Computational Learning Theory
Learning Decision Trees via the Fourier Transform

Lecturer: James Worrell

Introduction

In the following two lectures we present an algorithm, due to Kushilevitz and Mansour, for learning Boolean functions represented as decision trees. We work within a model in which the learner has query access to the target function and must produce with high probability a hypothesis that has low error with respect to the uniform distribution on the input space. Specifically, if $f : \{0,1\}^n \to \{-1,+1\}$ is the target function and $\varepsilon, \delta > 0$ are given accuracy and confidence parameters, we require the learner to output a hypothesis $h : \{0,1\}^n \to \{-1,+1\}$ such that with probability at least $1-\delta$ (with respect to the internal randomisation of the learner)

$$\Pr_{x \sim U_n}(h(x) \neq f(x)) \leq \varepsilon \,,$$

where U_n denotes the uniform distribution on $\{0,1\}^n$. We further require that learning algorithm to run in time polynomial in $n, \frac{1}{\varepsilon}, \frac{1}{\delta}$ and the size m of the smallest decision tree representing the function.

The basic idea of the algorithm of Kushilevitz and Mansour is to approximate the target function by using membership queries to compute a polynomial-size subset of its Fourier coefficients, namely those that are suitably large in magnitude. Note that no polynomial-time algorithm is known that PAC learns decision trees under the uniform distribution (i.e., when we only have access to uniformly distributed random examples rather than membership queries).

Background

A decision tree on n Boolean variables is a binary tree such that the leaves have labels chosen from $\{-1,+1\}$, and the internal nodes have exactly two children (respectively distinguished as the left child and right child) and labels chosen from $\{1,\ldots,n\}$. A vector $x \in \{0,1\}^n$ determines a path from the root to a unique leaf according to the rule *at an internal node with label i, take the left child if $x_i = 0$ and take the right child if $x_i = 1$*. Such a decision tree determines a function $\{0,1\}^n \to \{-1,+1\}$ in which each input is mapped to the label of the leaf that it determines.

We consider $\mathbb{R}^{\{0,1\}^n}$ as a real vector space, with inner product given by

$$\langle f, g\rangle := \mathop{\mathbb{E}}_{x\sim U_n}[f(x)g(x)] = \frac{1}{2^n}\sum_{x\in\{0,1\}^n} f(x)g(x)\,,$$

i.e., a scaled version of the usual inner product. We define the norm of $f \in \mathbb{R}^{\{0,1\}^n}$ to be

$$\|f\|_2 := \langle f, f\rangle^{1/2} = \mathbb{E}_x[f(x)^2]^{1/2}\,.$$

With respect to the inner product above, the collection of parity functions $\{\chi_\alpha : \alpha \in \{0,1\}^n\}$, given by

$$\chi_\alpha(x) = \begin{cases} +1 & \text{if } x\cdot\alpha \text{ is even} \\ -1 & \text{if } x\cdot\alpha \text{ is odd} \end{cases},$$

forms an orthonormal basis. (We leave it as an exercise to verify this claim.) In particular, every function $f : \{0,1\}^n \to \mathbb{R}$ can be written as a linear combination $f = \sum_\alpha \hat{f}(\alpha)\chi_\alpha$, where $\hat{f}(\alpha) := \langle f, \chi_\alpha\rangle$. The coefficients in this expansion are called the *Fourier coefficients* of f.

We conclude by stating two useful properties of the functions χ_α. We leave the proofs as exercises.

- Parseval's identity: for every f, $\|f\|_2^2 = \sum_\alpha |\hat{f}(\alpha)|^2$.
- For all $\alpha, \beta \in \{0,1\}^n$, $\chi_\alpha \chi_\beta = \chi_{\alpha \oplus \beta}$ (where the product of functions is pointwise and $\oplus$ denotes pointwise exclusive or).

Approximating Decision Trees

A Boolean function $f : \{0,1\}^n \to \{-1,+1\}$ necessarily satisfies $\|f\|_2 = 1$. For such an f we have $\sum_\alpha |\hat{f}(\alpha)|^2 = 1$ by Parseval's identity. However the sum $\sum_\alpha |\hat{f}(\alpha)|$ can potentially be exponential in n (specifically when the coefficients are "spread out", i.e., all roughly equal). Intuitively, the following proposition says that the vector of Fourier coefficients of a decision tree has relatively few large values.

Proposition 1. *If $f : \{0,1\}^n \to \{-1,+1\}$ is represented by a decision tree with m leaves then $\sum_\alpha |\hat{f}(\alpha)| \leq m$.*

Proof. Let L denote the set of leaves of the decision tree. Given $v \in L$, let ℓ_v be its label, $I(v) \subseteq \{0,1\}^n$ the set of inputs leading to v, and $\mathrm{Vars}(v) \subseteq \{1,\ldots,n\}$ the set of labels occurring along the unique path from the root to v. Then we have

$$\begin{aligned} \sum_{\alpha \in \{0,1\}^n} |\hat{f}(\alpha)| &= \frac{1}{2^n} \sum_{\alpha \in \{0,1\}^n} \Big| \sum_{x \in \{0,1\}^n} f(x)\chi_\alpha(x) \Big| \\ &= \frac{1}{2^n} \sum_{\alpha \in \{0,1\}^n} \Big| \sum_{v \in L} \sum_{x \in I(v)} \ell_v \, \chi_\alpha(x) \Big| \\ &\leq \frac{1}{2^n} \sum_{\alpha \in \{0,1\}^n} \sum_{v \in L} \underbrace{\Big| \sum_{x \in I(v)} \ell_v \, \chi_\alpha(x) \Big|}_{S_{\alpha,v}} . \end{aligned}$$

Now if α and v are such that $\alpha_i = 0$ for all $i \notin \mathrm{Vars}(v)$ then χ_α is constant on $I(v)$ and hence

$$S_{\alpha,v} = |I(v)| = 2^{n-|\mathrm{Vars}(v)|};$$

otherwise $S_{\alpha,v} = 0$. For each $v \in L$ the number of α such that $\alpha_i = 0$ for all $i \notin \mathrm{Vars}(v)$ is $2^{|\mathrm{Vars}(v)|}$. Thus we have

$$\sum_{\alpha \in \{0,1\}^n} |\hat{f}(\alpha)| \leq \frac{1}{2^n} \cdot m \cdot 2^{|\mathrm{Vars}(v)|} \cdot 2^{n-|\mathrm{Vars}(v)|} = m .$$

□

The benefit of the above concentration result is that we can approximate f by a Fourier expansion in which we take only the largest few Fourier coefficients.

Proposition 2. *Let $f : \{0,1\}^n \to \{-1,+1\}$ and m be such that $\sum_\alpha |\hat{f}(\alpha)| \leq m$ and let $\varepsilon > 0$ be given. Define $g : \{0,1\}^n \to \mathbb{R}$ by $\displaystyle\sum_{\alpha : |\hat{f}(\alpha)| \geq \frac{\varepsilon}{m}} \hat{f}(\alpha)\chi_\alpha$. Then $\|f - g\|_2^2 \leq \varepsilon$.*

Proof. We have that $f - g = \sum_{\alpha : |\hat{f}(\alpha)| < \frac{\varepsilon}{m}} \hat{f}(\alpha)\chi_\alpha$. Thus, by Parseval's identity,

$$
\begin{aligned}
\|f - g\|_2^2 &= \sum_{\alpha : |\hat{f}(\alpha)| < \frac{\varepsilon}{m}} |\hat{f}(\alpha)|^2 \\
&< \frac{\varepsilon}{m} \sum_{\alpha : |\hat{f}(\alpha)| < \frac{\varepsilon}{m}} |\hat{f}(\alpha)| \\
&\leq \frac{\varepsilon}{m} m \quad \text{by Proposition 1} \\
&= \varepsilon \,.
\end{aligned}
$$

□

Corollary 1. *Let f, g be as in Proposition 2. Define $h : \{0,1\}^n \to \{-1,+1\}$ by $h(x) := \operatorname{sgn}(g(x))$. Then*

$$\Pr_{x \sim U_n}(h(x) \neq f(x)) \leq \frac{\varepsilon}{4}$$

Proof. We have

$$\mathbb{I}(h(x) \neq f(x)) = \tfrac{1}{4}|h(x) - f(x)|^2 \leq \tfrac{1}{4}|g(x) - f(x)|^2 \,.$$

Hence

$$
\begin{aligned}
\Pr_x(h(x) \neq f(x)) &= \mathbb{E}_x[\mathbb{I}(h(x) \neq f(x))] \\
&\leq \frac{1}{4}\mathbb{E}[(g(x) - f(x))^2] \\
&= \frac{1}{4}\|g - f\|_2^2 \\
&\leq \frac{\varepsilon}{4} \quad \text{by Proposition 2} \,.
\end{aligned}
$$

□

Say that a Fourier coefficient $\hat{f}(\alpha)$ is "large" if it has magnitude at least ε/m. Since $\|f\|_2^2 = 1$, the number of large Fourier coefficients is at most m^2/ε^2, i.e., polynomial in $1/\varepsilon$ and m. The rough idea of the learning algorithm is to use membership queries to find all large Fourier coefficients and to form the hypothesis h described in Corollary 1. The tricky part, to be described in the next lecture, is to efficiently find these large elements within the set of all 2^n Fourier coefficients. Using membership queries and Hoeffding's inequality we can estimate any given Fourier coefficient to high precision, but brute-force examination of all Fourier coefficients would clearly yield a running time exponential in n.