Fourier Learning

Monday, November 1, 2021 6:38 AM

yhh

Motivated by learning DNF.

Still an open problem to PAC learn DNF or decision trees (lists we can learn).

So we consider special distributions

- uniform
- product
- Gaussian etc

and allow for membership queries, i.e., "what is the label of x?".

Assume $x \in \{-1, 1\}^n$ $f: \{-1, 1\}^n \to \{-1, 1\}$

Note that any such $f \in \{-1, 1\}^{2^n}$ (as a table)

Define inner product of f and g with respect to distribution D as

$$\langle f, g \rangle_D = \sum D(x) f(x) g(x)$$

$$\langle f, g\rangle_D = \sum_x D(x) f(x) g(x)$$

$$\langle f, f\rangle_D = \|f\|_D^2 = 1 \quad (\text{since } f(x)^2 = 1)$$

Viewing f as a vector, the standard basis is $e_1, e_2, \ldots, e_{2^n}$.

But we can use any basis and write

$$f(x) = \sum_v \langle f, v\rangle v$$ since $\{v\}$ is an orthonormal basis.

What's an interesting basis?

The set of parity functions!

$$\forall S \subseteq [n], \quad \chi_S(x) = \prod_{i \in S} x_i \qquad 2^n \text{ functions.}$$

$$\langle \chi_S, \chi_S\rangle_D = 1$$

$$\langle \chi_S, \chi_T\rangle_D = \mathbb{E}_D\left(\prod_{i \in S} x_i \prod_{j \in T} x_j\right) = 0$$

for a product distribution D.

for a product distribution D.

Hence $\{\chi_S\}$ is an orthonormal basis!

So any f can be written as

$$f(x) = \sum_S \hat{f}_S \chi_S(x) \quad \text{where } \hat{f}_S = \langle f, \chi_S \rangle_D .$$

Thm 1. (Parseval) $\langle f, f \rangle_D = \langle \hat{f}, \hat{f} \rangle$

Thm 2. (Plancherel) $\langle f, g \rangle_D = \langle \hat{f}, \hat{g} \rangle$.

Pf. $$\sum_x D(x) \sum_S \hat{f}_S \chi_S(x) \sum_T \hat{g}_T \chi_T(x)$$

$$= \sum_{S,T} \hat{f}_S \hat{g}_T \mathbb{E}_D (\chi_S(x) \chi_T(x))$$

$$= \sum_S \hat{f}_S \hat{g}_S = \langle \hat{f}, \hat{g} \rangle .$$

A decision tree is a Boolean function f.

We want to learn f by approximating all

We want to learn f by approximating ...
of its significant Fourier coefficients $\hat{f}_S$.

Our approx is g.

$$\Pr_D\left(g(x) \neq f(x)\right) \leq \mathbb{E}_D\left(\left(f(x)-g(x)\right)^2\right) = \sum_S \left(\hat{f}_S - \hat{g}_S\right)^2.$$

We will learn all $\hat{f}_S$ for which $\hat{f}_S \geq \tau$.

Note. $\sum_S \hat{f}_S^2 = 1 \Rightarrow |\hat{f}_S| \leq 1$

Lemma[DNF]. If a decision tree has m leaves
then $\|f\|_1 = \sum |\hat{f}_S| \leq 2m+1$.

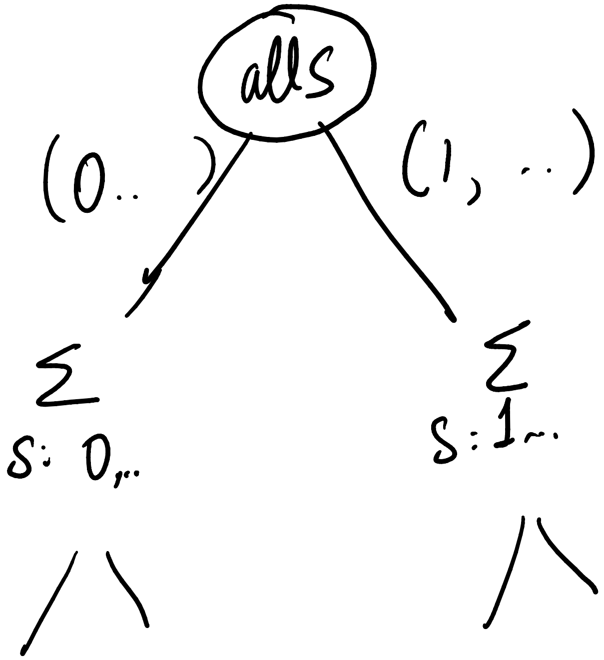
Thm. If we learn all $\hat{f}_S \geq \dfrac{\varepsilon}{\|f\|_1}$
then $\|\hat{f}_S - \hat{g}_S\|^2 \leq \varepsilon$.

Pf. $$\|\hat{f}_S - \hat{g}_S\|^2 = \sum_{S : |\hat{f}_S| \leq \frac{\varepsilon}{\|\hat{f}\|_1}} \hat{f}_S^2 \leq \sum_S |\hat{f}_S| \cdot \frac{\varepsilon}{\|\hat{f}\|_1}$$

$$S : |f_S| \geq \frac{\varepsilon}{\|\hat{f}\|_1}$$

$$= \varepsilon.$$

How to learn all large Fourier coefficients.



At each node estimate whether

$$\sum_{S:\ \text{prefix}\ \alpha} \hat{f}_S^2 \geq \tau.$$

width $\leq \frac{1}{\tau}$, depth $\leq n$

\# nodes $\leq \frac{n}{\tau}$.

How to estimate

$$\sum_{S \in S_\alpha} \hat{f}_S^2 \ ?$$

Suppose $\alpha = \underbrace{(0, \dots 0)}_{k}$.

Claim $\quad \sum \hat{f}^2 \quad \mathbb{E}\left(f(yx) f(zx)\right)$

Claim. $\sum_{S_\alpha} \hat{f}_S^2 = \mathbb{E}_{x \sim \{0,1\}^{n-k},\, y, z \sim \{0,1\}^k} \left(f(yx) f(zx) \right)$

Suppose f is a parity function
if f agrees with α, then $f(yx) = f(zx)$ so we get 1.
else $\Pr(f(yx) = f(zx)) = \frac{1}{2} \Rightarrow$ we get 0.
Any f can be written as a weighted sum of parities. So,

$$f = \sum_U \hat{f}_U \chi_U$$

$$\mathbb{E}(f(yx) f(zx)) = \mathbb{E}\left(\sum_U \hat{f}_U \chi_U(yx) \sum_V \hat{f}_V \chi_V(zx) \right)$$

$$= \sum_{U,V} \hat{f}_U \hat{f}_V \underbrace{\mathbb{E}\left(\chi_U(yx) \chi_V(zx)\right)}_{=0 \text{ if } U \neq V}$$

$$= \sum_U \hat{f}_U^2 \underbrace{\mathbb{E}(\chi_U(yx) \chi_U(zx))}_{=0 \text{ if } U \text{ does not}}$$

$$= \sum_{U \in U_\alpha} \hat{f}_U^2 .$$

$= 0$ if U does not agree with $\alpha = (0, \dots, 0)$

What about general α?

<u>Lemma</u>. $$\sum_{S \in S_\alpha} \hat{f}_S^2 = \underset{\substack{x \sim \{0,1\}^{n-k} \\ y, z \sim \{0,1\}^k}}{\mathbb{E}} \left(f(yx) f(zx) \chi_\alpha(y) \chi_\alpha(z) \right).$$

<u>Proof</u> [of Lemma [DNF]].

Consider a single conjunction T. Let $T(x) = 1$ if x satisfies it and $T(x) = 0$ o.w.

Then $\langle T, T \rangle_D = \mathbb{E}(T(x)^2) = \frac{1}{2^{|T|}}$

$$\hat{T}_S = \langle T, \chi_S \rangle_D$$

$$= \mathbb{E}_D (T(x) \chi_S(x))$$

$$\mathbb{E} (\chi_S(x) \mid T(x) = 1) = \begin{cases} 0 & \text{if } T \text{ contains } x_i \notin S \end{cases}$$

$$= \mathbb{E}_D(\chi_S(x) / T(x)=1) = \begin{cases} 0 & x_i \notin S \\ \frac{1}{2^{|T|}} & \text{o.w.} \end{cases}$$

So $\|\hat{T}\|_1 = 1$ $\quad \|\hat{T}\|_2^2 = 2^{|T|} \cdot \frac{1}{2^{2|T|}} = \frac{1}{2^{|T|}}$.

For a decision tree with m leaves, we can write

$$f(x) = 2(T_1(x) + \ldots T_m(x)) - 1$$

So $\|\hat{f}\|_1 \leq 2 \sum_{i=1}^{m} |\hat{T}_i|_1 + 1 \leq 2m+1$.