Maximizing the Margin with Boosting

Gunnar Rätsch
Australian National University
Canberra, Australia

Manfred K. Warmuth
University of California
Santa Cruz, California

THE AUSTRALIAN NATIONAL UNIVERSITY

UC SANTA CRUZ

Ensembles for Classification & Regression

- Ensemble for binary classification consists of
  - Weak hypotheses h_t, generated by a weak learner L
  - Hypothesis weights $\alpha_t \geq 0$
- Final hypothesis
  - $f_{\boldsymbol{\alpha}}(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x})$

⇒ Bagging [B96]

⇒ **AdaBoost** [FS94]

⇒ **Arc-GV** [B97]

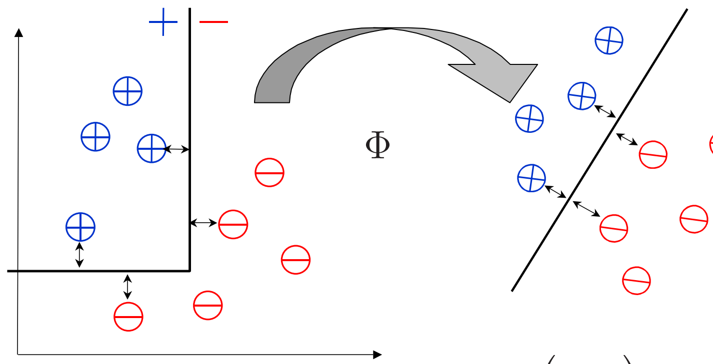
⇒ **Sparse Regression** [RBB02]

The AdaBoost$_\varrho$ Algorithm [B97]

1. **Input:** $S = \langle(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\rangle$, No. Iterations T, $\boxed{\text{margin target } \varrho}$
2. **Initialize:** $d_n^1 = 1/N$ for all $n = 1 \ldots N$
3. **Do for** $t = 1, \ldots, T$,
   (a) Train classifier on $\{S, \mathbf{d}^t\}$ and obtain hypothesis $h_t : \mathbf{x} \mapsto [-1, 1]$
   (b) Calculate the edge γ_t of h_t: $\gamma_t = \sum_{n=1}^{N} d_n^t y_n h_t(\mathbf{x}_n)$.
   (c) **if** $\gamma_t \leq \varrho$ or $\gamma_t = 1$, **then** break
   (d) Set $\alpha_t = \frac{1}{2} \log \frac{1+\gamma_t}{1-\gamma_t}$ $\boxed{-\frac{1}{2} \log \frac{1+\varrho}{1-\varrho}}$.
   (e) Update weights: $d_n^{t+1} = d_n^t \exp\{-\alpha_t y_n h_t(\mathbf{x}_n)\}/Z_t$, s.t. $\sum_{n=1}^{N} d_n^{t+1} = 1$.
4. **Output:** $f(\mathbf{x}) = \sum_{t=1}^{T} \frac{\alpha_t}{\sum_r \alpha_r} h_t(\mathbf{x})$, $\boxed{\gamma = \min_{t=1,\ldots,T} \gamma_t,\ \rho = \min_{n=1,\ldots,N} y_n f(\mathbf{x}_n)}$

Large Margin and Linear Separation

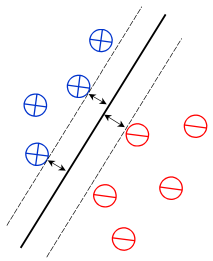
Input space $\mathcal{X}$ Feature space $\mathcal{F}$



Linear separation in $\mathcal{F}$ is nonlinear separation in input space

$$\Phi(\mathbf{x}) = \begin{pmatrix} h_1(\mathbf{x}) \\ h_2(\mathbf{x}) \\ \vdots \end{pmatrix}, \quad H = \{h_1, h_2, \ldots\}$$

Boosting and Large Margins



- Hyperplane = combined hypothesis

$$f_{\boldsymbol{\alpha}}(\mathbf{x}) = (\Phi(\mathbf{x}) \cdot \boldsymbol{\alpha}) = \sum_{j=1}^{J} \alpha_j h_j(\mathbf{x})$$

- margin $\rho(\boldsymbol{\alpha}) = \min_n \underbrace{\frac{y_n(\Phi(\mathbf{x}_n) \cdot \boldsymbol{\alpha})}{\|\boldsymbol{\alpha}\|_1}}_{\text{distance}}$
- distance = ℓ_∞-distance [M99]

Weak Learner Ł

- One uses an example weighting $d_1, \ldots, d_N$
- Weak learner Ł returns hypotheses $H \ni h : \mathbf{x} \mapsto \{\pm 1\}$

  …which are consistently *better than random guessing*

$$\underbrace{\sum_{n=1}^{N} d_n \, \mathrm{I}(h(\mathbf{x}_n) \neq y_n)}_{\text{weighted tr. error of } h} \leq \frac{1}{2} - \frac{1}{2}\gamma$$

  for any weighting $\mathbf{d}$ and some fixed $\gamma > 0$

- For real valued hypotheses $h : \mathbf{x} \mapsto [-1, 1]$

$$\underbrace{\sum_{n=1}^{N} d_n \; y_n h(\mathbf{x}_n)}_{\text{Edge of } h} \geq \gamma$$

Edge and Margin

- Edge of a weak learner Ł for a probability vector $\mathbf{d}$ on the examples

$$e(\mathbf{d}) := \max_{h \in H} \sum_{n=1}^{N} d_n \; y_n h(x_n)$$

- Margin of the combined hypothesis $f_{\boldsymbol{\alpha}}$

$$\rho(\boldsymbol{\alpha}) := \min_{n=1 \ldots N} y_n \frac{f_{\boldsymbol{\alpha}}(\mathbf{x}_n)}{\sum_t \alpha_t} = \min_{n=1 \ldots N} y_n \sum_{t=1}^{T} \frac{\alpha_t}{\sum_r \alpha_r} h_t(\mathbf{x}_n)$$

Min-Max-Theorem:

$$\min_{\mathbf{d}} e(\mathbf{d}) = \max_{\boldsymbol{\alpha}} \rho(\boldsymbol{\alpha})$$

Main Result

Modify AdaBoost for maximizing margin

- the algorithm requires

$$\mathcal{O}\left(\frac{\log(N)}{\epsilon^2} \log\left(\frac{1}{\epsilon}\right)\right) \text{ iterations to get margin in } [\rho^* - \epsilon, \rho^*]$$

- original AdaBoost finds a solution with large, but not maximal margin

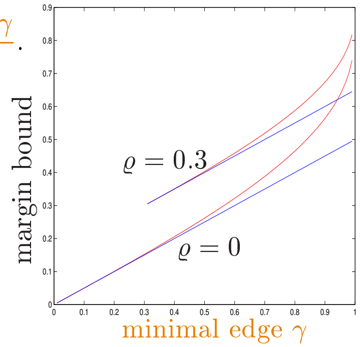
  … it might even not converge

- Arc-GV asymptotically maximizes the margin (no converge rate)

Asymptotic Margin for AdaBoost$_\varrho$ [ROM01]

Corollary 1 *Assume AdaBoost$_\varrho$ generates hypotheses $h_1, h_2, \ldots$ with edges $\gamma_1, \gamma_2, \ldots$ and coefficients $\alpha_1, \alpha_2, \ldots$. Let $\gamma = \inf_{t=1,2,\ldots} \gamma_t$ and assume $\boxed{\gamma > \varrho}$. Then, asymptotically ($t \to \infty$), the smallest margin ρ of the combined hypothesis is bounded from below by*

$$\rho \geq \frac{\log(1-\varrho^2) - \log(1-\gamma^2)}{\log\left(\frac{1+\gamma}{1-\gamma}\right) - \log\left(\frac{1+\varrho}{1-\varrho}\right)} \geq \frac{\varrho + \gamma}{2}.$$

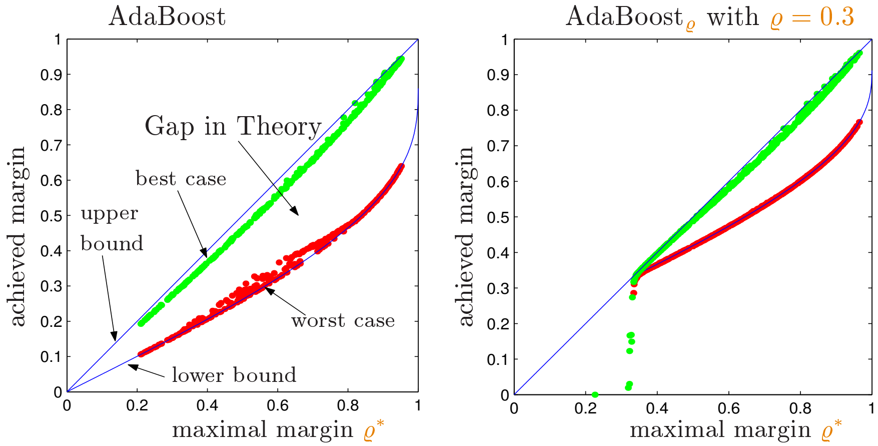


Arc-GV: Set $\varrho := \max_{r=1,\ldots,t-1} \rho_t$

$\Rightarrow$ easy convergence proof

$\Rightarrow$ no convergence rate

Illustration on Random Problems [R01]



$\Rightarrow$ There is a big GAP!

$\Rightarrow$ AdaBoost might not converge

$\Rightarrow$ Choose ϱ, then gap becomes small

$\Rightarrow$ Arc-GV $\Rightarrow$ No Convergence Rate

How to find ϱ^*? $\Rightarrow$ Marginal AdaBoost

- Run AdaBoost and construct interval $[a, b]$ for ϱ^*:

$$\max_t \rho_t = a \leq \varrho^* \leq b = \min_t \gamma_t$$

- Use a binary search to find $\varrho \approx \varrho^*$: chose $\varrho = \dfrac{a+b}{2}$
- Run AdaBoost$_\varrho$: if $\max_t \rho_t \geq \varrho$, then $a = \max_t \rho_t$
  else $b = \min_t \gamma_t + \epsilon$
- repeat $\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$ times

Complexity of the algorithm to find $\varrho \in [\rho^* - \epsilon, \varrho^*]$ and $f_{\boldsymbol{\alpha}}$

$$\mathcal{O}\left(\underbrace{\frac{\log(N)}{\epsilon^2}}_{T} \log\left(\frac{1}{\epsilon}\right)\right)$$

Some Results [SFBL98,RW02]

Lemma Assume that in a run of AdaBoost$_\varrho$ holds $\alpha_t \geq 0$ for all t. Then we have

$$\frac{1}{N}\sum_{n=1}^{N} \mathrm{I}\left(\mathbf{y}_n f_{\boldsymbol{\alpha}}(\mathbf{x}_n) \leq \theta\right) \quad \leq \quad \prod_{t=1}^{T} \exp\left\{\theta\alpha_t + \log Z_t\right\}$$

where Z_t is as in AdaBoost$_\varrho$.

Proposition Let γ_t be the edge of h_t in the t-th step of AdaBoost$_\varrho$. Assume $-1 \leq \varrho_t \leq \gamma_t$. Then holds for all $\theta \in [-1, 1]$

$$\exp\left\{\theta\alpha_t + \log Z_t\right\} \leq \exp\left(-\frac{1+\theta}{2}\log\left(\frac{1+\varrho_t}{1+\gamma_t}\right) - \frac{1-\theta}{2}\log\left(\frac{1-\varrho_t}{1-\gamma_t}\right)\right)$$

A Better Algorithm: AdaBoost*

Lemma If $\theta \leq \varrho_t \leq \gamma_t$:

$$\exp\left(-\frac{1+\theta}{2}\log\left(\frac{1+\varrho_t}{1+\gamma_t}\right)-\frac{1-\theta}{2}\log\left(\frac{1-\varrho_t}{1-\gamma_t}\right)\right)$$

$$\leq \exp\left(-\frac{1+\varrho_t}{2}\log\left(\frac{1+\varrho_t}{1+\gamma_t}\right)-\frac{1-\varrho_t}{2}\log\left(\frac{1-\varrho_t}{1-\gamma_t}\right)\right) \leq 1-\frac{1}{2}\frac{(\varrho_t-\gamma_t)^2}{1-\varrho_t{}^2}$$

- Would like to achieve $\theta = \varrho^* - \epsilon \quad \Rightarrow \quad \theta \leq \min_t \gamma_t - \epsilon$
- Set $\varrho_t = \min_t \gamma_t - \epsilon \quad \Rightarrow \quad \theta \leq \varrho_t \leq \gamma_t \quad \Rightarrow$ can apply Lemma

$$\frac{1}{N}\sum_{n=1}^{N} \mathrm{I}\left(\mathbf{y}_n f_{\boldsymbol{\alpha}}(\mathbf{x}_n) \leq \theta\right) \quad \leq \quad \prod_{t=1}^{T}\left(1-\frac{1}{2}\frac{(\varrho_t-\gamma_t)^2}{1-\varrho_t{}^2}\right) \leq \left(1-\frac{\epsilon^2}{2}\right)^T$$

- Hence, algorithm only needs $\dfrac{2\log(N)}{\epsilon^2}$ iterations (no log-term!)

Efficient Solution of Semi-infinite Problems

- Marginal AdaBoost and AdaBoost* can solve:

$$\max_{\boldsymbol{\alpha}\in\mathbf{R}^J} \rho \quad \text{with} \quad y_n \sum_{j=1}^{J} \alpha_j h_j(\mathbf{x}_n) \geq \rho \quad \text{for all} \quad n = 1,\ldots,N$$

  with $\|\boldsymbol{\alpha}\|_1 = 1$

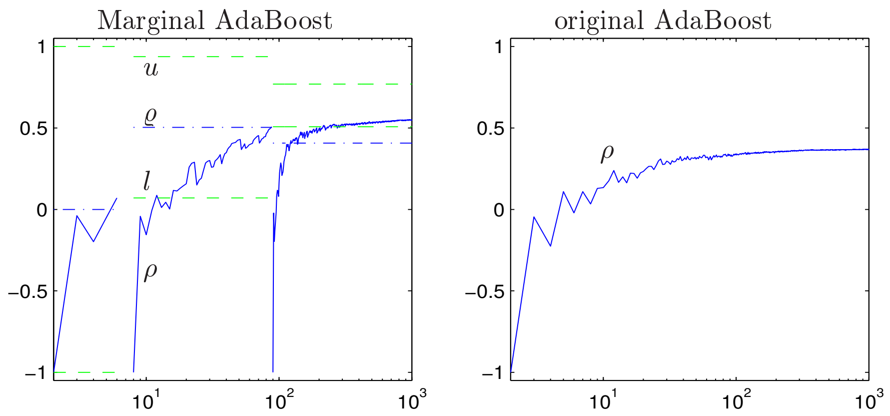
- Dual optimization Problem:

$$\min_{\mathbf{d}\in\mathbf{R}^N} \varepsilon \quad \text{with} \quad \sum_{n=1}^{N} d_n\, y_n h_j(\mathbf{x}_n) \leq \varepsilon \quad \text{for all} \quad h_j \in H$$

  and $\mathbf{d} \geq \mathbf{0}$, $\sum_{n=1}^{N} d_n = 1$

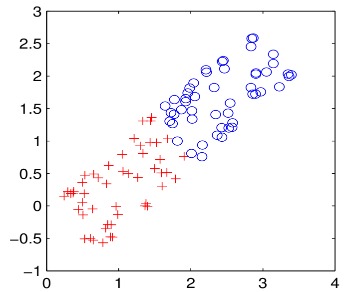
- if H has infinite number of elements, there are infinitely many variables/constraints
- AdaBoost works anyway and can solve the problem rather efficiently! (H needs to be compact)

Experimental Illustration



Does a Larger Margin Help?

- 100 dimensional data with two discriminative directions (other noise)
- C4.5, AdaBoost and Marginal AdaBoost on 200 realizations of data



	C4.5	AB	Marginal AB
E_{gen}	$7.4 \pm 0.11\%$	$4.0 \pm 0.11\%$	$\mathbf{3.6 \pm 0.10\%}$
ϱ	—	0.31 ± 0.01	$\mathbf{0.58 \pm 0.01}$
wins	1/200	59/200	**140/200**

- generalization error improves with larger margins
- this only holds for separable data ⇒ **Soft-Margins**!

Summary

- Developed a simple, yet effective technique for maximizing the margin:

  Marginal AdaBoost and **AdaBoost***

- Depends only logarithmically on sample size
- Works with infinitely many variables
- Observed empirical improvements on separable data
- Open problems:
  - Are better rates or constants possible? (second order version?)
  - How to show this for Soft-Margins?
  - Extend to probabilistic weak learners
  - Online version?