Today:

Learning via Fourier representation

- some fctns & their representation
- the low degree algorithm
- applications

Recall: $x_i \in \{\pm 1\}$

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

$$\langle f, g \rangle = \frac{1}{2^n} \sum_x f(x)\, g(x)$$

<u>def</u> Fourier coeffs: $\hat{f}(S) \equiv \langle f, \chi_S \rangle$

<u>Thm</u> $\hat{f}(S) = 1 - 2\Pr_x [f(x) \neq \chi_S(x)]$

<u>Thm</u> "Boolean" Parseval's $\sum_S \hat{f}(S)^2 = 1$ if f maps to ± 1

Learning via Fourier Representation

based on learning/approx Fourier representation of function

Approximation of one Fourier coeff:

lemma can approx any specific Fourier coeff S to w/in additive error γ

i.e. $|\text{output} - \hat{f}(S)| \leq \gamma$

with prob $\geq 1-\delta$ in $O\left(\frac{1}{\gamma^2}\log\frac{1}{\delta}\right)$ samples (from unif dist on x)

note: no queries needed, just random samples

Pf. $\hat{f}(S) = 2 \underbrace{\Pr_x[f(x) = \chi_S(x)]}_{\text{estimate from samples}} - 1$

+ Chernoff/Hoeffding ■

Can we find <u>all</u> Fourier coeffs?

yes, but:

1) time complexity is $\approx 2^n \times$ time for 1 coeff

2) sample complexity:

can reuse samples

<u>but</u> if you want all estimations to be good need union bnd

so need $\delta = \frac{1}{2^n}$

$\Rightarrow O\left(\frac{1}{\gamma^2} \log 2^n\right) =$

$O\left(\frac{n}{\gamma^2}\right)$ samples

$\Rightarrow$ exp time

poly samples possible

What about f with "few" heavy Fourier coeffs? "sparse"

Interesting? how do we find <u>which</u> are heavy?

What if f is sparse + heavy coeffs are in low degree (small S)

Fourier representations of important examples:

1) $\overline{\text{AND}}$ on $T \subseteq N$ $\quad |T| = k$

$T = \{i_1 \cdots i_k\}$

recall:

old $\left.\begin{matrix} 1 \\ 0 \end{matrix}\right\}$ And $\Rightarrow$ all $x_i = +1$

$0 \to +1$
$1 \to -1$

new $\left.\begin{matrix} -1 \\ +1 \end{matrix}\right\}$ And $\Rightarrow$ all $x_i = -1$

True $= -1$
False $= +1$

$$\overline{\text{AND}}(x_{i_1} \dots x_{i_k}) = \begin{cases} 1 & \text{if } \forall\, i_j \in T \in \{i_1 \cdots i_k\} \\ & \quad x_{i_j} = -1 \\ -1 & \text{o.w.} \end{cases}$$

First define

$$f(x) = \begin{cases} 1 & \text{if } \forall\, i \in T \quad x_i = -1 \\ 0 & \text{o.w.} \end{cases}$$ corresponds to AND over $\{0,1\}$

$$= \frac{(1 - x_{i_1})}{2} \cdot \frac{(1 - x_{i_2})}{2} \cdot \ldots \frac{(1 - x_{i_k})}{2}$$

$$= \sum_{S \subseteq T} \frac{(-1)^{|S|}}{2^k} \cdot \chi_S$$

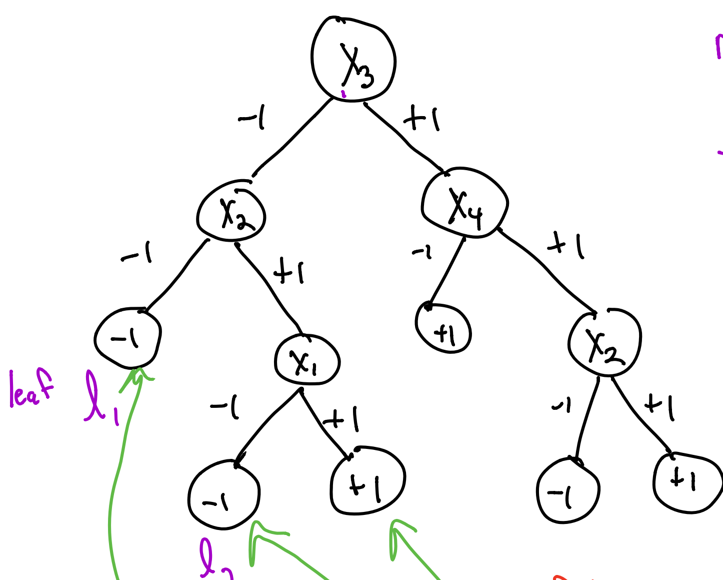
& so $\overline{AND}(x) = 2f(x) - 1$

$$= \underbrace{-1 + \frac{2}{2^k}}_{\text{Fourier coeff for } \chi_\phi} + \sum_{\substack{S \subseteq T \\ |S| > 0}} \frac{(-1)^{|S|}}{2^{k-1}} \cdot \chi_S$$

note all Fourier coeffs containing any var not in T are 0

Example 2 Decision Trees of depth $\leq k$

input $x = (x_1 \cdots x_n)$



can write: indicator function for inputs that reach leaf ℓ_i $\left(\begin{matrix} 1 & \text{if reach } \ell_i \\ 0 & \text{o.w.} \end{matrix}\right)$

example:

$$f_{\ell_1}(x) = \frac{(1-x_3)}{2} \cdot \frac{(1-x_2)}{2}$$

$$f_{\ell_2}(x) = \frac{(1-x_3)}{2} \cdot \frac{(1+x_2)}{2} \cdot \frac{(1-x_1)}{2}$$

$$f(x) = -1 \cdot f_{\ell_1}(x) + -1 \cdot f_{\ell_2}(x) + (+1)\, f_{\ell_3}(x) + \ldots$$

Consider path indicator fctns:

$$f_\ell(x) = \prod_{\underbrace{i \in V_\ell}_{\text{vars visited on path to leaf } \ell}} \frac{(1 \pm x_i)}{2}$$

left or right
$-1 \rightsquigarrow$ left
$+1 \rightsquigarrow$ right

$$= \frac{1}{2^{|V_\ell|}} \sum_{S \subseteq V_\ell} (\pm 1)\, \chi_S$$

$(-1)^{\#\text{ left turns } \in S \text{ taken in on path to } \ell}$

$$= \begin{cases} 1 & \text{if } x \text{ takes path } \ell \\ 0 & \text{o.w.} \end{cases}$$

So $f(x) = \sum_{\ell \in \text{leaves of } T} \underbrace{f_\ell(x)}_{\text{exactly one of these is 1, all the rest} = 0} \cdot \mathrm{val}(\ell)$

Fourier rep of f

$$\hat{f}_S(x) = \sum_\ell \hat{f}_\ell(x) \cdot \mathrm{val}(\ell)$$

since if $h = f + g$
then $\hat{h}(x) = \hat{f}(x) + \hat{g}(x)$

$\Rightarrow$ only coeffs corresponding to S s.t. $|S| \le$ max path length are non zero

Low degree algorithm:

def $f: \{\pm 1\}^n \to \mathbb{R}$ has $\alpha(\varepsilon, n)$ - Fourier concentration

if $\sum_{\substack{S \subseteq [n] \\ \text{s.t.} \\ |S| > \alpha(\varepsilon, n)}} \hat{f}(S)^2 \leq \varepsilon$

for Boolean f this implies via "Boolean Parseval's"

$$\sum_{\substack{S \text{ s.t.} \\ |S| \leq \alpha(\varepsilon, n)}} \hat{f}(S)^2 \geq 1 - \varepsilon$$

reminder:

Claim if f doesn't depend on x_i

then all $\hat{f}(S)$ for which $i \in S$ satisfy $\hat{f}(S) = 0$

why? usual pairing of inputs for $x_i = +1$, $x_i = -1$

examples

1) any fctn f on $\leq k$ vars

via claim

2) any fctn f on decision tree of depth $\leq k$

3) $f = \text{AND}$ on $T \subseteq \{1 \ldots n\}$ has $\log(\frac{4}{\varepsilon})$-F.C.

no requirement on $|T|$

- all $\hat{f}(S)^2 = 0$ for $|S| > |T|$ (by claim)

- if $|T| \leq \log \frac{4}{\varepsilon}$ ✓

- if $|T| \geq \log \frac{4}{\varepsilon}$

  f is almost always "False"

$$\hat{f}(\emptyset)^2 = (1 - 2\Pr(f(x) \neq \chi_\emptyset(x))^2$$

$\equiv 1$ which is "False"

$$= \left(1 - 2 \cdot \frac{1}{2^{|T|}}\right)^2 > 1 - \varepsilon$$

so $\sum_{S \neq \emptyset} \hat{f}(S)^2 \leq \varepsilon$ & f has 0-F.C.

Now, lets approx fctns with low F.C.:

Low degree Algorithm: goal approx f with $d = \alpha(\varepsilon, n)$ F.C.

Given
d degree
τ accuracy
δ confidence

Algorithm

- Take $m = O\left(\frac{n^d}{\tau} \ln \frac{n^d}{\delta}\right)$ samples
- $C_s \leftarrow$ estimate of $\hat{f}(s)$ for each S s.t. $|S| \le d$

(reuse samples for $\le \binom{n}{d}$ different S's)

- output $\text{sign}\left(\underbrace{h(x) = \sum_{|S| \le d} C_S \chi_S(x)}_{h \text{ not necessarily } \pm 1}\right)$

as hypothesis

Why does this work?

Two stages:

1) show if f has low F.C.
   then $E_x[(f(x)-h(x))^2]$ small

2) show $\Pr[f(x) \neq \mathrm{sign}(h(x))] \leq E_x[(f(x)-h(x))^2]$

$\uparrow$ Hamming dist

② First:

<u>Thm</u> $f: \{\pm 1\}^n \to \{\pm 1\}$
$h: \{\pm 1\}^n \to \mathbb{R}$

then $\Pr[f(x) \neq \mathrm{sign}(h(x))] \leq E[(f(x)-h(x))^2]$

<u>Pf</u>. $E[(f(x)-h(x))^2] = \frac{1}{2^n} \sum (f(x)-h(x))^2$ defn

$$\wedge|$$

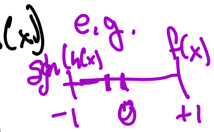
$\Pr[f(x) \neq \mathrm{sign}(h(x))] = \frac{1}{2^n} \sum_x \mathbb{1}_{\{f(x) \neq \mathrm{sign}(h(x))\}}$

show than "term by term"

2 cases: $f(x) = \mathrm{sign}(h(x))$
$(f(x)-h(x))^2 \geq 0$
$\mathbin{\rotatebox{90}{$\leq$}}$
$\mathbb{1}_{f(x) \neq \mathrm{sign}(h(x))} = 0$

| $f(x) \neq \mathrm{sign}(h(x))$
$(f(x)-h(x))^2 \geq 1$
$\mathbin{\rotatebox{90}{$\leq$}}$
$\mathbb{1}_{f(x) \neq \mathrm{sign}(h(x))} = 1$

e.g.



so $\forall x \quad (f(x)-h(x))^2 \geq \mathbb{1}_{f(x) \neq \text{sign}(h(x))}$

$\Rightarrow$ thm $\blacksquare$