CS 395T Computational Complexity of Machine Learning

Lecture 6: February 8, 2005 Lecturer: Adam Klivans

Scribe: Alex Sherstov

Learning Intersections of Halfspaces

6.1 Approximate Representations of Functions

In a previous lecture, we introduced the notion of a *weak representation* of a function. An alternative formalism for approximate representation is based on the real-valued difference between the function's true value and the output of the approximating polynomial. For example, we may be interested in finding a polynomial p satisfying $|p(x) - \textsc{or}(x)| \leq \epsilon$ for all x.

For the simpler functions (such as AND,OR), a degree-$\Theta(\sqrt{n})$ polynomial suffices. For more complicated functions (such as MAJORITY), a degree-n polynomial is necessary. Note that in this formalism, *any* function can be approximated to arbitrary accuracy with a degree-n polynomial. It suffices (i) to form the sum of the 2^n polynomials each of which evaluates to 1 on a distinct assignment to the variables and to 0 on all others; and (ii) to weigh these polynomials by the output of the actual function.

6.2 Intersections of Halfspaces

Previous lectures looked at the problem of learning a *halfspace*. In this lecture, we turn our attention to learning *intersections* of two or more halfspaces. We would like to generalize our results to arbitrary functions on multiple halfspaces.

Consider two halfspaces given by the PTF's $f(x) = \text{sign}(\hat{f}(x))$ and $g(x) = \text{sign}(\hat{g}(x))$, where $\hat{f}(x) = \sum_{i=1}^n \alpha_i x_i - \theta_1$ and $\hat{g}(x) = \sum_{i=1}^n \beta_i x_i - \theta_2$ are degree-1 polynomials in n variables. We would like to learn the function $f(x) \wedge g(x)$. As any other function on Booleans, $f(x) \wedge g(x)$ can be represented by a degree-n PTF. On the other hand, a lower bound on the degree of such a PTF is $\frac{\log n}{\log \log n}$. Our goal is to discover a low-degree PTF representation. This problem remains unsolved for the general case. Our solution will assume that the weights in $\hat{f}$ and $\hat{g}$ are bounded by W:

$$\sum_{i=1}^n |\alpha_i| + |\theta_1| \leq W \qquad \text{and} \qquad \sum_{i=1}^n |\beta_i| + |\theta_1| \leq W.$$

We will additionally assume that $f(x) \implies \hat{f}(x) \geq 1$ and $\neg f(x) \implies \hat{f}(x) \leq -1$; likewise for g. Without this additional requirement, the restriction W on the weights is irrelevant and can always be achieved by normalizing the weights in a suitable way. Under these assumptions, we will prove the existence of a degree-$O(\log W)$ PTF representing the intersection. This is a nontrivial result if the weights are small: $W \ll 2^n$.

6.3 Approximating the Sign Function

If it were possible to represent the sign function accurately with a low-degree polynomial q, our goal would be accomplished: the PTF representing the intersection would be simply $q(\hat{f}(x))+q(\hat{g}(x)) \geq 2$. This is the strategy we will adopt in this lecture. Our first step is to supply a low-degree polynomial approximation to the sign function. Namely, consider the following univariate functions:

$$p_l(x) = (x-1)\prod_{i=1}^{l}(x-2^l)^2 \quad \text{and} \quad s_l(x) = \frac{-p_l(-x)+p_l(x)}{-p_l(-x)-p_l(x)}.$$

Lemma 1 *The functions p_l and s_l above satisfy:*

(a) $0 \leq 4p_l(x) \leq -p_l(-x)$ *for* $1 \leq x \leq 2^l$;

(b) $1 \leq s_l(x) \leq 5/3$ *for* $1 \leq x \leq 2^l$, *and*
$-5/3 \leq s_l(x) \leq -1$ *for* $-2^l \leq x \leq -1$.

Proof: We prove (a) first. Trivially, $4p_l(x) \geq 0$. It remains to show that $4p_l(x) \leq -p_l(-x)$. Let $x \in [2^k, 2^{k+1}]$, where $k < l$. Then $4p_l$ and $-p_l(-x)$ can be rewritten as:

$$\begin{aligned} 4p_l(x) &= [4(x-2^k)^2] \cdot (x-1)(x-2)^2(x-4)^2 \cdots (x-2^{k-1})^2(x-2^{k+1})^2 \cdots (x-2^l)^2, \\ -p_l(-x) &= [(x+2^k)^2] \cdot (x+1)(x+2)^2(x+4)^2 \cdots (x+2^{k-1})^2(x+2^{k+1})^2 \cdots (x+2^l)^2. \end{aligned}$$

Every factor in the expression for $-p_l(-x)$, except for the first, is strictly greater than the corresponding factor in the expression for $4p_l(x)$. Thus, it is sufficient to show that $4(x-2^k)^2 \leq (x+2^{k-1})$. We have:

$$\begin{aligned} 4(x-2^k)^2 &\leq 4(2^{k+1}-2^k)^2 = 4 \cdot 2^{2k} = 2^{2k+2}, \qquad \text{and} \\ (x+2^k)^2 &\geq (2^k+2^k)^2 = 2^{2k+2}. \end{aligned}$$

This completes the proof of part (a).

To prove part (b), consider first the case $x \in [1, 2^l]$. An upper bound on $s_l(x) = \frac{-p_l(-x)+p_l(x)}{-p_l(-x)-p_l(x)}$ is obtained by substituting $-p_l(-x)/4$ for $p_l(x)$ and noting that $-p_l(-x) \neq 0$; this yields $s_l(x) \leq 5/3$. A lower bound is obtained by substituting 0 for $p_l(x)$, which yields $s_l(-x) \geq 1$. The argument for $x \in [-2^l, -1]$ is analogous. ■

6.4 A More Accurate Approximation of the Sign Function

Thus, the rational function $s_l(x)$ approximates the sign function within 2/3. The numerator and denominator both have degree $2l+1$. A better approximation can be obtained using:

$$s_l^{\log t}(x) \stackrel{\text{def}}{=} \frac{(-p_l(-x))^{\log t} + (p_l(x))^{\log t}}{(-p_l(-x))^{\log t} - (p_l(x))^{\log t}},$$

where $\log t \geq 1$ is an *odd* integer, to ensure that the sign is preserved. The quality of the approximation for $t \geq 3$ is given by the following lemma:

Lemma 2 *The function $s_l^{\log t}$ for $t \geq 3$ satisfies:*

(a) $1 \leq s_l^{\log t}(x) \leq 1 + 1/t$ *for* $1 \leq x \leq 2^l$, *and*

(b) $-1 - 1/t \leq s_l^{\log t}(x) \leq -1$ *for* $-2^l \leq x \leq -1$.

Proof: Assume first that $x \in [1, 2^l]$. Raising each constituent of inequality (a) in Lemma 1 to the *odd* power $\log t$ and simplifying yields:

$$0 \leq (p_l(x))^{\log t} \leq \frac{(-p_l(-x))^{\log t}}{t^2}.$$

An argument analogous to that in Lemma 1 yields the following lower and upper bounds: $1 \leq s_l^{\log t} \leq 1 + 2/(t^2-1)$. Noting that $2/(t^2-1) < 1/t$ for $t \geq 3$, we obtain: $1 \leq s_l^{\log t} \leq 1 + 1/t$.

The argument for $x \in [-2^l, -1]$ is analogous. ■

6.5 Application to Learning Intersections

We now revisit the problem of learning the intersection of two halfspaces $f(x) = \text{sign}(\sum_{i=1}^n \alpha_i x_i - \theta_1)$ and $g(x) = \text{sign}(\sum_{i=1}^n \beta_i x_i - \theta_2)$. Consider the rational functions $R_f(x) = s_{\log W}(\sum_{i=1}^n \alpha_i x_i - \theta_1)$ and $R_g(x) = s_{\log W}(\sum_{i=1}^n \beta_i x_i - \theta_2)$. Denote $R(x) = R_f(x) + R_g(x)$. We will show that $f(x) \wedge g(x) \equiv R(x) \geq 2$:

- $\Rightarrow$: Suppose $f(x) \wedge g(x)$. Then $\sum_{i=1}^n \alpha_i x_i - \theta_1 \geq 1$ and $\sum_{i=1}^n \beta_i x_i - \theta_2 \geq 1$. By Lemma 1, the sum $R(x)$ is at least 2.
- $\Leftarrow$: Suppose $\neg f(x) \vee \neg g(x)$. Then $\sum_{i=1}^n \alpha_i x_i - \theta_1 \leq -1$ or $\sum_{i=1}^n \beta_i x_i - \theta_2 \leq -1$. By Lemma 1, one of the summands can therefore be at most -1. Even if the other summand attains the maximum value of 5/3, the resulting sum will still be $5/3 - 1 < 2$.

To obtain a PTF from $R(x) \geq 2$, it suffices to multiply through by the squared product of the denominators. Namely, let $R_f(x) = p_f(x)/q_f(x)$ and $R_g(x) = p_g(x)/q_g(x)$, where $p_f(x), p_g(x), q_f(x), q_g(x)$

are all polynomials of degree $O(\log W)$, by construction of $s_{\log W}$. Then the corresponding PTF is $R(x) \cdot q_f^2(x)q_g^2(x) \geq 2q_f^2(x)q_g^2(x)$, or

$$p_f(x)q_f(x)q_g^2(x) + p_g(x)q_f^2(x)q_g(x) \geq 2q_f^2(x)q_g^2(x).$$

This inequality computes $f(x) \wedge g(x)$ because (i) $R(x) \geq 2$ computes $f(x) \wedge g(x)$, and (ii) multiplying through by the positive quantity $q_f^2(x)q_g^2(x)$ yields an equivalent inequality. The degree of the resulting PTF is $4 \cdot O(\log W) = O(\log W)$.

6.6 Learning Intersections of t Halfspaces

The result of the previous section can be generalized to t halfspaces by using a more accurate approximation of the sign function. Namely, consider t PTF's $h_1, h_2, \cdots, h_t$. Let $\hat{h}_1, \hat{h}_2, \cdots, \hat{h}_t$ be the corresponding linear functions s.t. $h_i = \text{sign}(\hat{h}_i)$ for $1 \leq i \leq t$. Form $R(x) = \sum_{i=1}^t s_{\log W}^{\log t}(\hat{h}_i)$. We will show that $\bigwedge_{i=1}^t h_i \equiv R(x) \geq t$:

- $\Rightarrow$: Suppose $\bigwedge_{i=1}^t h_i$. Then each $\hat{h}_i \geq 1$ and thus $1 \leq s_{\log W}^{\log t}(\hat{h}_i) \leq 1 + 1/t$. This implies that $R(x) \geq t$.
- $\Leftarrow$: Suppose $\bigvee_{i=1}^t \neg h_i$. Then $s_{\log W}^{\log t}(\hat{h}_i) \leq -1$ for at least one $\hat{h}_i$. Even if the remaining summands contribute the maximum value of $1 + 1/t$ each, the sum will be $R(x) \leq -1 + (1 + 1/t)(t-1) = -1 + t - 1 + (t-1)/t < t$.

To convert $R(x)$ into a PTF, we clear the t denominators by multiplying through by the square of their product. Since the numerator and denominator of each $s_{\log W}^{\log t}(\hat{h}_i)$ is a polynomial of degree $O(\log t \log W)$, clearing the denominators will yield a PTF of degree $2t \cdot O(\log t \log W) = O(t \log t \log W)$.

6.7 Learning Arbitrary Boolean Functions of t Halfspaces

The above method for learning *intersections* of t halfspaces generalizes to arbitrary Boolean functions on t halfspaces. Namely, let h be a Boolean function of halfspaces $h_1, h_2, \cdots, h_t$. Using the truth table for h, we can write h as a degree-t polynomial d in $h_1, h_2, \ldots, h_t$ via the standard conversion procedure. Specifically, if $\langle h_1 = 1, h_2 = -1, h_3 = 1\rangle$ is a satisfying assignment, we form the *elementary product* $h_1 \cdot \frac{h_2+1}{2} \cdot h_3$; the polynomial representing h is the sum of all such elementary products. By construction, each of the elementary products evaluates to 1 on its corresponding assignment, and to 0 otherwise. As a result, d evaluates to 1 if h holds, and to 0 otherwise.

To obtain a representation of h in the original variable set, we construct the polynomial $d(h_1, h_2, \ldots, h_t)$ and make the replacement $h_i = s_{\log W}^{\log 2^{3t}}(\hat{h}_i(x))$, for all i. For simplicity, we adopt the following notation for the resulting polynomial:

$$d\left(s_{\log W}^{\log 2^{3t}}(\hat{h}_1(x)), \quad s_{\log W}^{\log 2^{3t}}(\hat{h}_2(x)), \quad \ldots, \quad s_{\log W}^{\log 2^{3t}}(\hat{h}_t(x))\right) = d(x).$$

It remains to show how to obtain from $d(x)$ a PTF computing h. To this end, we need to understand how the inaccuracies in the representation of the h_i affect the output of $d(x)$. In other words, we need to bound $|d(h_1, \dots, h_t) - d(x)|$. We start with a basic algebraic fact.

Lemma 3 *For all reals $k \geq 1$ and for all integers $t \geq 1$,*

$$1 \leq \left(1 + \frac{1}{k}\right)^t \leq 1 + \frac{2^t}{k} \qquad \text{and} \qquad 1 - \frac{2^t}{k} \leq \left(1 - \frac{1}{k}\right)^t \leq 1.$$

Proof: Consider first the quantity $(1 + 1/k)^t$. Trivially, $(1 + 1/k)^t \geq 1$. The upper bound can be shown as follows:

$$\left(1 + \frac{1}{k}\right)^t = \sum_{i=0}^{t} \binom{t}{i} \frac{1}{k^i} = 1 + \sum_{i=1}^{t} \binom{t}{i} \frac{1}{k^i} \leq 1 + \sum_{i=1}^{t} \binom{t}{i} \frac{1}{k} \leq 1 + \frac{1}{k} \sum_{i=1}^{t} \binom{t}{i} \leq 1 + \frac{2^t}{k}.$$

Similarly, $(1 - 1/k)^t \leq 1$. The lower bound can be shown as follows:

$$\left(1 - \frac{1}{k}\right)^t = \sum_{i=0}^{t} \binom{t}{i} \frac{(-1)^i}{k^i} = 1 + \sum_{i=1}^{t} \binom{t}{i} \frac{(-1)^i}{k^i} \geq 1 - \sum_{i=1}^{t} \binom{t}{i} \frac{1}{k^i} \geq 1 - \frac{1}{k} \sum_{i=1}^{t} \binom{t}{i} \geq 1 - \frac{2^t}{k}.$$

■

To bound the representation error in $d(x)$, consider the contribution of each elementary product to this error. Let $k = 2^{3t}$ denote the sign-function accuracy parameter in the above representation of $d(x)$. Let $P(x)$ be an arbitrary elementary product. If x is the satisfying assignment for $P(x)$, the output of $P(x)$ is between $(1 - 1/k)^t$ and $(1 + 1/k)^t$. In this case, the contribution to the representation error is at most $\max\{1 - (1 - 1/k)^t, (1 + 1/k)^t - 1\} \leq 2^t/k$ (by Lemma 3). If x is not the satisfying assignment for $P(x)$, the output of $P(x)$ is between $(1/k)(1 + 1/k)^{t-1}$ and $-(1/k)(1 + 1/k)^{t-1}$. In this case, the contribution to the representation error is at most

$$\frac{1}{k}\left(1 + \frac{1}{k}\right)^{t-1} \leq \frac{1}{k}\left(1 + \frac{1}{k}\right)^{t-1} + \left(1 + \frac{1}{k}\right)^{t-1} - 1 = \left(1 + \frac{1}{k}\right)^t - 1 \leq \frac{2^t}{k} \quad \text{(by Lemma 3)}.$$

In any event, *the representation error due to a single elementary product is at most* $2^t/k = 2^{-2t}$. Since there are at most 2^t elementary products, the total representation error cannot exceed 2^{-t}. It immediately follows that $d(x) - 1/2 \geq 0$ is a valid inequality for the problem (assuming $t \geq 2$).

The degree of the PTF resulting from $d(x)$ is easy to compute. Every rational function $s_{\log W}^{\log 2^{3t}}(\hat{h}_i(x))$ in the definition of $d(x)$ has a numerator and denominator of degree $O(t \log W)$. Clearing the denominators in $d(x) - 1/2 \geq 0$ turns it into a PTF of degree $O(t^2 \log W)$.