Modern Topics in Learning Theory

Maria-Florina Balcan

04/24/2006

Modern Topics in Learning Theory

Semi-Supervised Learning

Active Learning

Kernels and Similarity Functions

Tighter Data Dependent Bounds

Outline

- AdaBoost
  - Algorithm
  - AdaBoost Behavior in Experiments

- Generalization error as a function of Margin Distributions
  - Classification Margin
  - Finite base-classifier spaces

- The effect of Boosting on Margin Distributions

AdaBoost recap

AdaBoost combines weak learners in a weighted majority voting scheme

- given a training set $(x_1, y_1), \cdots, (x_m, y_m)$
- $y_i \in \{-1, 1\}$ correct label of instance $x_i \in X$
- for $t = 1, \cdots, T$
  - construct a distribution D_t on $\{1, 2, \cdots, m\}$
  - find a weak hypothesis ("rule of thumb") $h_t : X \leftarrow \{-1, 1\}$ with small error ϵ_t on D_t, $\epsilon_t = \mathbf{Pr}_{D_t}[h_t(x_i) \neq y_i]$
- output final hypothesis H_{final}

- constructing D_t:
  - $D_1(i) = \frac{1}{m}$
  - given D_t and h_t

    $D_{t+1}(i) = \frac{D_t(i)}{Z_t} \cdot e^{-\alpha_t}$ if $y_i = h_t(x_i)$

    $D_{t+1}(i) = \frac{D_t(i)}{Z_t} \cdot e^{\alpha_t}$ if $y_i \neq h_t(x_i)$ where $\alpha_t = \frac{1}{2} \ln \left[\frac{1-\epsilon_t}{\epsilon_t}\right]$

- final hypothesis: $H_{final}(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$

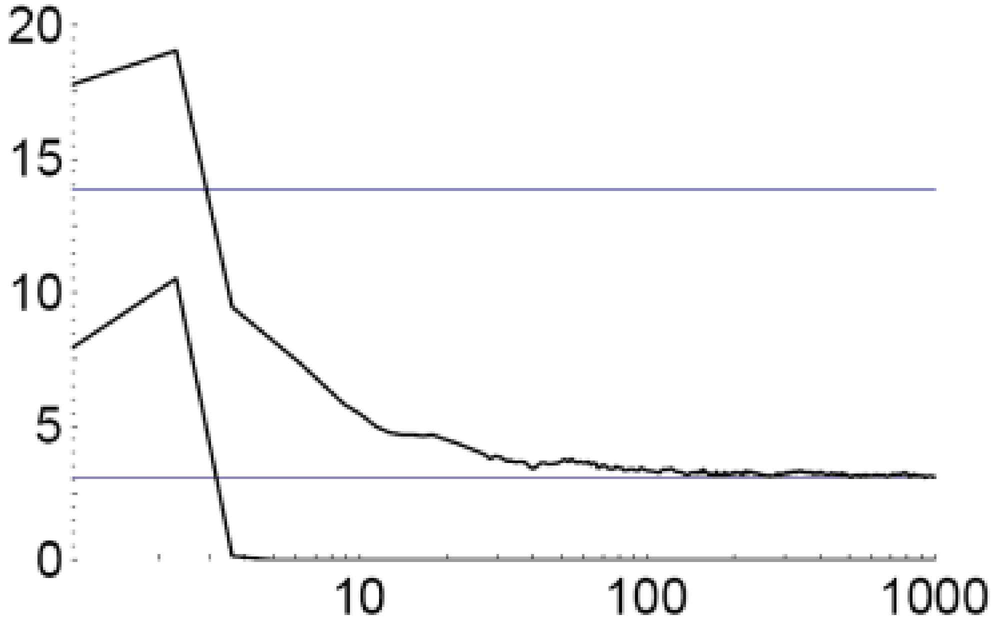
AdaBoost Behavior in Experiments

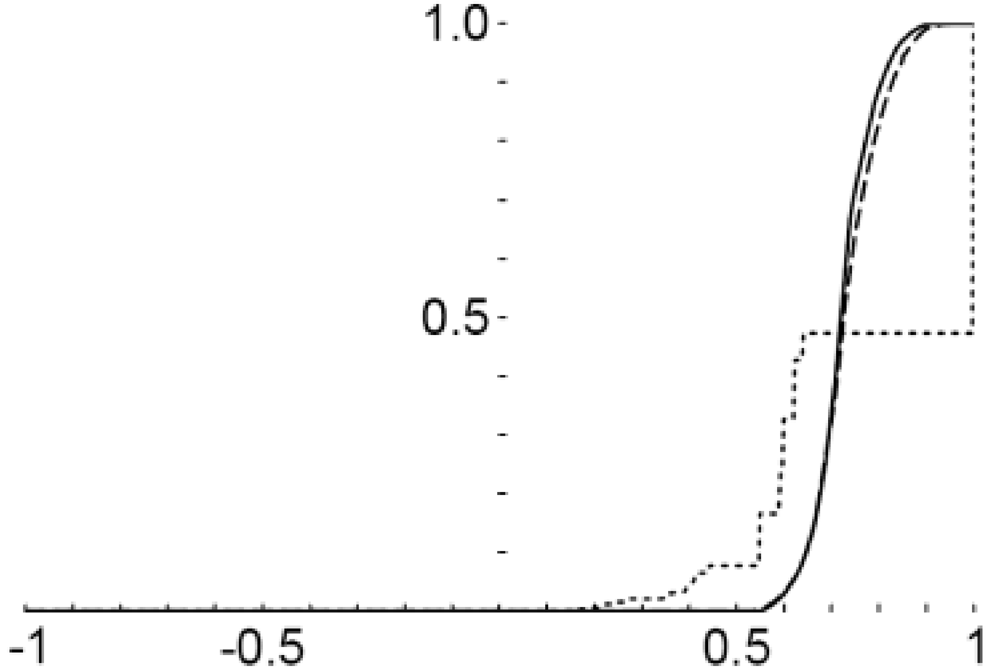
Experiments with boosting showed that the test error of the generated classifier usually does not increase as its size becomes very large.

Experiments with boosting showed also that continuing to add new weak learners after correct classification of the training set had been achieved could further improve test set performance!

These results seem to contradict Occam's razor: in order achieve good test error the classifier should be as simple as possible!

Error Curve, Margin Distr. Graph - Plots from [SFBL98]





Analyzing Generalization Error

Remember, usual sample complexity statements:

Theorem 1 *If H is a finite hypotheses space, then with probab. $1-\delta$, $\forall h \in H$ we have $|err(h) - \widehat{err}(h)| < \epsilon$ given that we see*

$$m \geq O\left(\frac{1}{\epsilon^2}\left[\ln|H| + \ln\frac{1}{\delta}\right]\right)$$

labeled examples.

Or, another way to state it: with probab. $1-\delta$, $\forall h \in H$

$$err(h) \leq \widehat{err}(h) + O\left(\sqrt{\frac{\ln|H| + \ln\left(\frac{1}{\delta}\right)}{m}}\right)$$

given that we see m labeled examples.

In general, with probab. $1-\delta$, $\forall h \in H$,

$$err(h) \leq \widehat{err}(h) + O\left(\sqrt{\frac{\ln(C[2m]) + \ln\left(\frac{1}{\delta}\right)}{m}}\right)$$

How can we explain the experiments?

R. Schapire, Y. Freund, P. Bartlett, W. S. Lee. present in "Boosting the margin: A new explanation for the effectiveness of voting methods" a nice theoretical explanation.

Main Idea:

Training error does not tell the whole story.

Need to also consider the classification confidence!

Classification Margin

Consider H to be the space of weak hypotheses. Define the convex hull of H to be

$$co(H) = \left\{ f = \sum_{t=1}^{T} a_t h_t, a_t \geq 0, \sum_{t=1}^{T} a_t = 1, h_t \in H \right\}$$

Let $f \in co(H)$, $f = \sum_{t=1}^{T} a_t h_t, a_t \geq 0, \sum_{t=1}^{T} a_t = 1$. The majority vote rule H_f associated with f (given by $H_f(x) = sign(f(x))$) gives a wrong prediction on the example (x, y) iff $yf(x) \leq 0$.

Define the margin of H_f(or of f) on example (x, y) to be $yf(x)$.

Note that $yf(x) = y \sum_{t=1}^{T} [a_t h_t(x)] = \sum_{t=1}^{T} [y a_t h_t(x)] = \sum_{t:y=h_t(x)} a_t - \sum_{t:y \neq h_t(x)} a_t$.

The margin is positive iff $y = H_f(x)$.

See $|yf(x)| = |f(x)|$ as the strength or the confidence of the vote.

Gen. error as a function of Margin Distributions

Assume that the examples are generated i.i.d. according to some distr. D over $X \times \{-1, 1\}$; denote by $\mathbf{Pr}_D[\cdot]$ the probability when (x, y) is chosen from D.

If S is a training set (a sample of size m, $S = \{(x_1, y_1), \cdots (x_m, y_m)\}$), then we denote by $\mathbf{Pr}_S[\cdot]$ the probability when (x, y) is chosen uniformly at random form S.

Theorem 2 *If H finite, then with probability at least $1 - \delta$, $\forall f \in co(H)$, $\forall \theta > 0$,*

$$\mathbf{Pr}_D\left[yf(x) \leq 0\right] \leq \mathbf{Pr}_S[yf(x) \leq \theta] + O\left(\frac{1}{\sqrt{m}}\sqrt{\frac{\ln m \ln |H|}{\theta^2} + \ln \frac{1}{\delta}}\right)$$

Theorem 3 *If H has VCdimension d then with probability at least $1 - \delta$, $\forall f \in co(H)$, $\forall \theta > 0$,*

$$\mathbf{Pr}_D\left[yf(x) \leq 0\right] \leq \mathbf{Pr}_S[yf(x) \leq \theta] + O\left(\frac{1}{\sqrt{m}}\sqrt{\frac{d \ln^2 \frac{m}{d}}{\theta^2} + \ln \frac{1}{\delta}}\right)$$

Note: no dependence on number of weak hypotheses !

A First Lemma

- $N > 0$, C_N - the set of unweighted averages over N elements from H, i.e.

$$C_N = \left\{ g | g(x) = \frac{1}{N} \sum_{j=1}^{N} h_j(x), h_j \in H \right\}$$

- **Lemma 4** *With probability at least* $1 - \delta_N$ *(over the random choice of the training set),* $\forall g \in C_N$, $\forall \theta > 0$,

$$\mathbf{Pr}_D \left[y g(x) \le \frac{\theta}{2} \right] \le \mathbf{Pr}_S \left[y g(x) \le \frac{\theta}{2} \right] + \epsilon_N$$

where

$$\epsilon_N = \sqrt{\frac{1}{2m} \ln \left[\frac{(N+1)|H|^N}{\delta_N} \right]}.$$

A First Lemma - Proof

Proof: For θ and g fixed

$$\mathbf{Pr}_{sample}\Big[\mathbf{Pr}_D\left[yg(x) \leq \tfrac{\theta}{2}\right] > \mathbf{Pr}_S\left[yg(x) \leq \tfrac{\theta}{2}\right] + \epsilon_N\Big]$$
$$\leq \exp\left[-2m\epsilon_N^2\right].$$

By union bound, the probability (taken over a random choice of S) that $\exists g \in C_N$ such that $\mathbf{Pr}_D\left[yg(x) \leq \frac{\theta}{2}\right] > \mathbf{Pr}_S\left[yg(x) \leq \frac{\theta}{2}\right] + \epsilon_N$ is at most $\leq |H|^N \exp\left[-2m\epsilon_N^2\right]$.

Since $yg(x)$ is always a multiple of $\frac{1}{N}$, we finally get that the probability (taken over a random choice of S) that $\exists \theta > 0$, $\exists g \in C_N$ such that $\mathbf{Pr}_D\left[yg(x) \leq \frac{\theta}{2}\right] > \mathbf{Pr}_S\left[yg(x) \leq \frac{\theta}{2}\right] + \epsilon_N$ is at most $\leq (N+1)|H|^N \exp\left[-2m\epsilon_N^2\right]$.

We finally set $(N+1)|H|^N \exp\left[-2m\epsilon_N^2\right] = \delta_N$ and get the desired result. ∎

A Second Lemma

Lemma 5 *With probability at least* $1-\delta$ *(over the random choice of the training set),*$\forall\theta > 0$*,* $\forall N > 0$*,* $\forall g \in C_N$*,*

$$\mathbf{Pr}_D\left[yg(x) \leq \frac{\theta}{2}\right] \leq \mathbf{Pr}_S\left[yg(x) \leq \frac{\theta}{2}\right] + \epsilon_N,$$

where

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

Proof: Just use lemma 1 and plug in $\delta_N = \frac{\delta}{N(N+1)}$. ∎

Main Result

If H finite, then with probability at least $1-\delta$, $\forall f \in co(H)$, $\forall \theta > 0$, we get

$$\mathbf{Pr}_D\left[yf(x) \leq 0\right] \leq \mathbf{Pr}_S[yf(x) \leq \theta] + O\left(\frac{1}{\sqrt{m}}\sqrt{\frac{\ln m \ln |H|}{\theta^2} + \ln\frac{1}{\delta}}\right).$$

Proof

Consider $f \in co(H)$, $f = \sum_{t=1}^{T} a_t h_t$; then f can be associated with a distr. D_f over H as defined by the coefficients a_t.

Moreover we can map f to a distribution Q_f over C_N; a function $g \in C_N$ distributed according to Q_f is generated by choosing $g_1, \cdots, g_N$ ind. at random according to D_f and then defining $g(x) = \frac{1}{N}\sum_{j=1}^{N} g_j(x)$.

Main Result, Proof

Note: If we fix x then $\mathbf{E}_{D_f}\Big[g_j(x)\Big] = \sum_{t=1}^{T} a_t h_t(x) = f(x)$ and $\mathbf{E}_{g\sim Q_f}[g(x)] = f(x)$.

Therefore

$$\mathbf{Pr}_{g\sim Q_f}\left[yg(x) > \frac{\theta}{2}, yf(x) \le 0\right] \le \exp\left[-N\theta^2/8\right]$$

and so

$$\mathbf{E}_D\Bigg[\mathbf{Pr}_{g\sim Q_f}\left[yg(x) > \frac{\theta}{2}, yf(x) \le 0\right]\Bigg] \le \exp\left[-N\theta^2/8\right]$$

or

$$\mathbf{Pr}_{D,g\sim Q_f}\left[yg(x) > \frac{\theta}{2}, yf(x) \le 0\right] \le \exp\left[-N\theta^2/8\right].$$

Similarly

$$\mathbf{Pr}_{S,g\sim Q_f}\left[yg(x) \le \frac{\theta}{2}, yf(x) > \theta\right] \le \exp\left[-N\theta^2/8\right].$$

Main Result, Proof - cont

Consider $f \in co(H)$. For any $g \in C_N$, for any $\theta > 0$ we have:

$$\mathbf{Pr}_D\left[yf(x) \le 0\right] \le \mathbf{Pr}_D\left[yg(x) \le \frac{\theta}{2}\right] + \mathbf{Pr}_D\left[yg(x) > \frac{\theta}{2}, yf(x) \le 0\right]$$

Therefore

$$\mathbf{E}_{g\sim Q_f}[\mathbf{Pr}_D\left[yf(x) \le 0\right]] \le \mathbf{E}_{g\sim Q_f}\left[\mathbf{Pr}_D\left[yg(x) \le \frac{\theta}{2}\right]\right] + \mathbf{E}_{g\sim Q_f}\left[\mathbf{Pr}_D\left[yg(x) > \frac{\theta}{2}, yf(x) \le 0\right]\right]$$

and so

$$\mathbf{Pr}_D\left[yf(x) \le 0\right] \le \mathbf{E}_{g\sim Q_f}\left[\mathbf{Pr}_D\left[yg(x) \le \frac{\theta}{2}\right]\right] + \mathbf{E}_D\left[\mathbf{Pr}_{g\sim Q_f}\left[yg(x) > \frac{\theta}{2}, yf(x) \le 0\right]\right]$$

and therefore

$$\mathbf{Pr}_D\left[yf(x) \le 0\right] \le \mathbf{E}_{g\sim Q_f}\left[\mathbf{Pr}_D\left[yg(x) \le \frac{\theta}{2}\right]\right] + \exp\left[-N\theta^2/8\right].$$

Main Result, Proof - finish

Therefore, by lemma 5 we know that with probability $1-\delta$ (over the random choice of the training set) we have

$$\begin{aligned}\mathbf{Pr}_D\left[yf(x) \le 0\right] \le \exp\left[-N\theta^2/8\right] &+ \\ \mathbf{E}_{g\sim Q_f}\left[\mathbf{Pr}_S\left[yg(x) \le \frac{\theta}{2}\right]\right] &+ \\ \sqrt{\frac{1}{2m}\ln\left[\frac{N(N+1)^2|H|^N}{\delta}\right]}\end{aligned}$$

and so

$$\mathbf{Pr}_D\left[yf(x) \le 0\right] \le 2\exp\left[-N\theta^2/8\right] + \mathbf{Pr}_S\left[yf(x) \le \theta\right] + \sqrt{\frac{1}{2m}\ln\left[\frac{N(N+1)^2|H|^N}{\delta}\right]}.$$

Choosing $N = \frac{4}{\theta^2}\ln\left[\frac{m}{\ln|H|}\right]$ we get the desired result.

Boosting increases the margin

Theorem 6 *Suppose the base learning algorithm, when called by AdaBoost, generates classifiers with weighted errors* $\epsilon_1, \cdots, \epsilon_T$*. Then for any* θ *we have*

$$\mathbf{Pr}_S\left[yf(x) \leq \theta\right] \leq 2^T \prod_{t=1}^{T} \sqrt{\epsilon_t^{(1-\theta)}(1-\epsilon_t)^{(1+\theta)}}$$

Interpretation: if $\forall t$, $\epsilon_t < \frac{1}{2} - \gamma$ and if $\theta < \gamma$, then $\mathbf{Pr}_S\left[yf(x) \leq \theta\right]$ goes to 0 as $T \to \infty$.

(If θ is not too large, then the fraction of the training examples for which $yf(x) \leq \theta$ decreases exponentially to 0 exponentially fast with the number of base classifiers.)