\psi(c_S) = \sum_{i \in S} e_i - (|S| - 1)\, e_{n+1}$$

If $c_S(x) = 1$, then

$$\langle \phi(x), \psi(c_S) \rangle = \underbrace{\sum_{i \in S} (2x_i - 1)}_{= |S|} - (|S| - 1)$$

$$= 1$$

If $c_S(x) = -1$

$$\langle \phi(x), \psi(c_S) \rangle = \underbrace{\sum_{i \in S} (2x_i - 1)}_{\leq |S| - 2} - (|S| - 1)$$

$$\leq |S| - 2 - |S| + 1$$

$$= -1 .$$

$$\|\phi(x)\| = \sqrt{n+1}$$

$$\|\psi(c)\| = \sqrt{|S| + (|S| - 1)^2} \leq n$$

Thm If C has a (W, G)-rep (in any dim), then
$$Ldim(C) \leq (WG)^2$$

Lower bound / limitation of linear learning!

Pf: $\forall x \; \forall c$

$c(x) \langle \phi(x), \psi(c) \rangle \geq 1$

norm of weight vector is $\leq W$

γ

norm of feature vector $\leq G$

mistake bound of Perceptron $\frac{G^2 W^2}{\gamma^2} = (GW)^2$. ⊠

$Ldim(LTF_{\mathbb{R}^n}) = \infty$ ∴

Our theorems about SGD do not imply that SGD can be used to learn $LTF_{\mathbb{R}^n}$.

$THR_D = \{ c_j \mid j \in \{0, 1, 2, \ldots, D\} \}$ $\quad X = [D]$

$c_j(x) = \mathbb{1}\{x > j\}$

$|THR_D| = D + 1$

$Ldim(THR_D) = \log D$ ← SGD requires some dependence on D

but $VCdim(THR_D) = 1$ ← but no dep. on D is necessary in general.

efficiently PAC-learnable

Q: Is there any C st. trivial rep is best possible?

$C = \text{PARITY}_n = \{c_S \mid S \subseteq [n]\} \subseteq \{-1,1\}^{\{0,1\}^n}$

$$c_S(x) = \begin{cases} 1 & \text{if } \sum_{i \in S} x_i \text{ is even} \\ -1 & \text{if } \sum_{i \in S} x_i \text{ is odd} \end{cases}$$

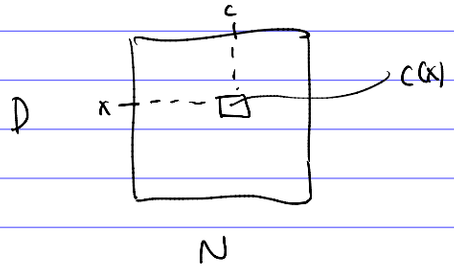
Parity fns are just linear functions over $\mathbb{F}_2$

VC dim $(\text{PARITY}_n) = \log_2 |\text{PARITY}_n| = n$

efficiently PAC learnable (via Gaussian elimination)

Thm (Forster 2002): every rep. for PARITY_n has dim $d \geq \sqrt{2^n}$

Easier Thm every (W,G)-rep for PARITY_n has $WG \geq \sqrt{2^n}$.



For PARITY_n, $D = N = 2^n$

$H_n(x,c) = C(x) \in \{-1,1\}$

$2^n \times 2^n$ Hadamard matrix

$n=1$

$H_1 =$

	ϕ	$\{1\}$
0	1	1
1	1	-1

$c_\phi(0) = c_\phi(1) = 1$

$c_{\{1\}}(0) = 1$

$c_{\{1\}}(1) = -1$

$n \geq 2$

$H_n =$

	$S \not\ni n$	$S \ni n$
strings ending in 0	H_{n-1}	H_{n-1}
strings ending in 1	H_{n-1}	$-H_{n-1}$

H_n is symmetric

Columns of H_n are orthogonal.

- Take any $S \neq S'$

$$\sum_{x \in \{0,1\}^n} c_S(x)\, c_{S'}(x) = 0$$

Rows of H_n are orthogonal

For any $x \neq x'$

$$\sum_{S \subseteq [n]} c_S(x)\, c_S(x') = 0.$$

Thm: Every (W, G)-rep for PARITY_n has $WG \geq \sqrt{2^n}$

Pf (ϕ, ψ)

$$1 \leq c_S(x) \langle \phi(x), \psi(c_S) \rangle$$

Sum over all S & x

$$(2^n)^2 \leq \sum_S \left\langle \sum_x c_S(x)\phi(x), \psi(c_S) \right\rangle$$

$$\leq \sum_S \left\| \sum_x c_S(x)\phi(x) \right\| \cdot \underbrace{\|\psi(c_S)\|}_{\leq W} \qquad \text{C-S}$$

$$\leq W \sqrt{\sum_S \left\| \sum_x c_S(x)\phi(x) \right\|^2} \sqrt{2^n} \qquad \text{C-S}$$

$$\leq W \sqrt{(2^n)^2 G^2} \sqrt{2^n}$$

$$= WG\,(2^n)^{3/2}$$

$\therefore WG \geq \sqrt{2^n}$. ☒

$$\sum_S \left(\sum_x c_S(x)^2 \|\phi(x)\|^2 + \sum_{x \neq x'} c_S(x) c_S(x') \langle \phi(x), \phi(x') \rangle \right)$$

$$= \sum_S \sum_x \|\phi(x)\|^2 + \sum_{x \neq x'} \sum_S c_S(x) c_S(x') \langle \phi(x), \phi(x') \rangle$$

$$\leq (2^n)^2 G^2$$