CS 395T Computational Complexity of Machine Learning

Lecture 14: October 10, 2005 — Lecturer: Adam Klivans

Scribe: Jason Davis

14.1 Learning the Intersection of Two Origin Centered Halfspaces

We wish to learn the intersection of two halfspaces in $\Re^n$, which we will express as $h_1 \wedge h_2$, where $h_1 = sign(\sum_i \alpha_i x_i)$ and $h_2 = sign(\sum_i \beta_i x_i)$. We will first note the exclusive-or of these two halfspaces, $h_1 \oplus h_2$ can be expressed by embedding our n variables into an n^2 dimensional space:

$$h_1 \oplus h_2 = \sum_i \sum_j \alpha_i \beta_j x_{ij},$$

where we define each new variable $x_{ij} = x_i \cdot x_j$.

Using this trick, we construct the following algorithm:

1. Find a halfspace h_a (using linear programming or any such polynomial algorithm) with all positive examples on one side.
2. Remove all points on the negative side of h_a, and use the exclusive or embedding to find a halfspace h_b.
3. Output final hypothesis $h_a \wedge h_b$.

This result is due to Eric Baum.

14.2 Boosting

Given a PAC learning algorithm with some (δ, ϵ) pair, we can easily 'boost' our confidence parameter, δ, by running the the PAC learning algorithm a polynomial number of times, and testing the final hypothesis between each run to see if it satisfies the given error ϵ. Boosting ϵ is non-trivial and was solved by Robert Schapire in 1991.

Definition 1 *A concept class C is weakly learnable if there exists a weka learning algorithm A such that for any distribution D on X, and for all $c \in C$, A outputs a hypothesis h such that $Pr_{x \in D}[h(x) \neq c(x)] \leq \frac{1}{2} - \eta$ and $\eta \geq \frac{1}{poly(n)}$. Furthermore, A should run in polynomial time and success w.h.p.*

Schapire's boosting algorithm implies that weak learnability and strong learnability are equivalent. We will next give a small example of how Schapire's original booster works.

Assume some concept class C to be weakly learnable with $\eta = .2$. Take a large set of examples, S, and run the weak learner on S to get a hypothesis h_1 with error at most $(1-\eta) = .3$. We define the initial distribution D_1 to be uniform, and we will re-weight this distribution to form D_2 such that half of the weight goes to points that h_1 predicted correctly, and the other half goes to points predicted incorrectly:

$$D_2(x) = \begin{cases} \frac{5}{7}D_1(x) & \text{if } x \text{ predicted correctly by } h_1 \\ \frac{5}{3}D_1(x) & \text{if } x \text{ predicted incorrectly by } h_1 \end{cases} \tag{14.1}$$

We will then run the weak learner again using distribution D_2 to obtain hypothesis h_2. Finally, for all points x such that $h_1(x) \neq h_2(x)$, we will run the weak learner a third time over these points (with the uniform distribution) to obtain the hypothesis h_3. Our final hypothesis is $h = majority(h_1, h_2, h_3)$.

We will partition S into four regions:

- R_1: h_1, h_2 correct
- R_2: h_1 correct, h_2 incorrect
- R_3: h_1, h_2 incorrect
- R_1: h_1 incorrect, h_2 correct

and we will define the weight that D_2 puts on R_2, $D_2(R_2)$ to be γ. From this, we have

$$\begin{aligned} D_2(R_3) &= .3 - \gamma \\ D_2(R_1) &= .5 + \gamma \\ D_2(R_4) &= .2 + \gamma. \end{aligned}$$

In particular, from our definition of D_2,

$$\begin{aligned} D_1(R_1) &= \frac{7}{5}(.5 - \gamma) \\ D_1(R_3) &= \frac{3}{5}(.3 - \gamma) \end{aligned}$$

Finally, we will bound the error of h in terms of our original distribution (as this is the distribution from which S is drawn). Consider some point x. If $x \in R_1$, then at least 2 of the 3 hypothesis will correctly classify it. If $x \in R_2$ or $x \in R_4$, then exactly one of h_1 or h_2 will correctly classify x, and with probability .7, hypothesis h_3 will classify x correctly. Finally, if $x \in R_4$, then x will always be classified incorrectly. Our new error is:

$$\begin{aligned} error(h) &= D_1(R_3) + .3(D_1(R_2) + D_1(R_4)) \\ &= D_1(R_3) + .3(1 - D_1(R_1) - D_1(R_3)) \\ &= .7D_1(R_3) + .3(1 - D_1(R_1)) \\ &= .7\left(\frac{3}{5}(.3 - \gamma)\right) + .3\left(.3 + \frac{7}{5}\gamma\right) \\ &= (.3)^2\left(1 + \frac{7}{5}\right) \\ &\approx .216 \end{aligned}$$

In general, for arbitrary η, the error of h is $3\beta^2 - 2\beta^3$, where $\beta = \frac{1}{2} - \eta$.