CS 395T Computational Complexity of Machine Learning

Lecture 14: March 8, 2005 Lecturer: Adam Klivans

Scribe: Alex Sherstov

The Harmonic Sieve

This lecture presents a polynomial-time algorithm based on boosting for learning DNF formulas to arbitrary precision. The algorithm is due to Jackson [1].

14.1 Preliminaries

Let $f = T_1 \vee T_2 \vee \cdots \vee T_s$ be a Boolean function in DNF on n variables. One of $T_1, T_2, \ldots, T_s$ must be true on at least $1/s$ of the satisfying assignments to f; denote some such term by T. The analysis below treats f and T as Boolean functions from $\{-1,1\}^n$ to $\{-1,1\}$; we denote by $f^{\{0,1\}}$ and $T^{\{0,1\}}$ their counterparts from $\{-1,1\}^n$ to $\{0,1\}$.

Let D be an arbitrary distribution over $\{-1,1\}^n$. We start by relating $\mathop{\mathrm{E}}_{D}[f]$ and $\mathop{\mathrm{E}}_{D}\left[f^{\{0,1\}}\right]$:

$$\begin{aligned}
\mathop{\mathrm{E}}_{D}[f] &= -1 \cdot \Pr_{D}[f = -1] + 1 \cdot \Pr_{D}[f = 1] \\
&= -1 \cdot \left(1 - \mathop{\mathrm{E}}_{D}\left[f^{\{0,1\}}\right]\right) + 1 \cdot \mathop{\mathrm{E}}_{D}\left[f^{\{0,1\}}\right] \\
&= 2\mathop{\mathrm{E}}_{D}\left[f^{\{0,1\}}\right] - 1.
\end{aligned}$$

Thus,

$$\mathop{\mathrm{E}}_{D}\left[f^{\{0,1\}}\right] = \frac{\mathop{\mathrm{E}}_{D}[f] + 1}{2}. \tag{14.1}$$

Second, we represent $T^{\{0,1\}}$ as its Fourier expansion. Let $V \subseteq \{x_1, x_2, \ldots, x_n\}$ be the subset of variables featured in $T^{\{0,1\}}$. Without loss of generality, assume T does not feature *negations* of variables.

$$\begin{aligned}
T^{\{0,1\}} &= \prod_{x \in V} \frac{1 - \chi_{\{x\}}}{2} = \sum_{A \subseteq V} \frac{(-1)^{|A|}\chi_A}{2^{|V|}} \\
&= \mathop{\mathrm{E}}_{A}\left[(-1)^{|A|}\chi_A\right].
\end{aligned} \tag{14.2}$$

The presence of any negated variables in T will affect only the *signs* of the terms in the above summation. Since the analysis below does not depend on these signs, no loss of generality is incurred.

14.2 Weakly Learning DNF's

This section will show that any DNF f with s terms has an $\Omega(1/s)$ correlation with some parity function χ. Combined with the KM algorithm for identifying large Fourier coefficients with queries, this result will yield a weak learner for polynomial-size DNF's.

Using (14.1), we obtain:

$$\begin{aligned}\mathop{\mathrm{E}}_{D}\left[f \cdot T^{\{0,1\}}\right] = \mathop{\mathrm{E}}_{D}\left[T^{\{0,1\}}\right] &\geq \frac{\mathop{\mathrm{E}}_{D}\left[f^{\{0,1\}}\right]}{s} \\ &= \frac{\mathop{\mathrm{E}}_{D}[f]+1}{2s}. \end{aligned} \tag{14.3}$$

On the other hand, (14.2) yields:

$$\begin{aligned}\mathop{\mathrm{E}}_{D}\left[f \cdot T^{\{0,1\}}\right] &= \mathop{\mathrm{E}}_{D}\left[f \cdot \mathop{\mathrm{E}}_{A}\left[(-1)^{|A|}\chi_A\right]\right] = \mathop{\mathrm{E}}_{A}\left[\mathop{\mathrm{E}}_{D}\left[f \cdot (-1)^{|A|}\chi_A\right]\right] \\ &\leq \mathop{\mathrm{E}}_{A}\left[\left|\mathop{\mathrm{E}}_{D}\left[f \cdot (-1)^{|A|}\chi_A\right]\right|\right] \\ &= \mathop{\mathrm{E}}_{A}\left[\left|\mathop{\mathrm{E}}_{D}[f \cdot \chi_A]\right|\right]. \end{aligned} \tag{14.4}$$

Combing (14.3) and (14.4) yields the following inequality:

$$\frac{\mathop{\mathrm{E}}_{D}[f]+1}{2s} \leq \mathop{\mathrm{E}}_{D}\left[f \cdot T^{\{0,1\}}\right] \leq \mathop{\mathrm{E}}_{A}\left[\left|\mathop{\mathrm{E}}_{D}[f \cdot \chi_A]\right|\right],$$

and

$$\frac{\mathop{\mathrm{E}}_{D}[f]+1}{2s} \leq \mathop{\mathrm{E}}_{A}\left[\left|\mathop{\mathrm{E}}_{D}[f \cdot \chi_A]\right|\right].$$

Therefore, for *some* parity function χ_A,

$$\left|\mathop{\mathrm{E}}_{D}[f \cdot \chi_A]\right| \geq \frac{\mathop{\mathrm{E}}_{D}[f]+1}{2s}.$$

Consider two cases:

- **Case 1:** $\mathrm{E}_D[f] \geq -1/(2s+1)$. The parity function χ_A above has a large correlation with f:

$$\left|\mathop{\mathrm{E}}_{D}[f \cdot \chi_A]\right| \geq \frac{\mathop{\mathrm{E}}_{D}[f]+1}{2s} \geq \frac{2s+1-1}{2s \cdot (2s+1)} = \frac{1}{2s+1}.$$

- **Case 2:** $\mathrm{E}_D[f] < -1/(2s+1)$. The parity function $\chi_\emptyset = 1$ has a large correlation with f:

$$\left|\mathop{\mathrm{E}}_{D}[\chi_\emptyset \cdot f]\right| = \left|\mathop{\mathrm{E}}_{D}[f]\right| \geq \frac{1}{2s+1}.$$

In either case above, there exists some parity function χ_A with correlation at least $1/(2s+1)$ with f over distribution D. In other words, $\Pr_D[f = \chi_A] \geq 1/2 + 1/(2s+1)$. Thus, identifying the parity function χ_A amounts to learning f over distribution D with advantage $1/(2s+1) = \Omega(1/s)$.

14.3 Boosting

Boosting is a technique for learning a concept class to arbitrary precision given only a weak learner. More specifically, a boosting algorithm receives:

1. a weak learner with advantage γ over an arbitrary distribution,
2. an accuracy requirement ϵ,
3. a success probability parameter δ,
4. access to the target function f;

and with probability $1-\delta$ produces a hypothesis h with $\Pr_x[h(x) = f(x)] \geq 1-\epsilon$.

Boosting works in stages. At stage 1, the algorithm uses the weak learner to produce hypothesis h_1 with accuracy $\Pr_{x\sim U}[h_1(x) = f(x)] \geq 1/2+\gamma$. At stage i, the algorithm constructs a distribution D_i that puts more weight on examples labeled incorrectly by $h_1, h_2, \ldots, h_{i-1}$, and uses the weak learner to obtain a hypothesis h_i with accuracy $\Pr_{x\sim D_i}[h_i(x) = f(x)] \geq 1/2+\gamma$. At the end, the algorithm outputs some combining function of the hypotheses $h_1, h_2, \ldots, h_l$ (e.g., a majority classifier). There are boosting algorithms that terminate within $l = O(\frac{1}{\gamma^2}\log\frac{1}{\epsilon})$ stages and weight no example more than $\frac{1}{2^n\epsilon}$.

14.4 Putting It All Together

A complication in using boosting to learn DNF formulas to arbitrary precision is the requirement that the weak learner operate over an arbitrary distribution D. In particular, the KM algorithm identifies large Fourier coefficients with respect to the uniform distribution, i.e., $|\mathrm{E}_{x\sim U}[f\cdot\chi_A]| \geq \theta$.

Observe that

$$\mathrm{E}_{x\sim D}[f(x)\chi_A(x)] = \sum_x D(x)f(x)\chi_A(x) = \frac{1}{2^n}\sum_x 2^n D(x)f(x)\chi_A(x) = \mathrm{E}_{x\sim U}[2^n D(x)f(x)\chi_A(x)].$$

Thus, identifying χ_A with $\mathrm{E}_D[f\cdot\chi_A] \geq \theta$ is tantamount to identifying a Fourier coefficient of $g(x) = 2^n D(x)f(x)$ with absolute value θ or more. The distributions D_i constructed in a boosting algorithm are *known*, so any query to g can be answered via a query to f.

An analysis of KM for non-Boolean functions, such as g, yields a running time polynomial in $1/\epsilon$, $1/\delta$, $1/\theta$, and $L_\infty(g)$. In our case, $1/\theta = 2s+1$ and $L_\infty(g) \leq 2^n \cdot 1/(2^n\epsilon)\cdot 1 = 1/\epsilon$; the latter bound follows because there are boosting algorithms that enforce $|D_i| \leq 1/(2^n\epsilon)$ for all i. As a result, DNF formulas can be learned to arbitrary precision via a boosting algorithm that uses KM as a weak learner. The time requirement of this implementation is polynomial in the DNF size s as well as the usual parameters $1/\epsilon$ and $1/\delta$.

References

[1] Jeffrey Charles Jackson. *The harmonic sieve: a novel application of Fourier analysis to machine learning theory and practice*. PhD thesis, Carnegie Mellon University, 1995.