COS 511: Foundations of Machine Learning

Rob Schapire
Scribe: Wei Dong

Lecture #11
March 14, 2006

1 Generalized Error of AdaBoost Based on Margin

The analysis of AdaBoost mainly includes two parts: to analyze the margin of training examples and to analyze if achieving certain margin is enough to guarantee the test error given enough training data. In the previous lecture, we outlined the proof of the upper bound of generalized error of AdaBoost as a function of margin. Today we are going to finish the proof. First, lets summarize the notations which will make the upcoming discussion easier.

$$
\begin{array}{lcl}
\mathcal{H} & \stackrel{\text{def}}{=} & \text{Weak hypothesis space} \\
co(\mathcal{H}) & \stackrel{\text{def}}{=} & \{f(x) = \sum_j a_j h_j(x) : a_j \geq 0, \sum_j a_j = 1, h_j \in \mathcal{H}\} \\
\mathcal{C}_N & \stackrel{\text{def}}{=} & \{g(x) = \frac{1}{N}\sum_{j=1}^{N} h_j(x) : h_j \in \mathcal{H}\} \\
\mathcal{D} & \stackrel{\text{def}}{=} & \text{Distribution on } X \times \{-1, +1\} \\
S & \stackrel{\text{def}}{=} & \text{Sample set} \\
Pr_{\mathcal{D}}[\cdot] & \stackrel{\text{def}}{=} & \text{probability over } \langle x, y\rangle \sim \mathcal{D} \\
Pr_{S}[\cdot] & \stackrel{\text{def}}{=} & \text{probability over } \langle x, y\rangle \text{ sampled in } S \text{ uniformly at random}
\end{array}
$$

And following is the theorem we are going to prove.

Theorem 1 *With probability* $\geq 1 - \delta$, $\forall f \in co(\mathcal{H})$, $\forall \theta > 0$,

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

This theorem says that the generalization error of AdaBoost can be bounded in terms of the number of training examples with margin below a threshold θ, plus an additional term which depends on the number of training examples, the size of $\mathcal{H}$, and the threshold θ(preventing us from choosing θ too close to zero).

Here we are interested in the function $f(x) = \sum_t a_t h_t(x)$, which can be approximated by sampling in the hypothesis space $\mathcal{H}$ with the probability given by a_t. Let

$$
\begin{array}{lcl}
g_j & = & \text{a hypothesis sampled from } \mathcal{H},\ h_t \text{ chosen with probability } a_t \\
g(x) & = & \frac{1}{N}\sum_{j=1}^{N} g_j(x) \in \mathcal{C}_N
\end{array}
$$

Then $g(x)$ can be regarded as a survey taken over the h_t's, which can be viewed as the voters. From the definition of g_j, we have

$$E_g[g_j(x)] = f(x)$$

And by Chernoff bound we know $g(x)$ better approximate $f(x)$. We will prove Theorem 1 by showing:

$$Pr_{\mathcal{D}}\left[yf(x) \le 0\right] \approx Pr_{\mathcal{D}}\left[yg(x) \le \frac{\theta}{2}\right]$$
$$Pr_{S}\left[yg(x) \le \frac{\theta}{2}\right] \approx Pr_{S}\left[yf(x) \le \theta\right]$$

and then

$$Pr_{\mathcal{D}}\left[yg(x) \le \frac{\theta}{2}\right] \approx Pr_{S}\left[yg(x) \le \frac{\theta}{2}\right]$$

Above is the high-level argument, we will go through the details of the proof in four steps.

Step 1 *For fixed x,*

$$Pr_{g}\left[|f(x) - g(x)| > \frac{\theta}{2}\right] \le \beta_\theta, \quad \textit{where } \beta_\theta = 2e^{-N\theta^2/8}$$

This step is just the formalization of the Chernoff bound argument.

Proof: Let $Z_j = g_j(x)$, then $g(x) = \frac{1}{N}\sum_j g_j(x) = \frac{1}{N}\sum_j Z_j$. By using Hoeffding's inequality, we can directly come to the result.

Step 2 *For random x with $(x, y) \sim P$,*

$$Pr_{P,g}\left[|yf(x) - yg(x)| > \frac{\theta}{2}\right] \le \beta_\theta.$$

This step is much the same as Step 1, except that $\langle x, y\rangle$ are chosen randomly according to some arbitrary distribution P.

Proof: By applying the marginalization trick $Pr_{x,y}[\Pi] = E_x[Pr_y[\Pi|x]]$, we have

$$\begin{aligned} & Pr_{P,g}\left[|yf(x) - yg(x)| > \frac{\theta}{2}\right] \\ = \;& E_P\left\{Pr_g\left[|yf(x) - yg(x)| > \frac{\theta}{2}|P\right]\right\} \\ \le \;& E_P\{\beta_\theta|P\} \\ = \;& \beta_\theta \end{aligned}$$

Step 3 *Fix g and $\theta > 0$, let*

$$P_{g,\theta} = Pr_{\mathcal{D}}[yg(x) \le \frac{\theta}{2}]$$
$$\hat{P}_{g,\theta} = Pr_{S}[yg(x) \le \frac{\theta}{2}],$$

then with respect to the choice of sample,

$$Pr_{sample}[P_{g,\theta} > \hat{P}_{g,\theta} + \epsilon] \le e^{-2\epsilon^2 m}.$$

Proof: We introduce a new variable Z_i for each example as follows

$$Z_i = \begin{cases} 1 & \text{if } y_i g(x_i) \leq \theta/2 \\ 0 & \text{else} \end{cases}$$

Then we have

$$E[Z_i] = P_{g,\theta}$$
$$\frac{1}{m}\sum_i Z_i = \hat{P}_{g,\theta}$$

Then we directly come to the conclusion by applying Hoeffding's Inequality.

Step 4

$$Pr_{sample}[\exists g \in \mathcal{C}_N, \theta > 0 : P_{g,\theta} > \hat{P}_{g,\theta} + \epsilon] \leq \delta$$

if

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

The technique for this step is union bound. Here we have $\mathcal{C}_N = |\mathcal{H}|^N$ and is thus finite. The problem lies in the real value θ. It turns out that even if there are infinity number of θ, there are only finite number of interesting ones. Actually we are interested in $yg(x) \leq \theta/2$. Plugging in the definition of $g(x)$, we have

$$\begin{aligned} & yg(x) \leq \theta/2 \\ \iff & \frac{y}{N}\sum_j g_j(x) \leq \frac{\theta}{2} \\ \iff & y\sum_j g_j(x) \leq \frac{N}{2}\theta. \end{aligned}$$

The left side of the inequity, $y\sum_j g_j(x)$, is an integer and the right side $\frac{N}{2}\theta$ is not necessarily an integer, thus the inequality will be true iff

$$y\sum_j g_j(x) \leq \left\lfloor \frac{N}{2}\theta \right\rfloor.$$

We do not need to consider all the θs, but only for the ones that make $\frac{N}{2}\theta$ an integer. Let

$$\tilde{\theta} = \frac{2}{N}\left\lfloor \frac{N}{2}\theta \right\rfloor.$$

Then $P_{g,\theta} = P_{g,\tilde{\theta}}$, and so

$$\begin{aligned}
& Pr_{\text{sample}}[\exists g, \theta, P_{g,\theta} > \hat{P}_{g,\theta} + \epsilon] \\
= \;& Pr_{\text{sample}}[\exists g, \theta, P_{g,\tilde{\theta}} > \hat{P}_{g,\tilde{\theta}} + \epsilon] \\
\leq \;& |\mathcal{H}|^N \left(\frac{N}{2} + 1\right) Pr_{\text{sample}}[P_{g,\tilde{\theta}} > \hat{P}_{g,\tilde{\theta}} + \epsilon] && \text{(Union bound)} \\
& && (\tilde{\theta} \text{ of the form } \frac{2}{N}v, v = 0, \ldots, \frac{N}{2}) \\
\leq \;& |\mathcal{H}|^N \left(\frac{N}{2} + 1\right) e^{-2\epsilon^2 m} && \text{(By Step 3)} \\
= \;& \delta
\end{aligned}$$

To put all the four steps together, we have with probability at least $1 - \delta$,

$$\begin{aligned}
& Pr_{\mathcal{D}}[yf(x) \leq 0] \\
= \;& Pr_{\mathcal{D},g}[yf(x) \leq 0] \\
= \;& Pr_{\mathcal{D},g}[yf(x) \leq 0 \wedge yg(x) \leq \theta/2] + Pr_{\mathcal{D},g}[yf(x) \leq 0 \wedge yg(x) > \theta/2] \\
\leq \;& Pr_{\mathcal{D},g}[yg(x) \leq \theta/2] + Pr_{\mathcal{D},g}[|yf(x) - yg(x)| > \theta/2] \\
\leq \;& E_g[Pr_{\mathcal{D}}[yg(x) \leq \theta/2|g]] + \beta_\theta \\
\leq \;& E_g[Pr_S[yg(x) \leq \theta/2|g] + \epsilon] + \beta_\theta \\
= \;& Pr_{S,g}[yg(x) \leq \theta/2] + \epsilon + \beta_\theta \\
= \;& Pr_{S,g}[yg(x) \leq \theta/2 \wedge yf(x) \leq \theta] + Pr_{S,g}[yg(x) \leq \theta/2 \wedge yf(x) > \theta] + \epsilon + \beta_\theta \\
\leq \;& Pr_{S,g}[yf(x) \leq \theta] + Pr_{S,g}[|yf(x) - yg(x)| > \theta/2] + \epsilon + \beta_\theta \\
\leq \;& Pr_S[yf(x) \leq \theta] + \beta_\theta + \epsilon + \beta_\theta
\end{aligned}$$

By plugging in

$$N = \left\lceil \frac{4}{\theta^2} \ln \frac{m}{\ln |\mathcal{H}|} \right\rceil$$

we get exactly the final result.

The bound proved above is not very meaningful unless m is very large. There are better bounds available. However, Theorem 1 does give a bound that predicts no overfitting.

2 Application of Boosting in text document classification

Boosting is used in text document classification, with one example being SPAM email detection. The weak learner often used is to test whether the document contains certain word or not.

Using such kind of weak learner means searching out a large space of hypotheses, because

$$|\mathcal{H}| = \text{number of words in the vocabulary}$$

This number can be even larger if we are to count in short phrases in addition to single words. Theorem 1 leads us to an easier life by making the bound only depend on the log of the size of the hypothesis space.

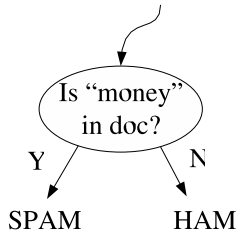

Figure 1: A Weak Learner Example

3 Introduction to Support Vector Machine

Boosting does not start out from maximizing the margin, but it seems that maximizing the margin is a good idea and this approach is taken by another learning algorithm, the Support Vector Machine, or SVM. Let's introduce the idea of SVM by looking back on the half-line learning problem (Figure 2). In this problem, we want to learn the separating point of the positive examples and the negative examples. Intuitively, we want the separating point to be half way between the closest positive and negative points. What if the instances are in a high dimension space? A straightforward idea is to generalize the separating point to a separating hyper-plane, which divides the whole instance space into two half-spaces. The basic idea of SVM is just to find such a separating hyper-plane, to maximize the margin between the hyper-plane and the data point. Here we assume that the data is linearly separable, which means we can always find such a separating hyper-plane. The model is formally defined as follows.

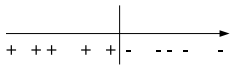

Figure 2: One-Dimensional Separating Problem

Given a set of pairs, $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)$, with $\mathbf{x}_i \in \mathbb{R}^n, ||\mathbf{x}||_2 \leq 1$ and $y \in \{-1, +1\}$, we want to find the hyper-plane $\mathbf{v} \cdot \mathbf{x} = 0, ||\mathbf{v}||_2 = 1$ (assume the hyper-plane goes through the origin). Later we simply represent the hyper-plane by its normal vector $\mathbf{v}$. By using a little geometry knowledge, for any point $\mathbf{x}$ we have

$$\mathbf{v} \cdot \mathbf{x} \begin{cases} > 0 & \text{if } \mathbf{x} \text{ is above the hyper-plane} \\ < 0 & \text{if } \mathbf{x} \text{ is below the hyper-plane} \\ = 0 & \text{if } \mathbf{x} \text{ is on the hyper-plane} \end{cases}$$

which gives a natural predicting hypothesis $h(\mathbf{x}) = sign(\mathbf{v} \cdot \mathbf{x})$.

The margin of SVM is related to that in boosting, but slightly different. Specifically, it is defined as follows:

$$\text{margin} \stackrel{\text{def}}{=} y(\mathbf{v} \cdot \mathbf{x})$$

The margin is > 0 if $(\mathbf{x}, y)$ is correctly classified. In SVM, we not only want the examples to be correctly classified, but also want its distance from the separating hyper-plane to be at least δ, i.e. $margin \geq \delta$. Thus we can express the problem of finding the separating

hyper-plane that maximize the margin as a mathematical programming:

$$\begin{aligned} \text{maximize} \quad & \delta \\ \text{s.t.} \quad & ||\mathbf{v}||_2 = 1 \\ & y_i(\mathbf{v} \cdot \mathbf{x}_i) \geq \delta, \quad \forall i \end{aligned}$$

We will see later that the result of this programming only depends on those examples with margin exactly equal to δ. The points $\mathbf{x}$ with $y(\mathbf{v} \cdot \mathbf{x}) = \delta$ are called the *support vectors*.

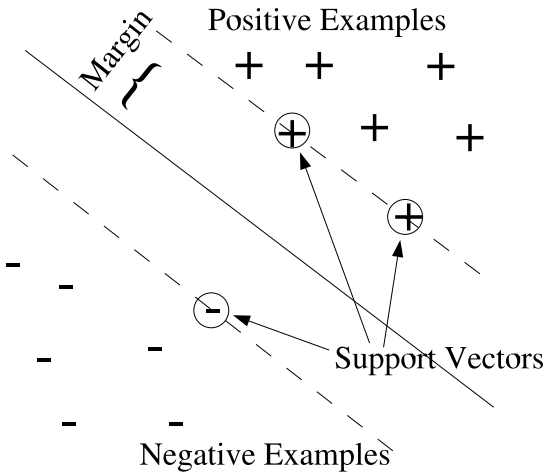

Figure 3: Support Vector Machine

4 Comparison of Support Vector Machine and Boosting

The following table summarizes the comparison of support vector machine and Boosting in terms of the space of instances/weak hypotheses(they are counterparts in the two algorithms), the parameters to optimize, the predicting method and the definition of margin. We can see from the table that the two algorithms have high similarity.

	SVM	**Boosting**
Instance/ Weak Hypo.	$\mathbf{x} \in \mathbb{R}^n$	$\mathbf{h}(x) =< h_1(x), h_2(x), \ldots >, \quad h_i \in \mathcal{H}$
	$\|\|\mathbf{x}\|\|_2 \leq 1$	$\|\|\mathbf{h}(x)\|\|_\infty = \max_i \|h_i(x)\| = 1, \quad h_i(x) \in \{-1, +1\}$
Search Goal	$\mathbf{v} \in \mathbb{R}^n$	$\mathbf{a} =< a_1, a_2, \ldots >$
	$\|\|\mathbf{v}\|\|_2 = 1$	$\|\|\mathbf{a}\|\|_1 = \sum_j \|a_j\| = 1$
Prediction	$sign(\mathbf{v} \cdot \mathbf{x})$	$sign(\mathbf{a} \cdot \mathbf{h}(x))$
Margin	$y(\mathbf{v} \cdot \mathbf{x})$	$y(\mathbf{a} \cdot \mathbf{h}(x))$

Table 1: Comparison of SVM and Boosting