Boosting Versus Covering

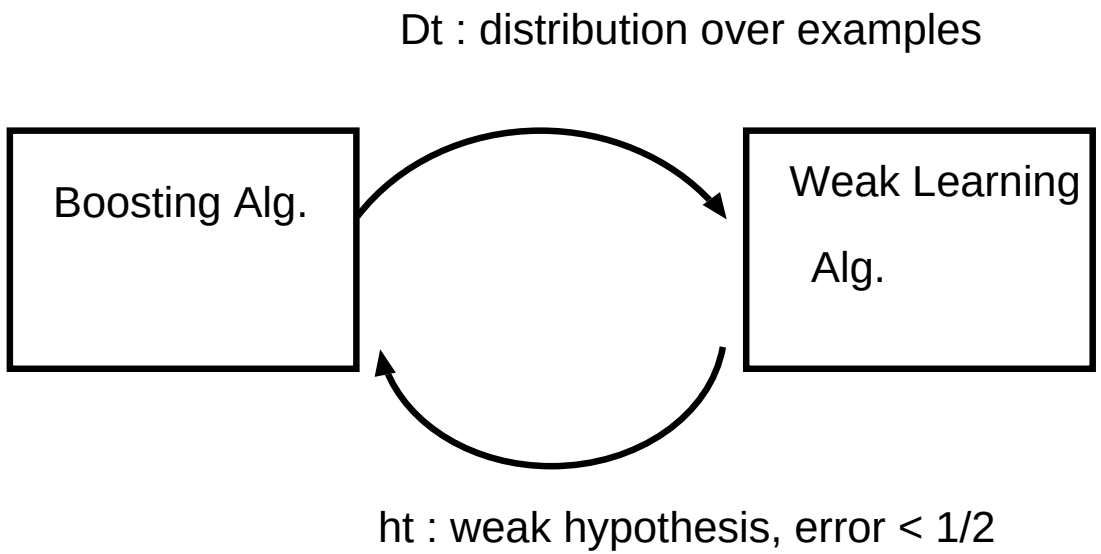
Kohei. Hatano Tokyo Institute of Technology

Manfred K. Warmuth UC Santa Cruz

Thanks to Gunnar Rätsch for many key ideas

Boosting

Given: m labeled examples $(x_1, y_1), \ldots, (x_m, y_m)$
$(x_i \in \mathcal{X},\ y_i \in \{-1, +1\})$



Goal: Construct consistent final hypothesis
- by combining weak hypotheses

Final hypothesis: $f(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$

Efficiency of Boosting

Assumption: Weak hypotheses have error at most $\frac{1}{2} - \frac{\gamma}{2}$

Size of final consistent hypothesis / # of iterations

- $O(\frac{1}{\gamma^2} \log m)$ [F95,FS97]
- $\Omega(\frac{1}{\gamma^2} \log m)$ [F95]
- If final hypothesis restricted to unweighted majority

Boosting with One-sided-Error

Hypothesis h is one-sided
iff positive predictions of h are always correct

$O(\frac{1}{\gamma}\log m)$ iterations suffice [N91]

Natarajan's "covering" algorithm
works only when each weak hypothesis one-sided

Our Motivation

Boosting algorithms that takes advantage of one-sidedness

Degree of one-sidedness	all $\longleftrightarrow$ none
Bound:	$O(\frac{1}{\gamma}) \longleftrightarrow O(\frac{1}{\gamma^2})$
Both types of one-sidedness	

Results

- AdaBoost [FS97] requires $\Omega(\frac{1}{\gamma^2}\log m)$ iterations even when weak hypotheses are one-sided (under some restrictions)
- InfoBoost [A01][a] finds consistent hypothesis in $O(\frac{1}{\gamma}\log m)$ iterations when weak hypotheses are one-sided
- Solving subtle technical problems with InfoBoost

[a] Equivalent to confidence-rated AdaBoost[SS99] with binary domain-partitioning hypotheses.

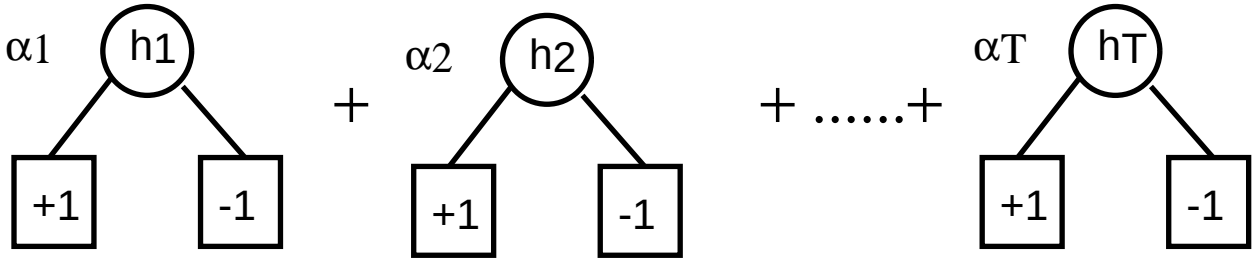
AdaBoost [FS97]

- Initialize $D_1(i) = \frac{1}{m}$
- At each iteration t,
  - Get $h_t : \mathcal{X} \to \{-1, +1\}$.
  - Update

$$D_{t+1}(i) = \frac{D_t(i) \exp\{-y_i h(x_i) \alpha_t\}}{Z_t}$$

where $\alpha_t = \frac{1}{2} \ln \frac{1+\gamma_t}{1-\gamma_t}$ and $\gamma_t = \sum D_t(i) y_i h_t(x_i)$

- Final hypothesis: the sign of $H(x) = \sum_t^T \alpha_t h_t(x_i)$



InfoBoost [A00]

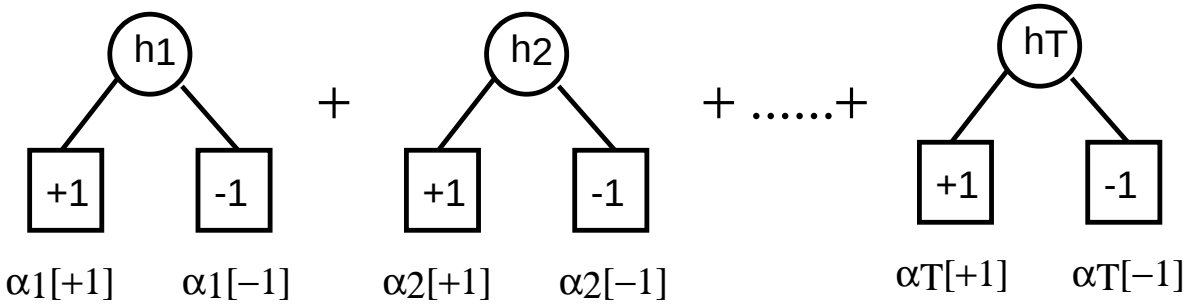
As AdaBoost except

- Update rule:

$$D_{t+1}(i) = \frac{D_t(i)\exp\{-y_i h(x_i)\alpha_t[h(x_i)]\}}{Z_t},$$

where $\alpha_t[\pm 1] = \frac{1}{2}\ln\frac{1+\gamma_t[\pm 1]}{1-\gamma_t[\pm 1]}$ and
$\gamma_t[+1] = \frac{\sum_{i:h_t(x_i)\geq 0} y_i h(x_i) D_t(i)}{\sum_{i:h_t(x_i)\geq 0} D_t(i)}$, and $\gamma_t[-1] = \frac{\sum_{i:h_t(x_i)<0} y_i h(x_i) D_t(i)}{\sum_{i:h_t(x_i)<0} D_t(i)}$

- Final hypothesis: Sign of $H(x) = \sum_t^T \alpha_t[h_t(x_i)]h_t(x_i)$



UpperBounds on training error

AdaBoost:

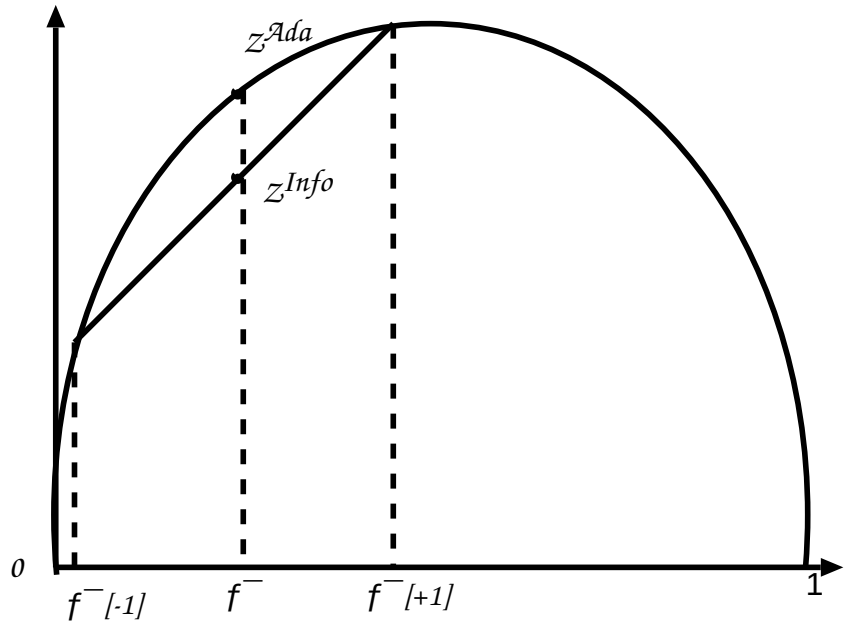
$$\prod_{t=1}^{T} Z_t^{Ada} = \prod_{t=1}^{T} \sqrt{1-\gamma_t^2}$$

InfoBoost:

$$\prod_{t=1}^{T} Z_t^{Info} =$$

$$\prod_{t=1}^{T} \{\Pr_{D_t}[h_t(x_i) \geq 0]\sqrt{1-\gamma_t[+1]^2} + \Pr_{D_t}[h_t(x_i) < 0]\sqrt{1-\gamma_t[-1]^2}\}$$

UpperBounds on training error(2)



By the concavity of $q(x) = \sqrt{1-x^2}$,

$$
\begin{aligned}
Z^{info} &= Pr[h(x) \geq 0]q(\gamma[+1]) + Pr[h(x) < 0]q(\gamma[-1]) \\
&\leq q(Pr[h(x) \geq 0]\gamma[+1] + Pr[h(x) < 0]\gamma[-1]) \\
&= q(\gamma) = Z^{Ada}
\end{aligned}
$$

Motivating update rules[KW98]

$$\max_{\alpha \in \mathbf{R}} (-\ln Z^{Ada}) = \min_{D: \sum_{i=1}^{m} D(i) y_i h_t(x_i) = 0} \Delta(D, D_t),$$

where $\Delta(D, D') = \sum_i D(i) \ln \frac{D(i)}{D'(i)}$ is relative entropy

Constraints

AdaBoost: $\sum_{i=1}^{m} D(i) y_i h_t(x_i) = 0$

InfoBoost: $\sum_{i=1:h_t(x_i)=+1}^{m} D(i) y_i h_t(x_i) = 0$

Second constraint for $h_t(x_i) = -1$

Constraints (the range of h_i is ± 1)

D_{t+1}:	$y_i \setminus h_t$	$+1$	-1
	$+1$	a	b
	-1	c	d

AdaBoost: $a + d = b + c$ (error $= 0$)

InfoBoost:

$a = c$ and $b = d$ $\iff$ $a + d = b + c$ and $a + b = c + d$

$\iff$ Mutual information between h_t and y is zero

$\iff$ Error $= 0$ and weight of positive examples is $\frac{1}{2}$

Simultaneous vs Sequential constraints

Info: $\sum_{i=1}^{m} D(i) y_i h_t(x_i) = 0$, $\sum_{i=1}^{m} D(i) y_i \mathbf{1} = 0$ (simultanious)

Let us consider sequential constraints (AdaBoost with Bias): $\sum_{i=1}^{m} D(i) y_i h_t(x_i) = 0$(at odd steps), then $\sum_{i=1}^{m} D(i) y_i \mathbf{1} = 0$(at even steps),...

AdaBoost with Bias is similar to AdaBoost except bias in the final hypothesis.

Illustration of Updates

D_t :

$y_i \setminus h_t$	$+1$	-1
$+1$	$\frac{2}{5}$	$\frac{2}{5}$
-1	0	$\frac{1}{5}$

D_t:**AdaB.**

$y_i \setminus h_t$	$+1$	-1
$+1$	$\frac{1}{3}$	$\frac{1}{2}$
-1	0	$\frac{1}{6}$

D_{t+1} :**InfoB.**

$y_i \setminus h_t$	$+1$	-1
$+1$	0	$\frac{1}{2}$
-1	0	$\frac{1}{2}$

D_{t+1} :**AdaB.w.Bias**

$y_i \setminus h_t$	$+1$	-1
$+1$	$\frac{1}{5}$	$\frac{3}{10}$
-1	0	$\frac{1}{2}$

Learning k-disjunctions

Find consistent hypothesis for m examples labeled by a k-disjunction over N boolean variables,

$$g = X_{i_1} \vee X_{i_2} \vee \cdots \vee X_{i_k}$$

Weak hypotheses: Literals and constants
Each X_{i_j} one-sided: $\gamma \geq \frac{1}{4k}$

AdaBoost: $O(k^2 \log m)$ iterations [FS97]

Greedy set-covering: $O(k \log m)$ iterations [J73]

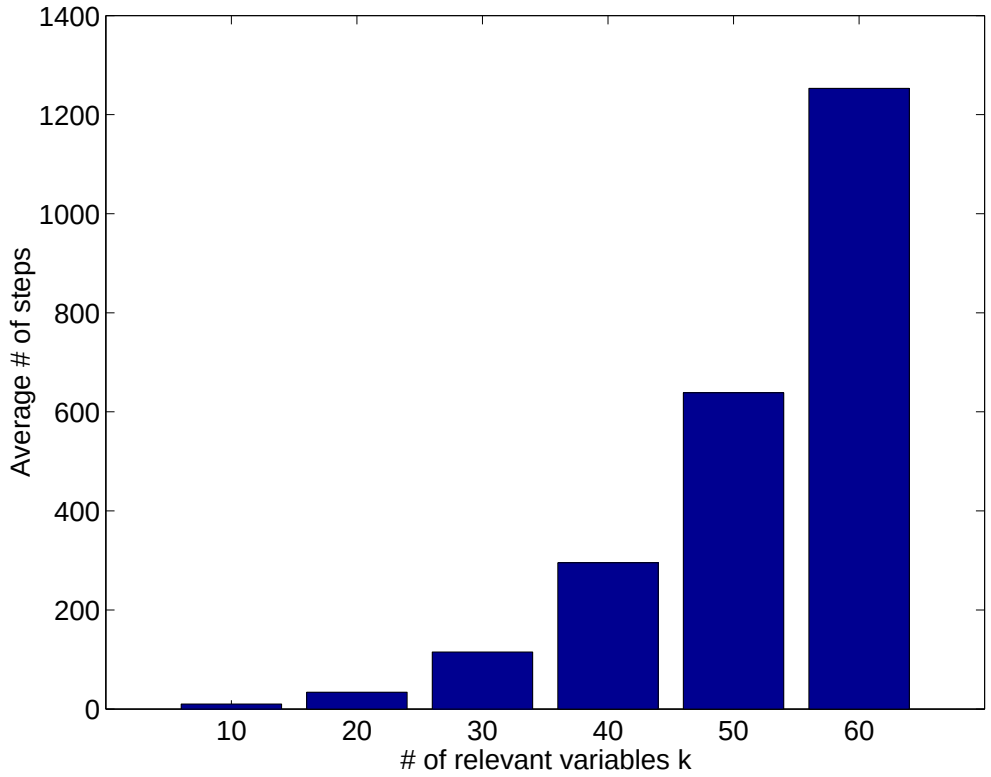
Lowerbound of AdaB(w. Bias) for k disjunctions

1. Greedy: Choose a hypothesis with minimum error from a pool of hypotheses (Experiment results)

2. Minimal: Choose hypothesis with error below a threshold $\delta = 1/2 - 1/(4k)$ (Bounds for this particular adversary)

Experimental results

Average # of iterations for AdaBoost with Bias:



Greedy-set-covering alg. and InfoBoost found consistent hypotheses in $O(k)$ iterations over this artificial data

Construction of Adversary

Thm. For any $\varepsilon \in (0, 1)$, and sufficiently large N, there exists an initial distribution over S, and sequences of choices by the adversary for which AdaBoost needs

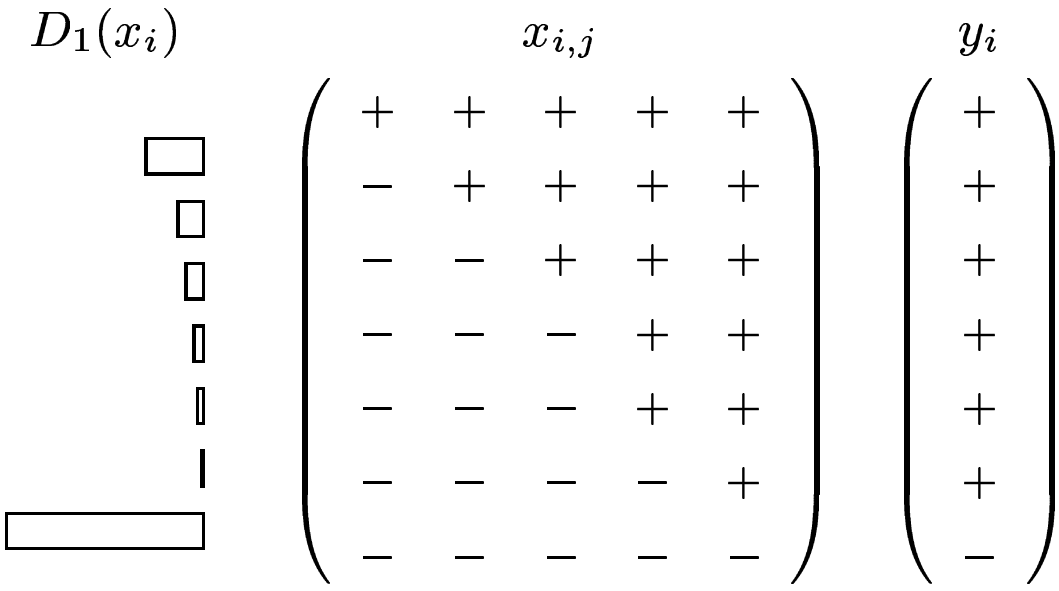
$$\frac{k^2}{2} \ln \frac{1}{2\varepsilon} - \frac{k}{2}$$

iterations to bound the training error of the final hypothesis below ε.

Idea of the proof

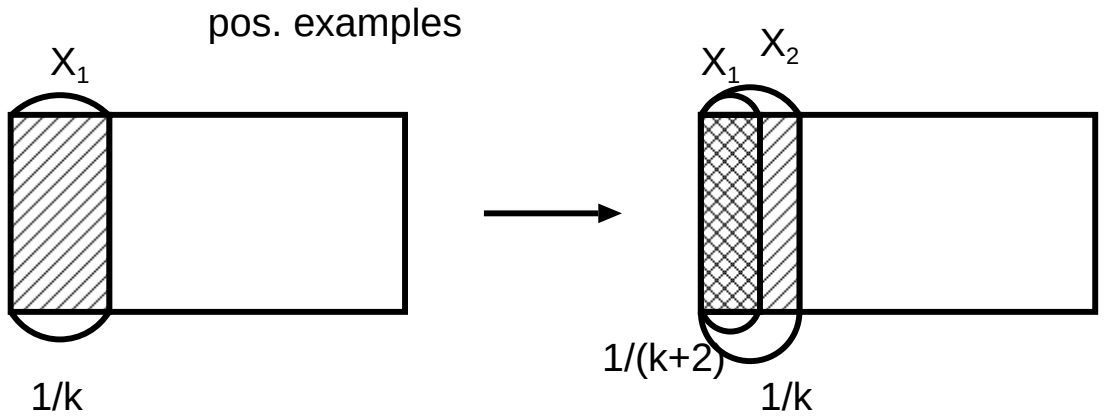
Initial distribution D_1:

$$D_1(x_t) := \begin{cases} \frac{1}{2k}, & \text{for } t = 1 \\ \frac{1}{k(k+1)} \left(1 - \frac{1}{k}\right) \left(1 - \frac{2}{k(k+1)}\right)^{t-2}, & \text{for } 2 \leq t \leq N-1 \\ \frac{1}{2} - \sum_{t=1}^{N-1} D(x_t), & \text{for } t = N \\ \frac{1}{2}, & \text{for t=}N+1. \end{cases}$$



NOTE: The set of examples are consistent with not only a k-disjunction but also a 1-disjunction.

Idea of the proof(cont.)



# of positive examples that are not "covered" by past literals (white) reduces by the factor of

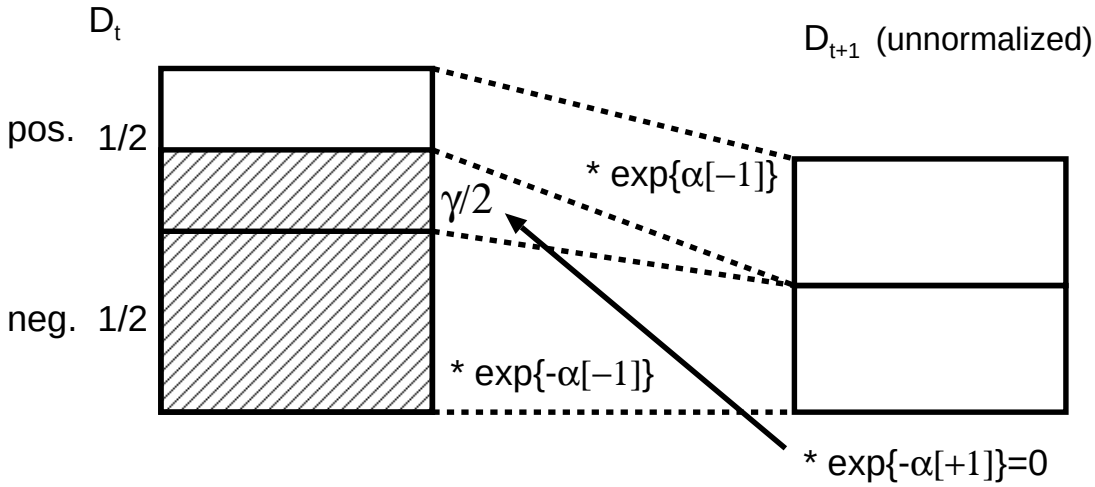
$$\frac{\frac{k-1}{k}}{\frac{k+1}{k+2}} = \frac{(k-1)(k+2)}{k(k+1)} = 1 - \frac{2}{k(k+1)}.$$

$O(\frac{1}{\gamma})$ bound of InfoBoost

Suppose that h_t is one-sided and has error $\frac{1}{2} - \frac{\gamma}{2}$

$\Rightarrow \gamma_t[+1] = 1$, and thus $\alpha_t[+1] = +\infty$

Tha fact $\alpha_t[+1]h_t(x) = +\infty$ implies $H(x) = +\infty$, i.e., x is correctly classified by H.



Since $D_t(pos.) = 1/2$ for $t \geq 2$, # of "uncovered" positive examples reduces by the factor of $1 - \gamma$

Technical problems of InfoBoost

- For positive (or negative) examples only, InfoBoost outputs a constant hypothesis ($+1$ or -1) and terminates
- However, in some applications, one may want more information, e.g., relevant variables

NOTE: Similar situations can happen during iterations when weights on positive and negative examples are highly unbalanced.

A modification: SemiBoost

The final hypotheisis of InfoBoost can be written as linear combination of h_t^+s and h_t^-s,

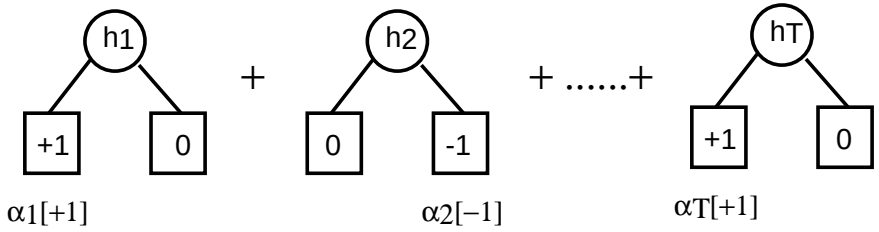
$$H(x) = \sum_t \alpha_t[+1]h_t^+ + \sum_t \alpha_t[-1]h_t^-,$$

where $h_t^{\pm}(x) = \pm 1$ if $h_t(x) = \pm 1$, and $h_t^{\pm}(x) = 0$ otherwise

SemiBoost: Instead of using the set of base hypotheses H, use

$$H^{\pm} = \{h^+, h^- | h \in H\},$$

and run AdaBoost



Avoids the technical problem

Discussion: Generalization Error

Occam Razor(# of iterations) doesn't give enough explanation on generalization error! [SFBW98]

Better explanation: (margin) $= \min_{(x_i,y_i)\in S} y_i H(x_i)/|H|$

Upperbound of the generalization error (for noise-free settings)[SS99]:

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

with probability $1-\delta$ (d: pseudodimension of the class of base hypotheses).

One too huge coefficient implies small margin,(and therefore) large upperbound of gen. error.

Conflict between Pos. and Neg. One-sidedness

Suppose that h_3 is (positively) one-sided and h_{10} is *negatively* one-sided.

$$H(x) = \cdots + \infty + \cdots - \infty + \cdots = ?$$

Desicion list interpretation: adopt the infinite coefficient with the minimum index (h_3)

If x is in the given sample, $H(x)$ is correct.

Open: How can we convert infinite coefficients to finite ones without losing consistency?

Open Problems

1. Does AdaBoost require $\Omega(k^2 \log m)$ iterations if it chooses hypotheses greedily?

2. We still don't know any intermediate upperbound between $O(1/\gamma)$ and $O(1/\gamma^2)$ for the cases where some of chosen weak hypotheses are not one-sided.

3. Practical performance of SemiBoost