Boosting and Relationship to SVM

By

Bhaskara Reddy Poluru

poluru@informatik.uni-freiburg.de

Outline

Introduction
- AdaBoost
- Error Analysis
- Relationship to SVM
- Literature

Introduction

- A general method for improving the accuracy of any given learning algorithm
- constructs an ensemble of weak learners
- What is a weak learner?

Outline

- Introduction
  AdaBoost
- Error Analysis
- Relationship to SVM
- Literature

AdaBoost

Given: $(x_1, y_1), \ldots, (x_m, y_m)$ where $x_i \in X$, $y_i \in Y = \{-1, +1\}$
Initialize $D_1(i) = 1/m$.
For $t = 1, \ldots, T$:

- Train weak learner using distribution D_t.
- Get weak hypothesis $h_t : X \to \mathbb{R}$.
- Choose $\alpha_t \in \mathbb{R}$.
- Update:

(Where)

$$\epsilon_t = \Pr_{i \sim D_t}[h_t(x_i) \neq y_i] = \sum_{i: h_t(x_i) \neq y_i} D_t(i).$$

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1 - \epsilon_t}{\epsilon_t}\right).$$

$_i))$

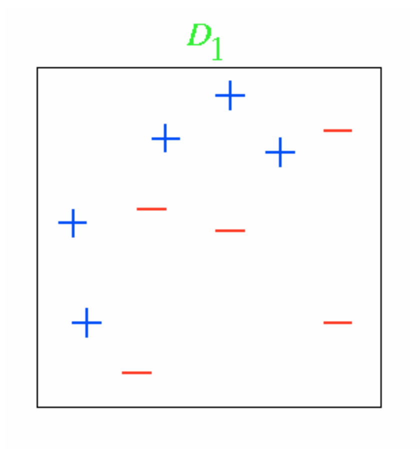
where Z_t is … D_{t+1} will be a distribu-
tion).

Output the fina…

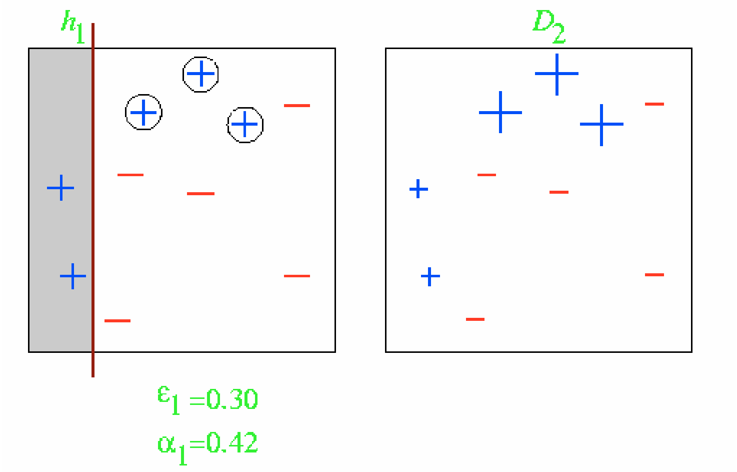
$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right).$$

Fig. 1. The boosting algorithm AdaBoost.

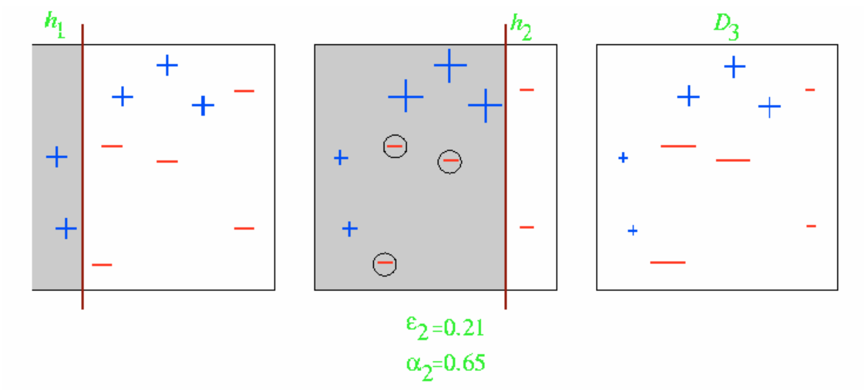
An Example on AdaBoost



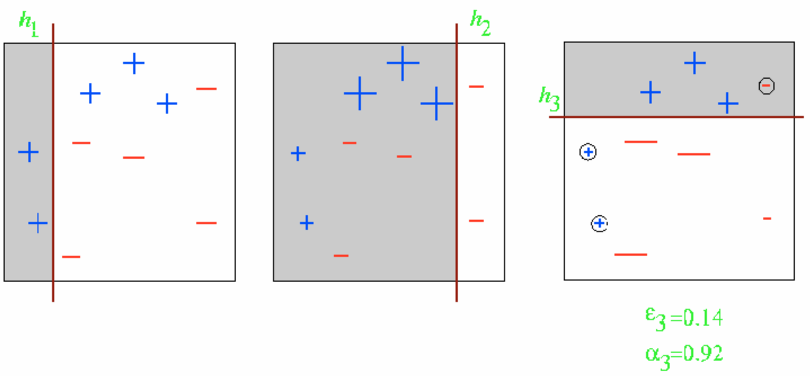
1st Round



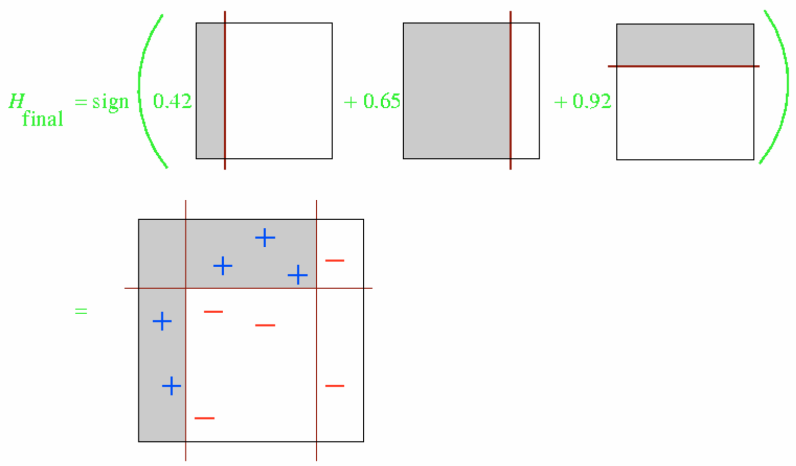
2nd Round



3rd Round



Linear Combination of the Hypotheses



Outline

- Introduction
- AdaBoost
  Error Analysis
- Relationship to SVM
- Literature

AdaBoost Training Error

(freund and shapire [23])

Let us write the error ϵ_t of h_t as: $\frac{1}{2} - \gamma_t$

The training error of the final hypothesis H is at most:

$$\prod_t \left[2\sqrt{\epsilon_t(1-\epsilon_t)}\right] = \prod_t \sqrt{1-4\gamma_t^2} \leq \exp\left(-2\sum_t \gamma_t^2\right)$$

Thus, if each weak hypothesis is slightly better than random so that $\gamma_t \geq \gamma$ for some $\gamma > 0$, then the training error drops exponentially fast.

AdaBoost is **adaptive** in the sense that it adapts to the error rates of particular weak learners

AdaBoost Training Error

- In the general case the training error of H is bounded by:

$$\frac{1}{m}|\{i : H(x_i) \neq y_i\}| \;\leq\; \frac{1}{m}\sum_i \exp(-y_i f(x_i)) \;=\; \prod_t Z_t$$

$$H(x) \;=\; \mathrm{sign}(f(x)).$$

- This equation suggests that the training error can be reduced most rapidly (in a greedy way) by choosing α_t and h_t in each round to minimize:

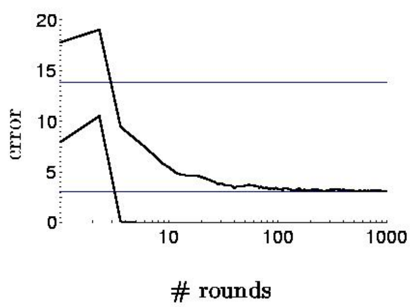
$$Z_t = \sum_i D_t(i)\exp(-\alpha_t y_i h_t(x_i)) \;=\; 2\sqrt{\epsilon_t(1-\epsilon_t)}$$

$$\alpha_t = \tfrac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right).$$

AdaBoost Generalization Error

- Training sample *S* of size *m*, VC-dimension *d and* With high probability the generalization is at most:

$$\hat{\Pr}\left[H(x) \neq y\right] + \tilde{O}\left(\sqrt{\frac{Td}{m}}\right)$$



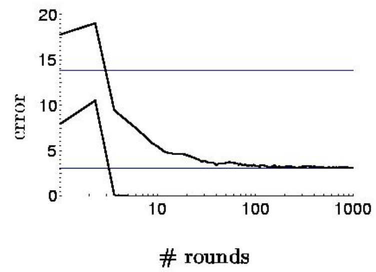
AdaBoost Generalization Error

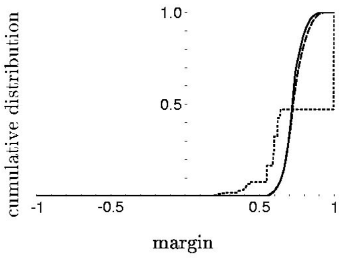
Margin of example (x,y) = $\frac{y \sum_t \alpha_t h_t(x)}{\sum_t |\alpha_t|}$.

New bound on the generalization error (independent of *T*):

$$\hat{\Pr}\left[\text{margin}_f(x,y) \le \theta\right] + \tilde{O}\left(\sqrt{\frac{d}{m\theta^2}}\right)$$

AdaBoost Generalization Error





Outline

- Introduction
- AdaBoost
- Error Analysis

  Relationship to SVM
- Literature

Comparison to SVM

- SVM works directly with margins, attempting to maximize the minimum margin of any training example (explicitly maximizes)
- AdaBoost tries to make the margins of all the training examples as large as possible (implicitly maximizes the minimum margin)

Comparison to SVM

(freund and shapire 1999)

let us denote the vector of weak-hypothesis predictions associated with the example (x, y) by $\mathbf{h}(x) \doteq \langle h_1(x), h_2(x), \ldots, h_N(x) \rangle$ which we call the *instance vector* and the vector of coefficients by $\boldsymbol{\alpha} \doteq \langle \alpha_1, \alpha_2, \ldots, \alpha_N \rangle$ which we call the *weight vector*. Using this notation and the definition of margin given in Eq. (2) we can write the goal of maximizing the minimum margin as

$$\max_{\boldsymbol{\alpha}} \min_i \frac{(\boldsymbol{\alpha} \cdot \mathbf{h}(x_i)) y_i}{||\boldsymbol{\alpha}|| \; ||\mathbf{h}(x_i)||} \qquad \text{(i)}$$

$$||\boldsymbol{\alpha}||_1 \doteq \sum_t |\alpha_t|, \qquad ||\mathbf{h}(x)||_\infty \doteq \max_t |h_t(x)| \, .$$

The explicit goal of SVM is also (i), where $h=\Phi$, and α= w :

$$||\boldsymbol{\alpha}||_2 \doteq \sqrt{\sum_t \alpha_t^2}, \qquad ||\mathbf{h}(x)||_2 \doteq \sqrt{\sum_t h_t(x)^2} \, .$$

Differences

- Both SVM and AdaBoost aim to find a linear combination in a high dimensional space But, the norms used to define the margin are different in the two cases and the precise goal is also different
- **Others:**
- **Different norms can result in very different margins**

  especially in high dimensional input spaces
- **The computational requirements are also different**

  SVM : Quadratic,

  AdaBoost : Linear

Differences

- **A different approach is used to search efficiently**

  SVM uses Kernels, while AdaBoost relies on a weak learning algorithm

  The re-weighting of the examples changes the distribution, guiding the weak learner to find different correlated coordinates

  Selecting the appropriate Kernel vs. Selecting an appropriate weak learning algorithm

Outline

- Introduction
- AdaBoost
- Error Analysis
- Relationship to SVM

  Literature

Related Papers

- A short introction to boosting
  Freund and shapire 1999

- boosting and SVM one class
  (Rätsch, Schölkopf, Mika & Müller, 2000)

- Marginal Boosting (maximizing the margin)
  (Rätsch & Warmuth, 2001)

Links

- http://www.boosting.org/tutorial.html
- http://www.cs.princeton.edu/~schapire/boost.html