CS 2416 - Machine Learning Theory

Lecture #9: 10 November, 2005

Lecturer: Toniann Pitassi

Scribe Notes by: Toniann Pitassi

1 Generalization error of Adaboost

Recall that in the last lecture we described the Adaboost algorithm. The algorithm is roughly as follows. First, we pick a polynomial sized sample of inputs from the original distribution D_1, say, $S = x_1, \ldots, x_m$. The algorithm then does the following:

Run weak learning algorithm on $x_1, \ldots, x_m$ distributed according to D_1, to get h_1

Run weak learning algorithm on $x_1, \ldots, x_m$ distributed according to D_2, to get h_2

...

Run weak learning algorithm on $x_1, \ldots, x_m$ distributed according to D_T, to get h_T

The algorithm then returns a weighted majority of $h_1, h_2, \ldots, h_T$. More specifically, if $\epsilon_t = Pr_{D_t}[h_t(x_i) \neq y_i]$, then the algorithm returns 1 if and only if

$$\sum_{i=1}^{T} \alpha_t h_t(x) > \frac{1}{2},$$

where $\alpha_t = \frac{\ln(\frac{1-\epsilon_t}{\epsilon_t})}{\sum_{i=1}^{T} \ln(\frac{1-\epsilon_i}{\epsilon_i})}$

We proved the following theorem, which shows that the empirical error is very small. (The empirical error is the error with respect to the samples, S.)

Theorem 1 $\widehat{error}(H) \leq \prod_{t=1}^{T} 2\sqrt{\epsilon_t(1-\epsilon_t)}$

Now let G be the class of functions of the form $\sum_{t=1}^{T} \alpha_t h_t(x)$. We want to determine the growth function of G, $\Pi_G(m)$.

Suppose that $h_1, \ldots, h_T \in \mathcal{H}$ where $|\mathcal{H}|$ is finite. Given $S = x_1, \ldots, x_m$, how many dichotomies can we get of these m points? This calculation is similar to the one that we did for neural networks a couple of lectures ago. Since the VC dimension of a linear combination of T variables is T, the growth function with $h_1, \ldots, h_T$ fixed is at most

$$(em/T)^T.$$

Since there are $|\mathcal{H}|^T$ choices for $h_1, \ldots, h_T$, the growth function is bounded as follows:

$$\Pi_G(m) \leq |\mathcal{H}|^T (em/T)^T.$$

Applying Theorem 2 of Lecture we can show that with probability at least $(1-\delta)$,

$$err(H) \leq \widehat{err}(H) + O(\sqrt{\frac{ln|\Pi_G(2m)| + ln(1/\delta)}{m}})$$

Now plugging in for Π_G we obtain

$$err(H) \leq \widehat{err}(H) + O(\sqrt{\frac{Tln|\mathcal{H}| + Tln(m/T) + ln(1/\delta)}{m}}).$$

Thus as long as m is chosen appropriately, the actual error $(err(H))$ will be about the same as the empirical error $(\widehat{err}(H))$.

A similar calculation can be done using VC dimension rather than $|\mathcal{H}|$ whenever $\mathcal{H}$ is infinite. Namely, Freund and Schapire prove the following theorem.

Theorem 2 *Let $\mathcal{H}$ be a class of binary functions of VC-dimension $d \geq 2$. Then the VC-dimension of the class of linear threshold of T functions in $\mathcal{H}$ is at most $2(d+1)(T+1)log_2(e(T+1))$ (where e is the base of the natural logarithm.) Therefore, if the hypotheses generated by the weak learner are chosen from a class of VC-dimension $d \geq 2$, then the final hypotheses generated by AdaBoost after T iterations belongs to a class of VC-dimension at most $2(d+1)(T+1)log_2[e(T+1)]$.*

The above theorem, combined with the following theorem of Vapnik, gives good bounds of generalization error as a function of VC-dimension.

Theorem 3 *Let $\mathcal{H}$ be a class of binary functions over domain X, and let d be the VC-dimension of $\mathcal{H}$. Let D be a distribution over labelled examples (over pairs $Xx\{+1,0\}$). Let S be a sample of m examples drawn from D. Then for any $\delta > 0$, we have*

$$Pr_S[\exists h \in \mathcal{H} \mid |Err(h) - \widehat{err}(h)| > 2\sqrt{\frac{d(ln(2n/d)+1) + ln(9/\delta)}{m}}] \leq \delta.$$

2 Overfitting

Note that whenever m is chosen appropriately, then Adaboost acheives low error. Thus, from the above calculations we get the equivalence between weak and strong learning.

However, as T increases (think now of m being fixed), the above equation predicts that there will eventually be overfitting. That is, Adaboost runs for longer and longer, the empirical error will continue to go down to zero at a very fast rate, but the true error will start to increase. This is consistent with Occam's razor, which says that the simplest hypothesis explaining the data is the best one. That is, as the hypthesis becomes more and more complex (as we take the weighted majority of lots of concepts), we will overfit and generalization will decrease.

Surprisingly, however, this does not happen with Adaboost in practice! In many examples it is observed that the generalization error continues to drop with the training error, and moreover, even when the training error is zero, if Adaboost continues to run, the generalization error continues to drop!

A recent "margins" analysis explains this superior performance of Adaboost. Define the margin as:

$$margin \ = \ yf(x)$$

where $y \in \{1, -1\}$. Remember that $y \in \{1, -1\}$ and $f(x)$ is a weighted majority of the predictions by several hypotheses. Thus, $f(x)$ is a number in $[1, -1]$. When $y = 1$ and $f(x) \geq 0$, $f(x)$ will give the correct value, and the closer $f(x)$ is to 1, the "more correct" it will be, and the larger the margin. Similarly when $y = -1$, as long as $f(x) < 0$, it will give the correct value, but the closer $f(x)$ is to -1, the larger the margin. Thus the margin is associated with confidence–the higher the margin the greater the strength or confidence of the weighted votes.

Recall that what we have proven so far about adaboost is that the generalization error can be bounded by the empirical error, plus a term that depends on T, the number of iterations, as well as on the number of samples, m, and some measure of the complexity of $\mathcal{H}$. Thus, as T increases (well beyond when the empirical error is zero), the second term will get larger, and thus it may seem that the generalization error will in fact start to decrease because of overfitting.

However, this does not happen! We will now present an improved bound on the generalization error. Roughly, we want to show that the generalization error can be upper bounded by the sum of two terms, where the first term bounds the number of examples with margin below a certain threshold θ, plus a second term that depends on the number of training examples, m, some measure of complexity of $\mathcal{H}$, and the threshold θ. The important thing to notice is that the second term no longer depends on T, and thus we see that in fact adaboost continues to improve in terms of generalization as T increases, since the first term will decrease with increasing T!

We will first show, using the same analysis as in the previous lecture, that not only does the empirical error decrease as T increases, but the margin also increases as T increases. (This will be the first term described above.)

That is, we can prove that:

$$Pr_S[yf(x) \leq \theta] \leq \Pi_{t=1}^{T} \sqrt{2\epsilon_t^{1-\theta}(1-\epsilon_t)^{1+\theta}}.$$

Let's see this when all ϵ_i's have the same value, $\epsilon_1 = \ldots = \epsilon_T = \epsilon$.

Claim 1. The total weight goes down. This argument is the same, so we have $w_T = (2\epsilon)^T m$.

Claim 2. Any bad example (where margin is smaller than θ) has to have high weight.

Recall that the margin, θ, on a particular example x equals $(C - I)/T$, where C is the number of hypothesis that label x correctly, and I is the number of hypotheses that label x incorrectly. Thus,

$$\theta \ = \ (C - I)/T \ = \ (C - (T - C))/T \ = \ (2C - T)/T.$$

Solving for C gives $C = T(1 + \theta)/2$. Thus if x is bad (has margin less than θ), then at most C hypotheses get x correct. Every time x is correctly labelled, the weight goes down by a factor of $(\epsilon/(1 - \epsilon)$ and thus the final weight of a bad x is at least

$$((\epsilon/(1-\epsilon))^{T(1+\theta)/2}.$$

If E is the fraction of bad x's, then we have

$$w_T \geq ((\epsilon)/(1-\epsilon))^{T(1+\theta)/2} Em.$$

Putting this together, we have

$$(2\epsilon)^T m \geq ((\epsilon/(1-\epsilon))^{T(1+\theta)/2} Em.$$

Solving for E gives

$$E \leq (\sqrt{2\epsilon^{1-\theta}(1-\epsilon)^{1+\theta}})^T.$$

3 Margin analysis for Adaboost

We want to prove the following theorem.

Theorem 4 *With probability* $(1-\delta)$, *for all* $f \in co(\mathcal{H})$ *and for all* $\theta > 0$,

$$Pr_D[yf(x) \leq 0] \leq Pr_S[yf(x) \leq \theta] + \left(\frac{1}{\sqrt{m}} \sqrt{\frac{ln(m)ln(|\mathcal{H}|) + ln(1/\delta)}{\theta^2}} \right).$$

As mentioned earlier, the $O()$ term is independent of the number of rounds of boosting! Since the first term can be bounded, overall we have:

$$Pr_D[yf(x) \leq 0] \leq \Pi_{t=1}^T \sqrt{2\epsilon_t^{1-\theta}(1-\epsilon_t)^{1+\theta}} \;+\; \left(\frac{1}{\sqrt{m}} \sqrt{\frac{ln(m)ln(|\mathcal{H}|) + ln(1/\delta)}{\theta^2}} \right).$$

The main intuition, as explained in the Boosting the Margin paper, is as follows. The idea brings us back to Occam's razor though in an indirect way. Recall that an example is classified correctly if its margin is positive. If an example is classified by a large margin (either positive or negative) then small changes to the weights in the majority vote are unlikely to change the label. If most of the examples have a large margin, then the classification error of the original majority vote and the perturbed majority vote will be similar. Suppose now that we have a small set of weighted majority rules that was fixed ahead of time, called the "approximation set." One way of perturbing the weights of the classifier majority vote is to find a nearby rule within the approximating set. As the approximating set is small, we can guarantee that the error of the approximating rule on the training set is similiar to its generalization error, and as its error is similar to that of the original rule, the generalization error of the original rule should also be small. Thus we are back to an Occam's razor argument in which instead of arguing that the classification rule itself is simple, we argue that the rule is close to a simple rule.

Let $co(\mathcal{H})$ be the set of all functions f of the form $\sum_{i=1}^T a_t h_t(x)$, where $a_t \geq 0$ and $\sum a_t = 1$, and $h_t \in \mathcal{H}$. Let D be the distribution over labelled examples, and let S be a sample of size m.

As mentioned above, we want to approximate $co(\mathcal{H})$ with a much smaller class. Let C_N be the set of unweighted averages over N elements from $\mathcal{H}$. That is, C_N is the set of all functions g of the form $\frac{1}{N}\sum_{j=1}^N h_j(x)$.

Given $f(x) = \sum a_t h_t(x)$, we will construct a $g \in C_N$ that approximates f. We will construct g randomly, and write $Pr_g[]$ to denote the probability over a random choice of g. Notice that $|C_N| \leq |\mathcal{H}|^N$ because we have $|\mathcal{H}|$ choices for h_j and there are N of them. Think of N as much smaller than T.

We construct each g by randomly choosing N elements, $g_1, \ldots, g_N$, from the set of h_t's in f, where we choose each g_i independently, picking it to be h_t with probability α_t.

The main idea of the argument is to show that:

(1) There exists a $g \in C_N$ such that for most x, $|f(x) - g(x)| \leq \theta/2$.

(2) With probability at least $(1 - \delta)$,

$$Pr_D[yg(x) \leq \theta/2] \leq Pr_S[yg(x) \leq \theta/2] + \epsilon_N.$$

Then this will allow us to show:

$$\begin{aligned} Pr_D[yf(x) \leq 0] &= Pr_D[yg(x) \leq \theta/2] \\ &= Pr_S[yg(x) \leq \theta/2] \\ &= Pr_S[yf(x) \leq \theta] \end{aligned}$$

(The above equalities should be approximate equalities but I don't know how to typeset them.)

$$Pr_D[yf(x) \leq 0] \;\leq\; Pr_{D,g}[yf(x) \leq 0 \;\wedge\; yg(x) > \theta/2] \;+\; Pr_{D,g}[yg(x) \leq \theta/2] \tag{1}$$

The first term can be rewritten as:

$$Pr_{D,g}[yf(x) \leq 0 \wedge yg(x) > \theta/2] = E_D[Pr_g[yf(x) \leq 0 \wedge yg(x) > \theta/2])$$

Now this is at most $E_D[Pr_g[yg(x) > \theta/2 \mid yf(x) \leq 0]]$. Here's a way to think about what is going on. Since $f(x)$ and y are fixed, consider the case where $f(x)$ gives equal weight to each hypothesis. Then $f(x)$ consists of the sum of T numbers, where each number is either -1 or +1, and where there are at least $T/2$ -1's (since $f(x)$ is at most 0.) A random g is obtained by choosing N out of the T numbers, and we want to calculate the probability that we end up with a lot more 1's – more precisely with an excess of at least $\delta/2$ 1's. This can be viewed as a balls-in-bin experiment. There are $T/2$ red balls and $T/2$ black balls. We choose N balls with replacement from the bin, and we want to calculate the probability that we pick at least $\delta/2$ extra black balls than red balls. The expected number of excess black balls is 0. By Chernoff bounds, we get

$$Pr_g[yg(x) > \delta/2 \mid yf(x) \leq 0] \leq e^{-N\theta^2/8}.$$

Thus the first term is bounded by

$$E_D[e^{-N\theta^2/8}] = e^{-N\theta^2/8}.$$

For the second term, we want to show that with probability at least $(1 - \delta_N)$,

$$Pr_{D,g}[yg(x) \leq \theta/2] \leq Pr_{S,g}[yg(x) \leq \theta/2] + \epsilon_N.$$

Fix g and θ. Applying Chernoff bounds we obtain: $\Pr_S(Pr_D[yg(x) \leq \theta/2] > Pr_S[yg(x) \leq \theta/2] + \epsilon_N]) \leq e^{-2\epsilon_N^2 m}$,

Intuitively, $yg(x) \leq \theta/2$ is the proportion of $\mathcal{D}$ where the margin is small – think of this proportion as the "true" proportion. Now we choose a sample S and compute its proportion of S where the margin is small – this is the "empirical" proportion. A sample S is "bad" if its empirical proportion plus an extra proportion, ϵ_N, is still smaller than the true proportion.

Now we want to calculate the overall bad probability as we vary over all choices of g and θ. There are $|\mathcal{H}|^N$ choices of g. Since $yg(x)$ is always a multiple of $1/N$ and is between 0 and 1, the different possible values of θ are $2i/N$, for $i = 0, \ldots, N$. Thus there are $N+1$ choices of θ. By the union bound, we thus have that the probability over S that there exists any g and θ such that

$$Pr_D[yg(x) \leq \theta/2] > Pr_S[yg(x) \leq \theta/2] + \epsilon_N$$

is at most $(N+1)|\mathcal{H}|^N e^{-2m\epsilon_N^2}$.

As long as we choose ϵ_N so that

$$(N+1)|\mathcal{H}|^N e^{-2\epsilon_N^2 m} = \delta,$$

we have with probability at least $1-\delta$, for all $g \in C_N$ and for all $\theta > 0$,

$$Pr_D[yg(x) \leq \theta/2] \leq Pr_S[yg(x) \leq \theta/2] + \epsilon_N.$$

Solving for ϵ_N gives

$$\epsilon_N = \sqrt{\frac{1}{2m} ln(\frac{(N+1)|\mathcal{H}|^N}{\delta})}.$$

Plugging in these quantities for term 1 and term 2, we now have that with probability at least $(1-\delta_n)$,

$$Pr_D[yf(x) \leq 0] \leq e^{-N\theta^2/8} + Pr_{S,g}[yg(x) \leq \theta/2] + \epsilon_n.$$

To finish the argument, we want to relate the fraction of the training set on which $yg(x) \leq \theta/2$ to the fraction on which $yf(x) \leq \theta$, which is the quantity that we measure.

$$\begin{aligned} Pr_S[yg(x) \leq \theta/2]] &\leq Pr_{S,g}[yf(x) \leq \theta] + Pr_{S,g}[yg(x) \leq \theta/2 \wedge yf(x) > \theta] \\ &= Pr_S[yf(x) \leq \theta] + E_S[Pr_g[yg(x) \leq \theta/2 \wedge \wedge yf(x) > \theta]] \\ &\leq Pr_S[yf(x) \leq \theta] + E_S[Pr_S[yg(x) \leq \theta/2 \mid yf(x) > \theta]] \end{aligned}$$

To bound the expression inside the expectation we use the Chernoff bound as we did to bound term 1 to get

$$Pr_S[yg(x) \leq \theta/2 \mid yf(x) > \theta] \leq e^{-N\theta^2/8}.$$

Thus we have overall

$$Pr_D[yg(x) \leq 0] \leq 2e^{-N\theta^2/8} + Pr_S[yf(x) \leq \theta].$$

Let $\theta_n = \theta/(N(N+1))$ so that the probability of failure for any N will be at most $\sum_N \theta_N = \theta$. Then we get:

$$N = \frac{4}{\theta^2} ln(\frac{m}{ln|\mathcal{H}|}).$$

3.1 Margins analysis for infinite classs

The above argument can be modified to give a similar result based on VC dimension. In particular, in the Boosting the Margins paper, the authors prove the following theorem.

Theorem 5 *Let D be a distribution over labelled examples, with input space X. (Outputs are either $+1$ or -1.) Let S be a set of m examples from D chosen independently and at random. Suppose the hypothesis space $\mathcal{H}$ has VC-dimension d, and let $\delta > 0$. Assume that $m \geq d \geq 1$. Then with probability at least $1 - \delta$ (over the random choice of S), every weighted average function $f \in C$ satisfies the following bound for all $\theta > 0$:*

$$Pr_D[yf(x) \leq 0] \leq P_S[yf(x) \leq \theta] \ + \ O(\frac{1}{\sqrt{m}}(\frac{d log^2(m/d)}{\theta^2} \ + \ \log(1/\delta))^{1/2}).$$