Lecture 13

- Basics of Fourier Analysis on Boolean cube (some review, some new)
- Analysis of linearity test
- learning Boolean functions - a model

Recall from last time:

def. $f: \{\pm 1\}^n \to \{\pm 1\}$ is "linear" (homomorphism) if

$$\forall\, x, y \in \{\pm 1\}^n \quad f(x) \cdot f(y) = f(x \odot y)$$

coordinatwise multiplication $x \odot y = (x_1 y_1, x_2 y_2, \ldots)$

def f is "ε-linear" if $\exists$ linear g s.t. f & g agree on $\geq 1-\varepsilon$ fraction of inputs.

A useful observation:

$$\forall\, a, y \in \{\pm 1\}^n \quad \Pr_x[\, y = a \odot x \,] = \frac{1}{2^n}$$

Linear fctns are:

$$S \subset \{1..n\}$$

$$\chi_S(x) = \prod_{i \in S} x_i$$

"Parity fctns"

1 if $\#\, i \in S$ s.t. $x_i = -1$ even

-1 if " " " " odd

e.g. $f(x) = x_1 \cdot x_3 \cdot x_5 \cdot x_7$

$g(x) = x_1 \cdot x_2$

def. "linearity tester"

given: query access to fct f & parameter ε

requirements:

- if f linear, $\Pr[\text{tester passes}] = 1$
- if f <u>not</u> ε-linear, $\Pr[\text{tester fails}] \geq 3/4$

arbitrary constant

Proposed test:

Pick random $x, y \in \{\pm 1\}^n$

Test $f(x) \cdot f(y) = f(x \circ y)$

repeat how many times?

Express event that test passes as algebraic fctn:

$$f(x)\cdot f(y)\cdot f(x\odot y) = \begin{cases} 1 & \text{if test accepts} \\ -1 & \text{" " rejects} \end{cases}$$

$f(x)\cdot f(y) = f(x\odot y)$ ⇕

$f(x)\cdot f(y) \neq f(x\odot y)$

⇕

0/1 indicator var

$$\frac{1 - f(x)\,f(y)\,f(x\odot y)}{2} = \begin{cases} 0 & \text{if accepts} \\ 1 & \text{o.w.} \end{cases}$$

Now we have a new way to express rejection probability:

rejection probability

$$\begin{aligned} \delta_f &\equiv \Pr_{x,y}\left[f(x)\odot f(y) \neq f(x\odot y)\right] \\ &= E_{x,y}\left[\frac{1 - f(x)f(y)f(x\odot y)}{2}\right] \end{aligned}$$

Fourier Analysis on Boolean Cube

want basis to describe all fctns. $f: \{\pm 1\}^n \to \{\pm 1\}$

Linear fctns:

$$S \subseteq \{1..n\}$$

parity fctns

for $x \in \{\pm 1\}^n$, $\boxed{\chi_S(x) = \prod_{i \in S} x_i}$

define $\boxed{\langle f, g \rangle = \frac{1}{2^n} \sum_{x \in \{\pm 1\}^n} f(x)\, g(x)}$ inner product (but normalized)

Fact parity (linear) fctns $\{\chi_S\}$ is orthonormal basis w.r.t. inner product!

Proof of fact:

- $\langle \chi_S, \chi_S \rangle = \frac{1}{2^n} \sum_x (\chi_S(x))^2 = \frac{2^n}{2^n} = 1$ normal

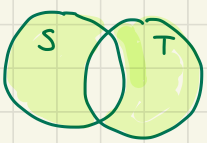
  ($\chi_S(x) = \pm 1$, so $(\chi_S(x))^2 = +1$)

- if $S \neq T$:

$$\langle \chi_S, \chi_T \rangle = \frac{1}{2^n} \sum_x \chi_S(x) \cdot \chi_T(x)$$

if $i \in S \cap T$ then $x_i \cdot x_i = 1$ "drops out" so can ignore

$S \Delta T$:



$$= \frac{1}{2^n} \sum_x \chi_{S \Delta T}(x)$$

nonempty since $S \neq T$ so pick $j \in S \Delta T$

$$= \frac{1}{2^n} \sum_{\substack{\text{pairs} \\ x, x^{\oplus j}}} \left(\chi_{S \Delta T}(x) + \chi_{S \Delta T}(x^{\oplus j}) \right)$$

$x^{\oplus j} = x$ with jth bit flipped

$$= \frac{1}{2^n} \sum_{\substack{\text{pairs} \\ x, x^{\oplus j}}} x_j \cdot \prod_{i \in (S \Delta T) \setminus \{j\}} x_i + \overline{x}_j \cdot \prod_{i \in (S \Delta T) \setminus \{j\}} x_i$$

equal

$$= \frac{1}{2^n} \sum_{\text{pairs}} 0$$

One is $+1$ the other is -1 so sum to 0

$$= 0$$

Orthogonal!

So $\{\chi_S\}$ is an orthonormal basis

Thm f is uniquely expressible as linear comb. of χ_S.

Def. $\hat{f}(S) \equiv \langle f, \chi_S \rangle$

$$= \frac{1}{2^n} \sum_{x \in \{\pm 1\}^n} f(x)\chi_S(x)$$

Fourier Coefficients of f

Thm $\forall f \quad f(x) = \sum \hat{f}(S)\chi_S(x)$

Fourier coefts of linear fctns:

Fact f linear $\Leftrightarrow$ $\exists\ S \subseteq [n]$ s.t. $\hat{f}(S) = 1$ ← one is really big

$\&\ \forall\ T \neq S \quad \hat{f}(T) = 0$ ← others are 0

e.g. if $f(x) = x_1 \cdot x_2$

$f(x) = 0 \cdot \chi_\emptyset + 0 \cdot \chi_{\{1\}} + 0 \cdot \chi_{\{2\}} + 1 \cdot \chi_{\{1,2\}}$

Fourier coeffs characterize distance to linear:

Lemma $\forall S \subseteq [n]$

$$\hat{f}(S) = 1 - 2\,\mathrm{dist}(f, \chi_S)$$

$$= 1 - 2 \Pr_{x \in \{\pm 1\}^n} [f(x) \neq \chi_S(x)]$$

Pf $2^n \cdot \hat{f}(S) = \sum_x f(x)\chi_S(x)$ def of Fourier coeff

$$= \sum_{\substack{x \text{ s.t.} \\ f(x) = \chi_S(x)}} 1 \; + \sum_{\substack{x \text{ s.t.} \\ f(x) \neq \chi_S(x)}} -1$$

$$= 2^n [1 - \mathrm{dist}(f, \chi_S)] - 2^n [\mathrm{dist}(f, \chi_S)]$$

$$= 2^n (1 - 2 \cdot \mathrm{dist}(f, \chi_S))$$ ▨

example: f = all -1's

for $S = \emptyset$: $\mathrm{dist}(f, \chi_\emptyset) = 1$ so $\hat{f}(\emptyset) = -1$

$\forall S \neq \emptyset$ $\mathrm{dist}(f, \chi_S) = \frac{1}{2}$ (why?) so $\hat{f}(S) = 0$

Observation: Any two distinct linear fctns differ on exactly $\frac{1}{2}$ of inputs

pf. let $f = \chi_T$ for $T \neq S$
$g = \chi_S$

note $\hat{f}(S) = \hat{\chi}_T(S) = \langle \chi_T, \chi_S \rangle$

$$0 = \langle \chi_T, \chi_S \rangle = 1 - 2\,\text{dist}(\chi_T, \chi_S)$$

↑ since orthonormal ↑ lemma

$$\Rightarrow \text{dist}(\chi_T, \chi_S) = \frac{1}{2}$$ ∎

Cordlary: if $S \neq \emptyset$, $\chi_S(x) = +1$ on exactly $\frac{1}{2}$ the inputs

Very useful tools:

Plancherel's identity

$$\langle f, g \rangle = \left\langle \sum_{S \subseteq [n]} \hat{f}(S)\chi_S, \sum_{T \subseteq [n]} \hat{g}(T)\chi_T \right\rangle$$

$$= \sum_{S,T} \hat{f}(S)\,\hat{g}(T) \langle \chi_S, \chi_T \rangle$$ bilinearity of $\langle \cdot, \cdot \rangle$

$\langle \chi_S, \chi_T \rangle$: 0 if $S \neq T$; 1 if $S = T$

$$= \sum_{S} \hat{f}(S)\,\hat{g}(S)$$

Parseval's identity:

$$\langle f, f \rangle = \sum_S \hat{f}(S)^2$$

"Boolean Parseval's"

if $f: \{0,1\}^n \to \{\pm 1\}$ $\quad \langle f, f \rangle = \frac{1}{2^n} \sum_x f(x) \cdot f(x) = \frac{1}{2^n} \cdot 2^n = 1$

$f(x) \cdot f(x) = +1$ since f is Boolean

So $\sum_S \hat{f}(S)^2 = 1$

Analysis of linearity test

if δ_f is small, can we conclude that f is close to linear?

doesn't this contradict Coppersmith's lower bound? this is only for $f: \{\pm 1\}^n \to \{\pm 1\}$

YES! rejection probability gives upper bound on distance.

Thm. $f: \{\pm 1\}^n \to \{\pm 1\}$ is δ_f-close to some linear fctn

Pf.

$$E_{x,y}[f(x)f(y)f(x\circ y)]$$

$$= E_{x,y}\left[\left(\sum_S \hat{f}(S)\chi_S(x)\right)\cdot\left(\sum_T \hat{f}(T)\chi_T(y)\right)\cdot\left(\sum_U \hat{f}(U)\chi_U(x\circ y)\right)\right]$$

$$= E_{x,y}\left[\sum_{S,T,U} \hat{f}(S)\hat{f}(T)\hat{f}(U)\,\chi_S(x)\chi_T(y)\chi_U(x\circ y)\right]$$

what is this?

if $S=T=U$: $\chi_S(x)\chi_T(y)\chi_U(x\circ y)$ ($x_i \cdot y_i$)

$$= \prod_{i\in S} x_i \cdot \prod_{i\in S} y_i \cdot \prod_{i\in S} (x\circ y)_i$$

$$= \prod_i x_i \cdot y_i \cdot (x_i \cdot y_i) = \prod_i x_i^2 y_i^2 = \prod_i (1\cdot 1) = 1$$

if $\neg(S=T=U)$:

$$E_{x,y}[\chi_S(x)\chi_T(y)\chi_U(x\circ y)$$

$$= E_{xy}\left[\prod_{i\in S} x_i \prod_{j\in T} y_j \prod_{k\in U}(x_k\cdot y_k)\right]$$

$$= E_{xy}\left[\prod_{i\in S\Delta U} x_i \cdot \prod_{j\in T\Delta U} y_j\right]$$

$$= E_{xy}\left[\prod_{i\in S\Delta U} x_i\right]\cdot E_{xy}\left[\prod_{j\in T\Delta U} y_j\right]$$ since x,y indep

assumption $\Rightarrow$ one of these holds so $\rightarrow$

if $S\neq U$ then $S\Delta U\neq\emptyset \Rightarrow E_{xy}\left[\prod_{i\in S\Delta U} x_i\right]=0$

analogously if $T\neq U$, $E_{xy}\left[\prod_{j\in T\Delta U} y_j\right]=0$

$= 0$ so all these terms drop out!

$$E_{x,y}[f(x)f(y)f(x\circ y)]$$

$$= E_{xy}\left[\left(\sum_S \hat{f}(S)\chi_S(x)\right)\cdot\left(\sum_T \hat{f}(T)\chi_T(y)\right)\cdot\left(\sum_U \hat{f}(U)\chi_U(x\circ y)\right)\right]$$

$$= E_{xy}\left[\sum_{S,T,U} \hat{f}(S)\hat{f}(T)\hat{f}(U)\chi_S(x)\chi_T(y)\chi_U(x\circ y)\right]$$

$$= \sum_{S=T=U} \hat{f}(S)^3$$

$$\leq \max_S \hat{f}(S) \cdot \sum_S \hat{f}(S)^2$$

$\sum_S \hat{f}(S)^2 = 1$ by "Boolean Parseval's"

$$= \max_S \hat{f}(S)$$

$$= \max_S \left(1 - 2 \cdot \text{dist}(f, \chi_S)\right)$$

$$= 1 - 2 \cdot \min_S \left(\text{dist}(f, \chi_S)\right)$$

So $$\delta_f \geq \frac{1 - \left(1 - 2 \min_S \left(\text{dist}(f, \chi_S)\right)\right)}{2}$$

$$= \min_S \text{dist}(f, \chi_S)$$

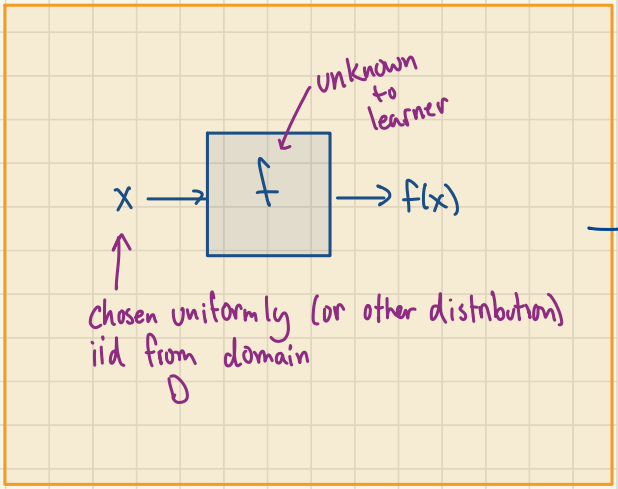
$\Rightarrow \exists S$ s.t. f is δ_f-close to χ_S

■

Learning

Learn from random, uniform examples ← How to formalize?

lots of other models – will mention some others later



Example oracle $Ex(f)$

After seeing several examples, learner should output hypothesis h.

what do we hope h satisfies?

- hopefully $h = f$ ← asking too much?
- at least $\text{dist}(h,f) \leq \varepsilon$

$$\text{dist}(h,f) = \Pr_{x \in D}[h(x) \neq f(x)]$$

in what distribution on inputs do we use?

today: uniform

in general, match distribution of example oracle

Valiant's PAC model "probably approximately correct"

Common terms for same thing:

- $\text{dist}_{x \in D}(h,f)$
- $\text{error}_{x \in D}(h)$ (w.r.t. f is understood from context)
- h is ε-close to f (dist D understood from context)

Note in above:

$x \in D$ can be chosen according to uniform or any other prespecified distribution

Note if f is arbitrary, there is nothing you can do that is "efficient" in terms of sample complexity

(e.g. you can't learn a random fctn f without seeing the value of f for most inputs)

However, if you know something about f, there may be hope.

here: what if you know that f is a member of fctn family C?

e.g. C = linear fctns
k-term DNF
⋮

def uniform distribution learning algorithm for concept class $\mathcal{C}$ is algorithm $\mathcal{A}$ s.t.

- $\mathcal{A}$ given $\varepsilon, \delta > 0$
  access to $Ex(f)$ for $f \in \mathcal{C}$
- $\mathcal{A}$ outputs h s.t. with prob $\geq 1-\delta$
  error (h) w.r.t. f is $\leq \varepsilon$
  (h is ε-close to f)

Parameters of interest:

- m — # samples used by $\mathcal{A}$ "sample complexity"
- ε — accuracy parameter
- δ — confidence parameter
- runtime? — hope for poly $(\log(\text{domainsize}), \frac{1}{\varepsilon}, \frac{1}{\delta})$

- description of h?
  - should it be similar to description of fctns in $\mathcal{C}$? "proper learning"
  - at least should be relatively compact + efficient to evaluate
    compact: $O(\log|\mathcal{C}|)$

Remarks

- as before, dependence on δ needn't be more than $O(\log(1/\delta))$ why?

- uniform case is special case of PAC-model:
  given $EX_{\mathcal{D}}(f)$ for unknown $\mathcal{D}$
  output h with small error with respect to same $\mathcal{D}$
  (some $\mathcal{D}$ can be harder than others)

Efficient learning algorithm for conjunctions:

C = conjunctions over $\{0,1\}^n$

ie. $f(x) = x_i x_j \bar{x}_k$

note:

- can't hope for 0-error from subexponential # of random examples

  e.g. how to distinguish $f(x) = x_1 x_2 \cdots x_n$ from $f'(x) = 0 \ \forall x$?

Poly time algorithm:

- draw $poly(\frac{1}{\varepsilon})$ random examples
- estimate $Pr[f(x)=1]$ to additive error $\pm \frac{\varepsilon}{4}$
- if estimate $< \varepsilon/2$, output "$h(x)=0$" & halt

  else estimate $\geq \varepsilon/2$ so $Pr[f(x)=1] > \frac{\varepsilon}{4}$

  $\Rightarrow$ see new random positive example every (at most) $O(1/\varepsilon)$ examples
- collect $\underline{?}$ more positive examples
- let $V = \{$vars set same way in each positive example$\}$
- Output $h(x) = \bigwedge_{i \in V} x_i^{b_i}$ ← b_i tells us if x_i complemented or not

Behavior of poly time algorithm:

for i in conjunction:

must be set same way in each positive example $\Rightarrow$ in V

for i <u>not</u> in conjunction:

$$\Pr[i \in V] \leq \Pr[i \text{ set same way in each of } k \text{ positive examples}]$$

$$\leq \frac{1}{2^{k-1}}$$

$$\Pr[\text{any } i \text{ } \underline{\text{not}} \text{ in conjunction manages to survive}]$$

$$\leq \frac{n}{2^{k-1}}$$

$$\leq \delta \text{ if pick } k = \log \frac{n}{\delta}$$

So if use $\Omega(\log \frac{n}{\delta})$ positive examples

or $\Omega(\frac{1}{\varepsilon} \log \frac{n}{\delta})$ total examples, will suffice to rule out all $i \notin$ conjunction.