Lecture 13

RECAP

Definition (SQ dimension) The SQ-dimension of a class $\mathcal{C}$ wrt a distribution D over X is the size of the largest subset $\mathcal{C}' \subseteq \mathcal{C}$ s.t. for all $f, g \in \mathcal{C}'$,

$$\left| \Pr_{x \sim D} [f(x) = g(x)] - \frac{1}{2} \right| \leq \frac{1}{|\mathcal{C}'|}.$$

Thm2: If $SQ\text{-}DIM_D(\mathcal{C}) > poly(d)$ then you **cannot** efficiently learn $\mathcal{C}$ over D by SQ-algorithms (even "weak-learning" to error $\leq \frac{1}{2} - \frac{1}{poly(d)}$ is impossible).

Parity function:

$X^d = \{0,1\}^d$

$Y = \{0,1\}$

$\mathcal{C} = \{ w(x) = \langle w, x \rangle \bmod 2 : w \in \{0,1\}^d \}$

Thm: PARITIES are efficiently PAC learnable.

Let U be the uniform distribution over $\{0,1\}^d$.

Claim: Any two parity functions $\chi_{w_1}(x)$ & $\chi_{w_2}(x)$ (where $w_1 \neq w_2$) are uncorrelated:

$$\Pr_{x \sim U}\left[\chi_{w_1}(x) = \chi_{w_2}(x)\right] = \frac{1}{2}.$$

TODAY

It will be convenient to define PARITIES as a function on $\{-1,+1\}^d \to \{-1,1\}$

$$\chi^d = \{\pm 1\}^d$$

$$Y = \{\pm 1\}$$

$$\mathcal{C} = \left\{ \chi_S(x) = \prod_{i \in S} x_i \; : \; S \subseteq \{1,\dots,d\} \right\}$$

New Claim: If $S \neq T$

$$\Pr_{x \sim U}\left[\chi_S(x) = \chi_T(x)\right] = \frac{1}{2}.$$

Proof

$$\mathbb{E}_{x \sim U}\left[\chi_S(x) \cdot \chi_T(x)\right] = \mathbb{E}_{x \sim U}\left[\prod_{i \in S} x_i \cdot \prod_{i \in T} x_i\right]$$

$$= \mathbb{E}_{x \sim U}\left[\prod_{i \in S \Delta T} x_i\right]$$

$$(S \Delta T = \{S - T\} \cup \{T - S\})$$

$$= 0 \quad \text{if } S \neq T \text{ (unif dist.)}$$

$$\Pr[c_S(x) = c_T(x)] + \Pr[c_S(x) \neq c_T(x)] = 1$$

$$\mathbb{E}_{x \sim U}[c_S(x) \cdot c_T(x)] = \Pr[c_S(x) = c_T(x)] - \Pr[c_S(x) \neq c_T(x)]$$

$$\therefore \mathbb{E}[c_S(x)\, c_T(x)] = 0$$

$$\Rightarrow \Pr[c_S(x) = c_T(x)] = \frac{1}{2}.$$

Corollary of Thm 2 It is not possible to efficiently learn PARITIES in the SQ model over the uniform distribution

Proof $SQ\text{-}DIM_U(\mathcal{C}) = 2^d$.

We'll now prove Thm 2 for the special case of PARITIES.

Thm (hardness of parities in SQ): Any SQ algorithm for learning PARITIES over $D = U$, which makes queries of tolerance $\tau \geq \tau_{min}$ must make $\Omega(\tau_{min}^2\, 2^d)$ queries to $STAT(c, D)$.

Proof

correlation SQ (CSQ) oracle:

for any query function $\psi : \mathcal{X} \to \{-1, +1\}$, & tolerance τ

let $P_\psi = \mathbb{E}[\psi(x) \cdot c(x)]$

oracle returns $\hat{P}_\psi \in [P_\psi - \tau, P_\psi + \tau]$

Lemma: If learner knows target distribution D, can simulate SQ oracle with CSQ oracle.

Proof we can decompose any SQ ϕ into:

$$\mathop{\mathbb{E}}_{x \sim D}[\phi(x, c(x))] = \mathop{\mathbb{E}}_{x \sim D}[\phi(x, 1) \cdot \mathbb{1}(c(x) = 1)]$$

$$+ \mathop{\mathbb{E}}_{x \sim D}[\phi(x, -1) \cdot \mathbb{1}(c(x) = -1)]$$

$$= \mathop{\mathbb{E}}_{x \sim D}\left[\phi(x, 1)\left(\frac{1 + c(x)}{2}\right)\right]$$

$$+ \mathop{\mathbb{E}}_{x \sim D}\left[\phi(x, -1)\left(\frac{1 - c(x)}{2}\right)\right]$$

"target-independent query" ← $= \frac{1}{2}\left(\mathop{\mathbb{E}}_{x \sim D}[\phi(x, 1)] + \mathop{\mathbb{E}}_{x \sim D}[\phi(x, -1)]\right)$

"correlational query" ← $+ \frac{1}{2}\left(\mathbb{E}\left(\phi(x, 1) c(x)\right) - \mathbb{E}\left(\phi(x, -1) \cdot c(x)\right)\right)$

∎

Since we consider $D = U$ (fixed), suffices to show hardness for CSQ oracle.

Basics of Boolean Function Analysis

Think of any function $f : \{\pm 1\}^d \to \{\pm 1\}$ as a vector $\vec{f}$ of 2^d entries.

$$\left(\frac{1}{2^{d/2}} f(-1,-1,-1), \frac{1}{2^{d/2}} f(-1,-1,1), \ldots, \frac{1}{2^{d/2}} f(1,1,1) \right)$$

Note that $\langle \vec{f}, \vec{g} \rangle = \mathop{\mathbb{E}}_{x \sim U} [f(x) \cdot g(x)]$

$$\langle \vec{f}, \vec{g} \rangle = \sum_{i=1}^{2^d} \frac{1}{2^{d/2}} f(x_i) \cdot \frac{1}{2^{d/2}} g(x_i)$$

$$= \mathop{\mathbb{E}}_{x \sim U} [f(x) \cdot g(x)].$$

In particular $\langle \vec{f}, \vec{f} \rangle = 1$

Fourier analysis: change "basis" to understand f.

Recall that an orthonormal basis for a vector space is a set of orthogonal unit vectors that span the space.

If v_1, v_2 are orthonormal basis for $\mathbb{R}^2$, we can write any vector

$$w = \langle w, v_1 \rangle v_1 + \langle w, v_2 \rangle v_2 .$$

Claim PARITIES form an orthonormal basis for our vector space.

Proof Note that for $S \neq T$

$$\langle \vec{\chi}_S(x), \vec{\chi}_T(x) \rangle = \mathbb{E}\left(\chi_S(x) \cdot \chi_T(x)\right) = 0.$$

$$\langle \vec{\chi}_S(x), \vec{\chi}_S(x) \rangle = 1. \quad \forall S$$

■

For any CSQ $\psi : \{-1,1\}^d \to \{-1,1\}$

$$\vec{\psi} = \sum_{S : S \subseteq [n]} \alpha_S \vec{\chi}_S$$

where $\vec{\chi}_S$ is the parity function over S.

Note that $\alpha_S = \mathbb{E}_{x \sim U}\left[\psi(x) \cdot \chi_S(x)\right]$

Expected response to CSQ ψ if the target function is $\chi_S(x)$.

Since $\langle \vec{\psi}, \vec{\psi} \rangle = 1$

$$\sum \hat{\psi}_s^2 = 1$$

$\therefore$ There can be atmost $\frac{1}{\tau^2}$ S s.t. $|\hat{\psi}_s| \geq \tau$.

Note that if target is S^*, then CSQ oracle can just answer 0 to this query ψ if $|\hat{\psi}_{S^*}| < \tau$.

We draw the target parity function (the set S^*) uniformly at random from all possible subsets.

Claim If alg. makes $\leq \tau^2 2^d$ queries, whp over choice of S^*, CSQ can answer 0 to all these queries.

Proof. There are 2^d options

Any query ψ is non-zero on atmost $\frac{1}{\tau^2}$ options.

If s^* is not among these, just answer 0.

Since s^* is random, whp cannot "find" it in $O(\tau^2 2^d)$ queries. [Exercise]

This finishes proof. ∎

Thm Over the unif dist. D, in the presence of $R<N$ with noise level η, PARITIES are learnable $O\left(\frac{d}{(1-2\eta)^2}\right)$ samples, whp.
(information-theoretically)

Proof Let $\varepsilon = \frac{1}{2} - \eta$

We keep label w.p. $\frac{1}{2} + \varepsilon$
" flip " " $\frac{1}{2} - \varepsilon$

Get $m = O\left(\frac{d}{\varepsilon^2}\right)$ samples from $Ex^{\eta}(c, D)$.

(where $c(x) = c_{s^*}(x) : \prod_{i \in s^*} x_i$)

Claim: with high prob,

① S^* is consistent with $\geq \left(\frac{1}{2} + \frac{\varepsilon}{2}\right)$ fraction of examples

② Any $S \neq S^*$ is consistent with $< \left(\frac{1}{2} + \frac{\varepsilon}{2}\right)$ fraction.

Proof Exercise. Use Chernoff / Hoeffding & property that parity functions are uncorrelated for $S \neq S'$.

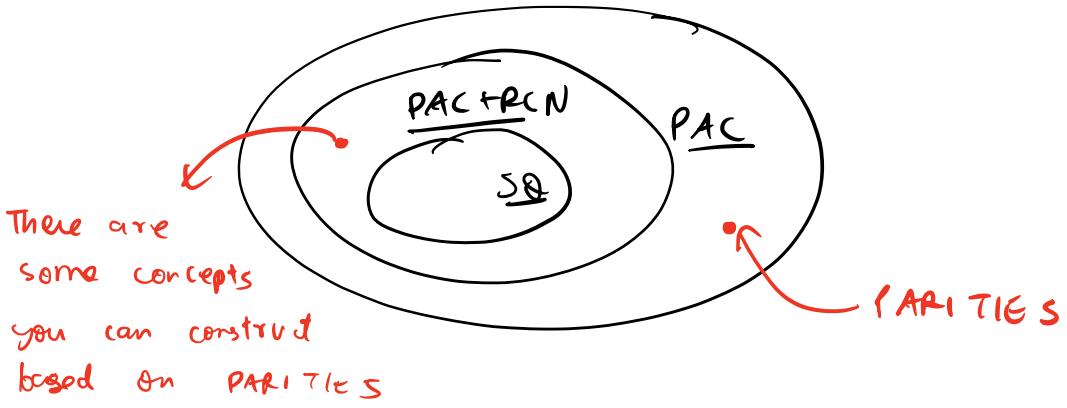
Best known algorithm for learning parity with noise: $2^{d/\log d}$ time

(Blum-Wasserman-Kalai '03)

Slightly less than exponential, non-SQ algorithm.

↳ PNC (Learning Parity w/ Noise) is believed to be hard.

Efficient also.



$2^{d/\log d}$ alg. => we can learn parities over $O(\log d \log\log d)$ co-ordinates with RCN in poly-time

$$SQ \subset PAC + RCN \subseteq PAC$$

↑ conditioned on hardness of LPN, this is proper.

With exception of Gaussian elimination, almost all known algo. can be seen in SQ model.

SQ is sort of frontier of algorithmic knowledge.

∴ To show that some learning problem is hard, many recent papers show hardness in SQ models.

Thanks to SQ-dim, we have an information theoretic way to show hardness in SQ.

Boosting

Recall our definition of weak-learning.

Weak-learning: An algorithm A is a weak learner with **edge / advantage** γ for class $\mathcal{C}$ if: for any dist. D & any target $c \in \mathcal{C}$, given access to $EX(c, D)$, w.p. $(1-\delta)$ A produces a hypothesis with $\text{error}(h, c, D) \leq \frac{1}{2} - \gamma$.

If A runs in time $\text{poly}(d, \frac{1}{\delta})$ & $\frac{1}{\gamma} > \frac{1}{\text{poly}(d)}$

then $\mathcal{C}$ is efficiently weakly-PAC learnable

Thm If $\mathcal{C}$ is weakly-PAC learnable (efficiently), then $\mathcal{C}$ is PAC-learnable (efficiently.

Proof AdaBoost (Freund & Schapire)

(Godel prize '03, very influential in practice)