Is boosting a universal learning algorithm?

Yoav Freund

AT&T Research

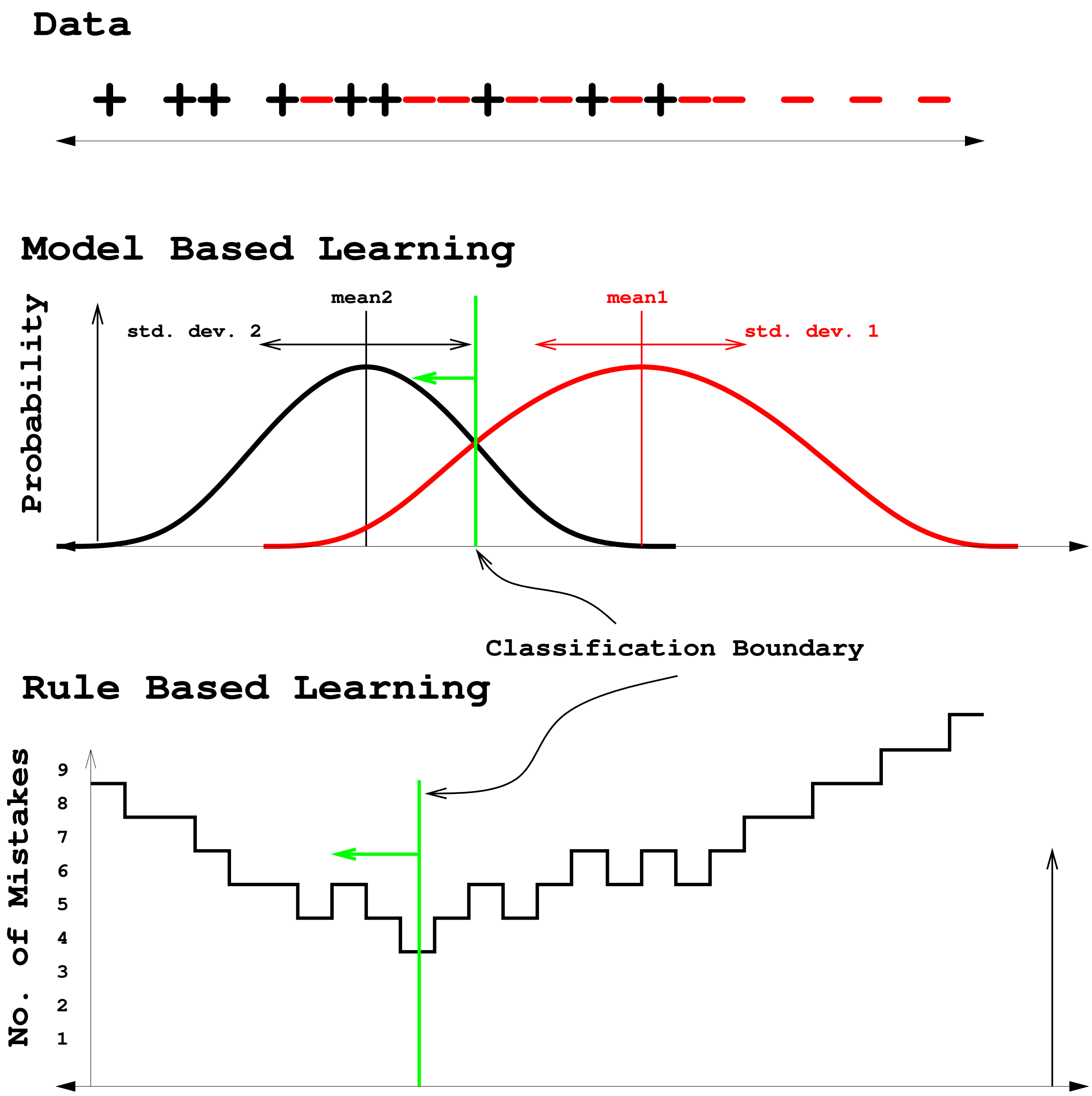
Papers and transparencies in:
http://www.research.att.com/∼yoav

No!

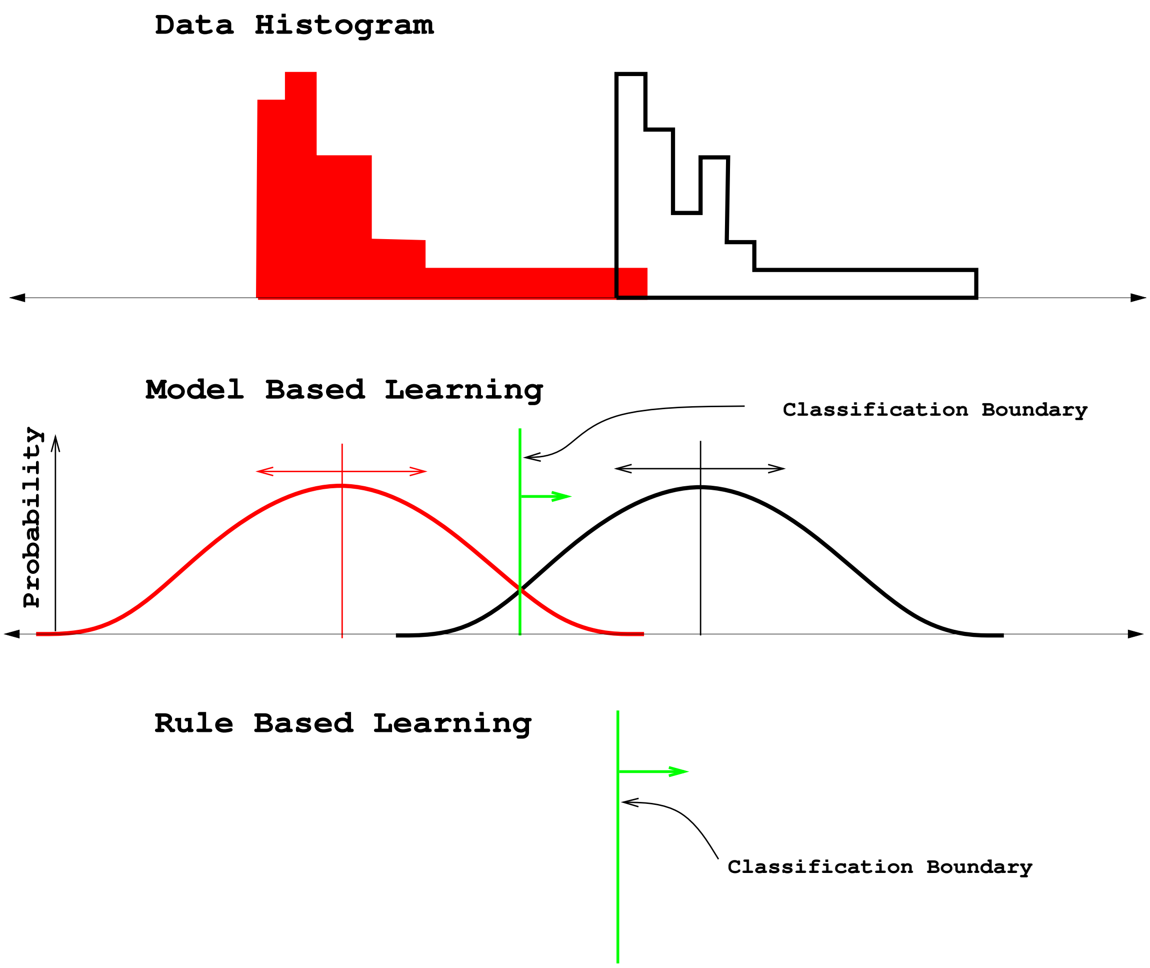
Plan of talk

1. Training error
   (a) Two methods for learning classification rules.
   (b) The assumptions underlying boosting.
   (c) Conditions under which boosting is guaranteed to work.
   (d) Example where boosting fails.
2. Test error
   (a) Can boosting over-fit?
   (b) The Margins explanation.
   (c) Do the Margins explain everything?
3. Summary: how do theory and experiments relate?

Two approaches to learning classification rules



Using an inappropriate distribution model



The assumptions underlying boosting

"Weak" learner:

- A learning method that can generate classification rules that are *slightly* correlated (or anti-correlated) with the truth.
- Should work for altered distributions (weighted training set).

Main (old) result regarding training error of **AdaBoost**

Theorem:

If running **AdaBoost** generates hypotheses whose weighted training errors are:

$$\text{error}_{D_t}(h_t) = \epsilon_t = 1/2 - \gamma_t$$

Then

$$\text{Training error}(h_{\text{final}}) \leq \prod_{t=1}^{T} \sqrt{1 - 4\gamma_t^2} \leq e^{-2(\sum_t \gamma_t^2)}$$

Sufficient condition for boosting to work

Theorem:

If there exists a rule of the form

$$f(x) = \mathrm{sign}\left(\sum_{i=1}^{n} h_i(x)\right), \quad h_i \in H, \quad n \text{ is odd}$$

that is consistent with the training set:
$\forall (x, y) \in T,\ f(x) = y.$

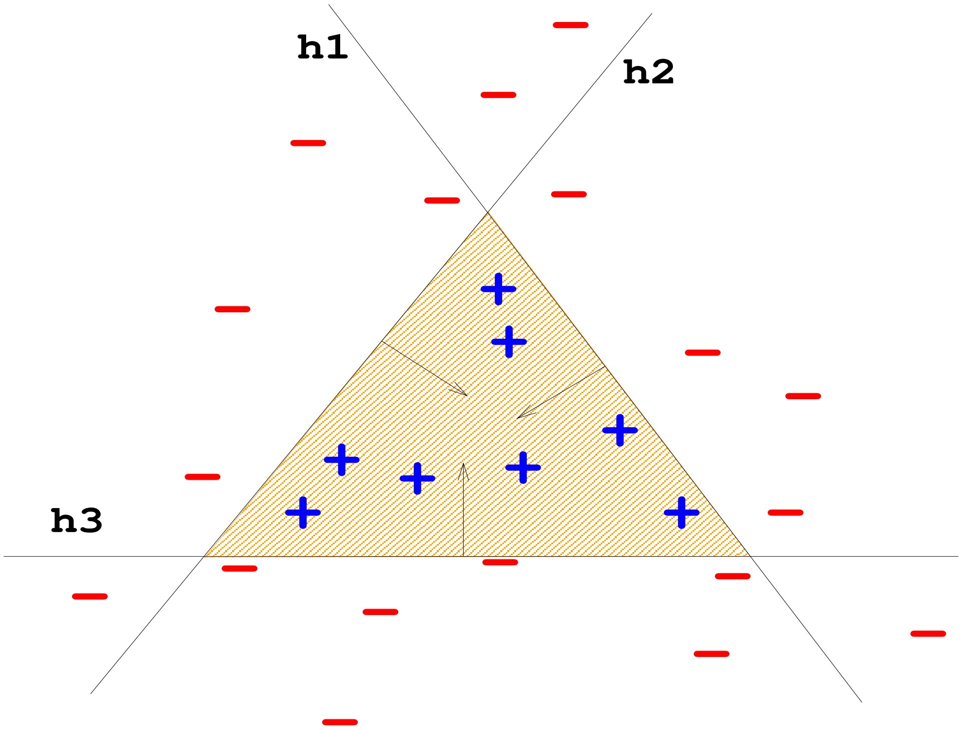
Then $\forall P$ distribution over T,
$\exists i, 1 \le i \le n$ such that

$$\sum_{(x,y)\in T} P(x, y)\ \mathbf{1}\left(h_i(x) = y\right) \ge \frac{1}{2} + \frac{1}{2n}$$

Corollary:

If AdaBoost uses H as the set of base hypotheses
Then it will generate a consistent hypothesis in at most $2n^2 \ln |T|$ iterations.

An example of the sufficient condition



Let $h_4(x) = -1$ for all x.

A consistent rule:

$$\text{sign}(2h_1(x) + 2h_2(x) + 2h_3(x) + 5h_4(x))$$

$n = 11 \Rightarrow$ boosting a learner that select the linear discriminator with smallest weighted error generates a consistent rule within $2(11^2)\ln|T| = 242\ln|T|$ iterations.

Note: boosting a weak learner that does not minimize the weighted training error (such as logistic regression) might get stuck.

Proof of the lemma

1.

$$\forall x \in T \quad \sum_{i=1}^{n} \mathbf{1}\left(h_i(x) = y\right) > \frac{n}{2}$$

2. For any distribution P over T:

$$\sum_{(x,y)\in T} P(x,y) \left(\sum_{i=1}^{n} \mathbf{1}\left(h_i(x) = y\right) \right) \geq \left\lceil \frac{n}{2} \right\rceil = \frac{n}{2} + \frac{1}{2}$$

3.

$$\sum_{i=1}^{n} \left(\sum_{(x,y)\in T} P(x,y)\, \mathbf{1}\left(h_i(x) = y\right) \right) \geq \frac{n}{2} + \frac{1}{2}$$

4. $\exists i, 1 \leq i \leq n$ such that

$$\sum_{(x,y)\in T} P(x,y) \mathbf{1}\left(h_i(x) = y\right) \geq \frac{1}{2} + \frac{1}{2n}$$

A case in which boosting will not work

$x_1, \ldots, x_n$ binary features in $\{-1, +1\}$.
y a binary label in $\{-1, +1\}$.

The joint distribution over $x_1, \ldots, x_n, y$ is such that

- $\forall i, 1 \leq i \leq n,\ E(x_i y) = 0$
- for some $1 \leq i, j \leq n$,

$$\forall (x_1, \ldots, x_n, y), \quad x_i x_j = y$$

Then

- If the base learner searches for the best single feature (n possibilities), boosting will fail.
- If the base learner searches for the best feature pair ($n(n-1)/2$ possibilities) the base learner succeeds immediately.
- Other greedy learning algorithms, such as **C4.5** or **CART** suffer from the same limitation.

Generalization error

Is it true that boosting never over-fits?

No!

Example of over-fitting with boosting

H is the set of all delta-functions:

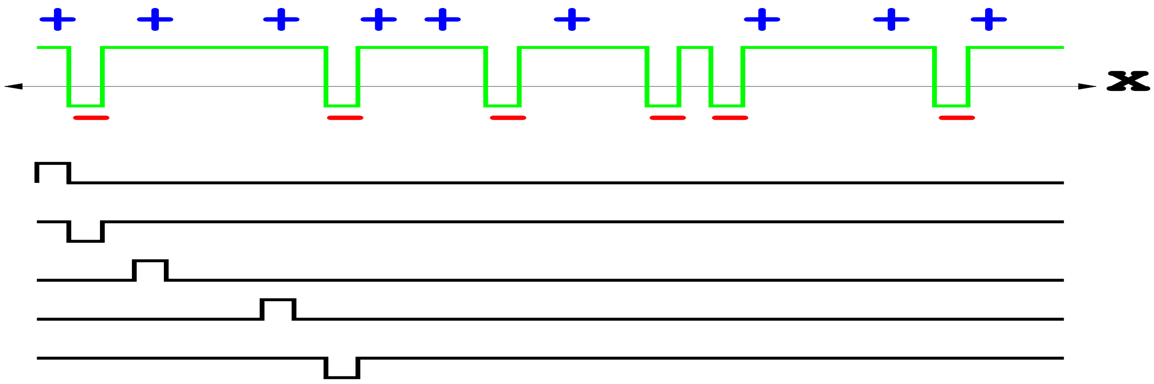
$$h_{a+}(x) = \begin{cases} +1, & \text{if } x = a \\ -1, & \text{otherwise} \end{cases}.$$

$$h_{a-}(x) = \begin{cases} -1, & \text{if } x = a \\ +1, & \text{otherwise} \end{cases}.$$

For any distribution P over a *randomly labeled* training set T, there exists $h \in H$ whose weighted error is at most $1/2 - 1/(2|T|)$

$\Rightarrow$

AdaBoost will find a consistent hypothesis which is a vote over at most $2m^2 \ln m$ functions from H



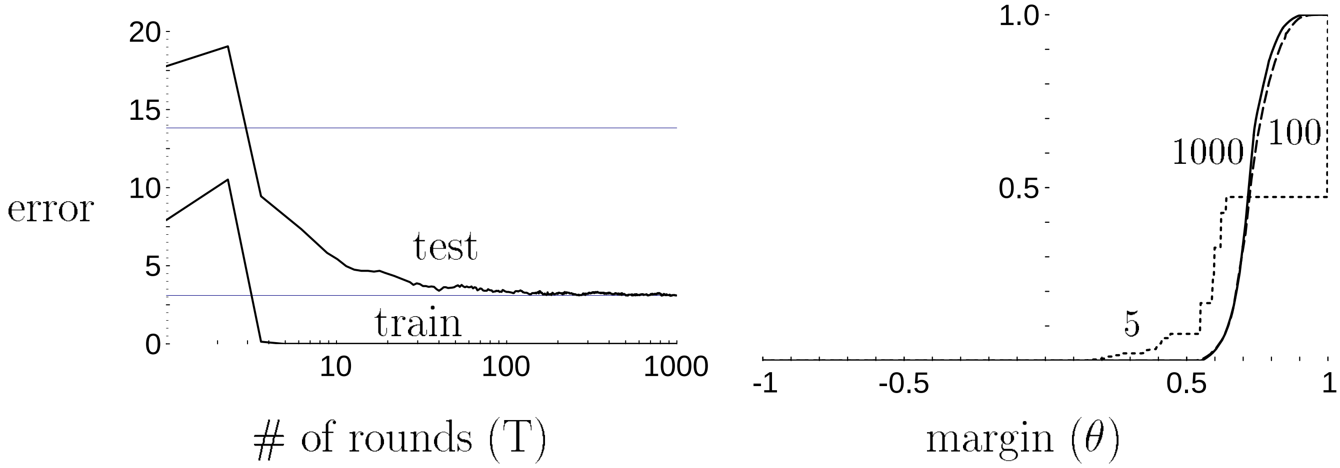
Obviously, nothing can be learned from a randomly labeled set of examples

$\Rightarrow$

The expected *test error* is $1/2$.

Experimental margins

- margin distribution = cumulative distribution of margins on training set



	# rounds		
	5	100	1000
% train error	0.0	0.0	0.0
% test error	8.4	3.3	3.1
% margin ≤ 0.5	7.7	0.0	0.0
$\theta_{\min}$	0.14	0.52	0.55

- between rounds 5 and 100:
  - training error unchanged
  - BUT significant shift in margin distribution
- In the previous example the margin is very small $O(1/|T|)$.

Theory of margins

$\mathcal{H}$ - hypothesis class of weak learner, $\text{VCdim}(\mathcal{H}) = d$
$\mathcal{C}$ - convex combinations of $h \in \mathcal{H}$.

$$\mathcal{C} \doteq \left\{ f : x \mapsto \sum_{h \in \mathcal{H}} a_h h(x) \,\middle|\, a_h \geq 0; \;\; \sum_h a_h = 1 \right\}$$

Theorem 1
S = Training set of m examples, i.i.d. from D.
with high probability over choice of S:
$\forall f \in \mathcal{C}, \forall \theta > 0$*:*

$$P_{(x,y)\sim D}\left[sign(f(x)) \neq y\right] \;\leq\; P_{(x,y)\sim S}\left[margin_f(x,y) \leq \theta\right] + \tilde{O}\left(\frac{\sqrt{d/m}}{\theta}\right)$$

- No dependence on number of weak hypotheses (= boosting iterations).
- This theorem is relevant for *any voting method*
- Not just for the *minimum* margin $P_{(x,y)\sim S}\left[\text{margin}_f(x,y) \leq \theta\right] = 0$
  (Used as goal in [Breiman97, Grove & Schuurmans 98])

Do the margins explain everything?

No!

- If the base rules are too complex, boosting might not help.
- VC dimension gives an upper bound on over-fitting, but there are tighter bounds.
- The margin is one way for upper bounding the test error of rules that are based on thresholding a convex sum.
  There might be other measures that yield tighter bounds!

Summary: how do theory and experiments relate?

False rumors:

- Boosting is a universal learning method.
- Boosting never over-fits.
- The margin analysis suggests maximizing the *minimal* margin.
- The margins contain all the information about the generalization error.

Conclusions:

- The story is more complex, read the papers!
- Rigorous and correct theory gives *bounds* which relate different quantities. Experiments cannot falsify these bounds.
- Proven bounds have explicit and clear assumptions. The validity of these *assumptions* can be tested in experiments.
- Experiments can compare different bounds.
- Experiments can suggest new measures and bounds.