Semi-Supervised MarginBoost

F. d'Alché-Buc, Yves Grandvalet, Chrisophe Abroise
Université P. et M. Curie
and
Université de Technologie de Compiègne

NIPS - December 2001

Related Papers

Experiments with a new boosting algorithm

- Yoav Freund and Robert E. Schapire
- *ICML 96*

Boosting the margin: A new explanation for the effectiveness of voting methods

- Robert E. Schapire and Yoav Freund and Peter Bartlett and Wee Sun Lee
- *ICML 97*

Boosting Algorithms as Gradient Descent

- Llew Mason, Jonathan Baxter, Peter Bartlett and Marcus Frean
- *NIPS 99*

Overview

- Discriminative Framework
- Extend the notion of margin to unlabeled data - they suggest two possibilities
- Modify MarginBoost to try to maximize the combined margin of the labeled and unlabeled data
- Implement and evaluate this technique

MarginBoost-L1Algorithm

Combined Classifier:

$$g_t(x) = \sum_{k=1}^{t} c_k . h_k(x)$$

Cost Function:

$$C(g_t) = \sum_{i=1}^{i=l} c(\rho(g_t(\mathbf{x}_i), y_i)))$$

with $\rho(g_t(\mathbf{x}_i), y_i) = y_i g_t(\mathbf{x}_i)$

Equivalent Cost Function:

$$J_t^L = \sum_{i \in S} w_t(i).y_i(h_{t+1}(\mathbf{x}_i) - g_t(\mathbf{x}_i))$$

with $w_t(i) = \dfrac{c'(\rho(g_t(\mathbf{x}_i), y_i))}{\sum_{j \in S} c'(\rho(g_t(\mathbf{x}_j), y_j))}$

MarginBoost-L1Algorithm

Let $w_0(i) = 1/l$, $i = 1, \ldots, l$.
Let $g_0(x) = 0$
For $t = 1 \ldots T$ (do the gradient descent):

1. Learn a gradient direction $h_{t+1} \in H$ with a high value of $J_t^S = \sum_{i \in S} w_t(i) y_i (h_{t+1}(\mathbf{x}_i) - g_t(\mathbf{x}_i))$
2. Apply the stopping rule : if $J_t^S \leq 0$ then return g_t else go on.
3. Choose a step-length in the obtained direction: $c_{t+1} = \epsilon$
4. Add the new direction to obtain $g_{t+1} = \frac{(|c_t| g_t + c_{t+1} h_{t+1})}{|c_{t+1}|}$
5. Fix the weight distribution : $w_{t+1} = \frac{c'(\rho(g_{t+1}(\mathbf{x}_i), y_i))}{\sum_{j \in S} c'(\rho(g_{t+1}(\mathbf{x}_j), y_j))}$

Margins for Unlabeled Data

Expected Margin:

$$\rho_U(g_t(x)) = \mathbf{E}_y(\rho_L(g_t(x), y))$$

Hypothesis as label posterior:

$$\rho_U{}^g(g_t(\mathbf{x})) = \frac{(g_t(\mathbf{x}) + 1)}{2} g_t(\mathbf{x}) + (1 - \frac{(g_t(\mathbf{x}) + 1)}{2})(-g_t(\mathbf{x}))$$

$$\rho_U{}^g(g_t(\mathbf{x})) = g_t(\mathbf{x})^2$$

Hypothesis as maximum a posterior label estimate:

$$\rho_U^s(g_t(\mathbf{x})) = g_t(\mathbf{x})\mathrm{sign}(g_t(\mathbf{x})) = |g_t(\mathbf{x})|$$

Boosting with Unlabeled Data

Cost Function:

$$C(g_t) = \sum_{i \in L} c(\rho_L(g_t(x_i), y_i) + \sum_{i \in U} c(\rho_U(g_t(x_i)))$$

Weight Computation:

$$w_t(i) = \begin{cases} \frac{c'(\rho_L(g_t(\mathbf{x}_i), y_i))}{|w_t|} & \text{if } i \in L \\ \frac{c'(\rho_U(g_t(\mathbf{x}_i)))}{|w_t|} & \text{if } i \in U \end{cases} \text{ with } |w_t| = \sum_{i \in S} w_t(i)$$

Boosting a Mixture Model

Mixture Model:

$$f(\mathbf{x}|\Phi) = \sum_{k=1}^{K} p_k f_k(\mathbf{x}; \theta_k)$$

Log Likelihood:

$$L(\Phi; \{\mathbf{x}_i, \mathbf{z}_i\}_{i=1}^{n}) = \sum_{i=1}^{n} w_t(i) \log\left(f(\mathbf{x}_i, \mathbf{z}_i; \Phi)\right)$$

EM for Gaussian Mixture Model

E-Step Compute the expectation of $L(\Phi; \{\mathbf{x}_i, \mathbf{z}_i\}_{i=1}^n)$ conditionally to $\{\mathbf{x}_i, \mathbf{z}_i\}_{i=1}^n$ and the current value of Φ (denoted Φ^q):

$$
\begin{aligned}
Q(\Phi|\Phi^q) &= \sum_{i=1}^{n} \sum_{k=1}^{K_1+K_2} w_t(i) u_{ik} \log\left(p_k f_k(\mathbf{x}_i; \theta_k)\right) \\
\text{with } u_{ik} &= \frac{z_{ik} p_k f_k(\mathbf{x}_i; \theta_k)}{\sum_\ell z_{i\ell} p_\ell f_\ell(\mathbf{x}_i; \theta_\ell)}
\end{aligned}
$$

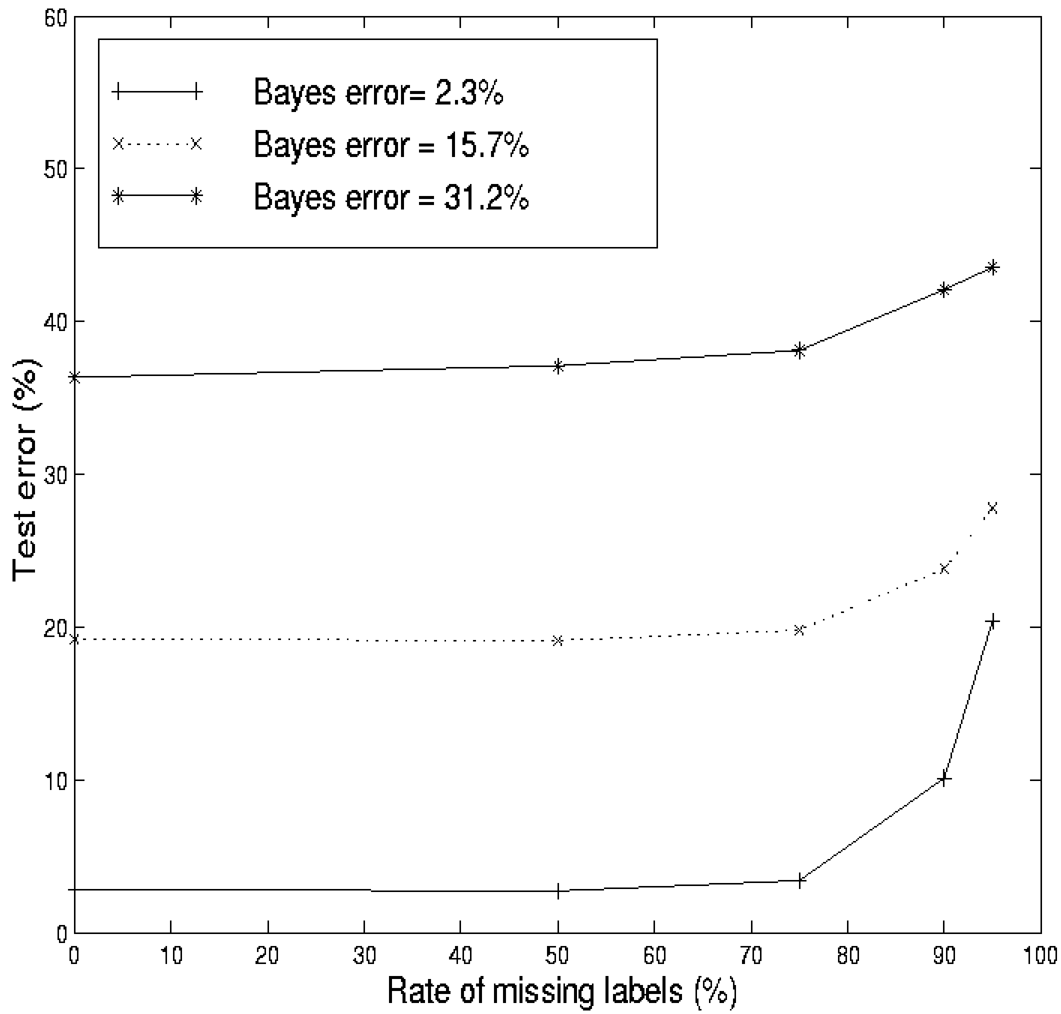
M-Step Maximize $Q(\Phi|\Phi^q)$ with respect to Φ.
Assuming that each mode k follows a normal distribution with mean $\boldsymbol{\mu}_k$, and covariance $\boldsymbol{\Sigma}_k$, $\Phi^{q+1} = \{\boldsymbol{\mu}_k^{q+1}; \boldsymbol{\Sigma}_k^{q+1}; p_k^{q+1}\}_{k=1}^{K_1+K_2}$ is given by:

$$
\begin{aligned}
p_k^{q+1} &= \frac{\sum_i w_t(i) u_{ik}}{\sum_i w_t(i)} \; ; \; \boldsymbol{\mu}_k^{q+1} = \sum_i \frac{w_t(i) u_{ik} \mathbf{x}_i}{\sum_i w_t(i) u_{ik}} \\
\boldsymbol{\Sigma}_k^{q+1} &= \frac{1}{\sum_i w_t(i) u_{ik}} \sum_{i=1}^{n} u_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k^{q+1})(\mathbf{x}_i - \boldsymbol{\mu}_k^{q+1})^t
\end{aligned}
$$

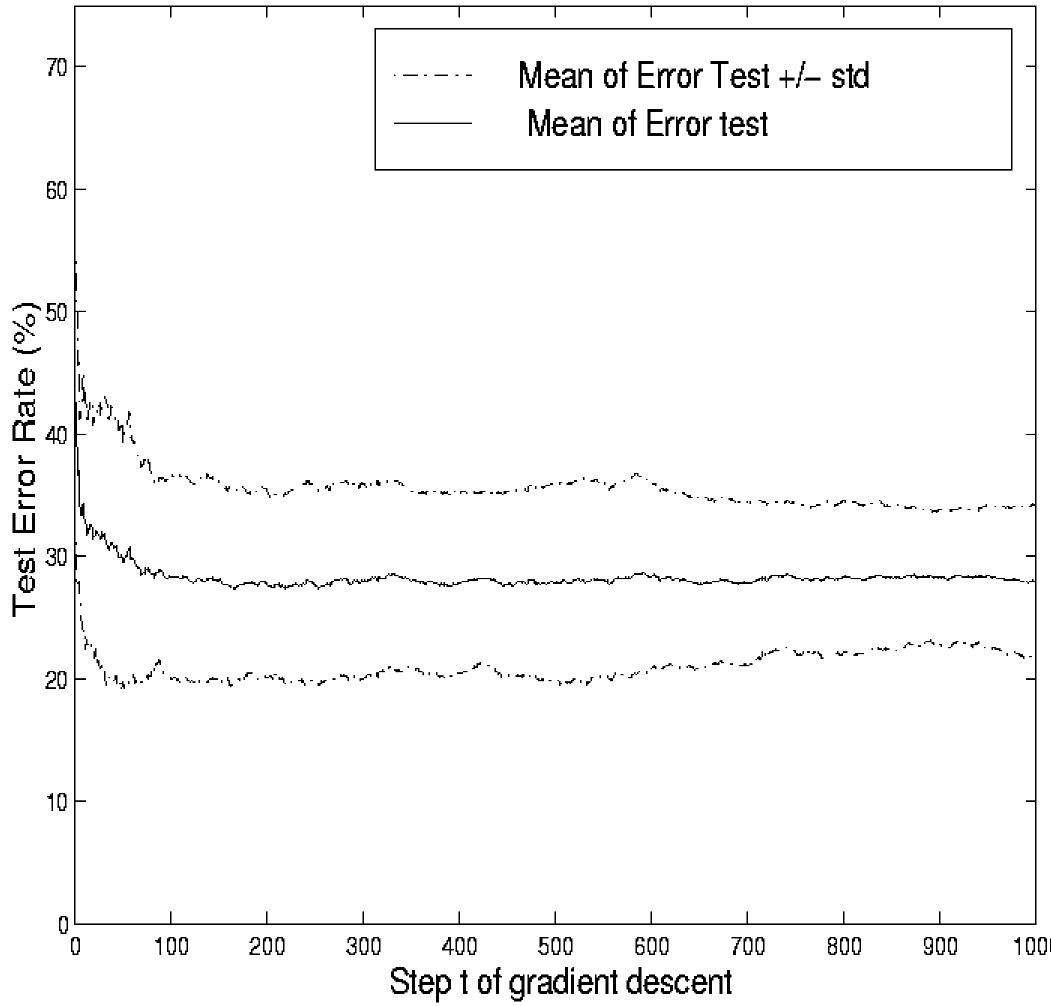
Experimental Results

Ringnorm	50%	75%	90%	95%
EM	2.1[1.7, 2.1]	4.3[1.9, 5.7]	9.5[2.7,12.0]	23.7[14.5,27.0]
AB	1.8[1.6, 2.0]	3.1[1.9, 4.1]	11.5[4.2,12.1]	28.7[11.5,37.6]
SSMBoost ρ^s	**1.7**[1.5, 1.8]	**2.0**[1.5, 2.4]	**3.7**[2.1, 4.8]	**6.9**[5.6,10.7]
SSMBoost ρ^g	**1.7**[1.6, 1.8]	**2.0**[1.4, 2.5]	4.5[2.2, 3.6]	8.1[4.2, 9.0]
Twonorm	**50%**	**75%**	**90%**	**95%**
EM	3.2[2.7, 3.1]	6.5[3.0, 9.0]	20.6[10.3,22.5]	24.8[18.3,31.9]
AB	3.2[2.9, 3.2]	**3.2**[3.0, 3.5]	11.0[5.2,14.2]	38.9[29.4,50.0]
SSMBoost ρ^s	**2.7**[2.5, 2.9]	3.4[2.8, 4.3]	**10.1**[5.8,13.6]	**20.4**[11.9,32.3]
SSMBoost ρ^g	**2.7**[2.5, 2.8]	3.4[2.8, 4.2]	11.0[5.6,16.2]	21.1[12.5,30.8]
Banana	**50%**	**75%**	**90%**	**95%**
EM	18.2[16.7,18.6]	21.8[18.0,25.0]	26.1[20.7,29.8]	31.7[23.8,35.8]
AB	**12.6**[11.7,13.1]	**15.2**[13.0,16.8]	**22.1**[18.0,24.3]	37.5[32.2,42.2]
SSMBoost ρ^s	13.3[12.7,14.3]	17.0[15.3,17.8]	22.2[18.0,28.0]	**28.3**[20.2,35.2]
SSMBoost ρ^g	13.3[12.8,14.2]	16.9[15.6,17.8]	22.8[18.3,29.3]	28.6[21.5,34.2]

Test Error vs. Missing Label Rate



Test Error vs. Iteration Count



Contributions

Pros:

- Extension of Margin notion to unlabeled data
- Reasonable empirical results

Cons:

- Neighborhood model is based on hypothesis class

Some Notes About Margins...

Basic Notation

- Input space $x \in X$
- Output space $y \in Y = \{-1,+1\}$
- Hypothesis $h \in H$
- Real-valued: $f : X \to \mathrm{R}$
- Training Set $S = \{(x_1, y_1), \ldots, (x_i, y_i), \ldots.\}$
- Test error ε
- Dot product $\langle x, z \rangle$

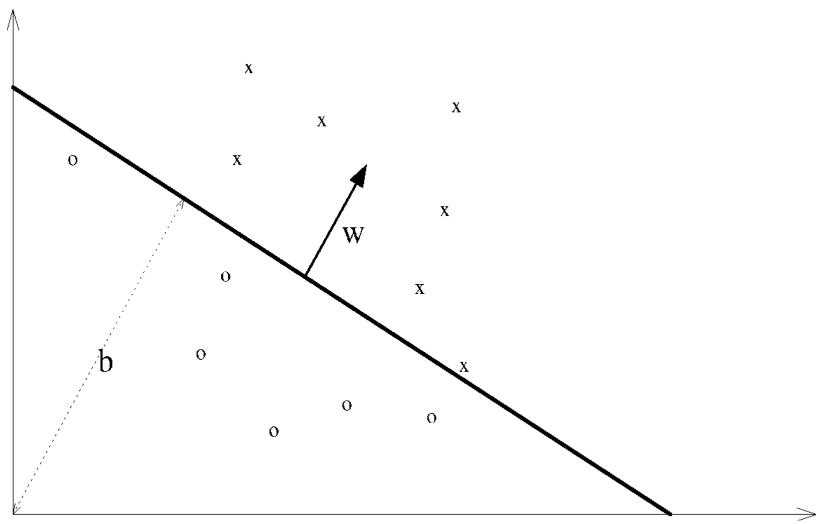
Linear Algebra

- Inner product between vectors

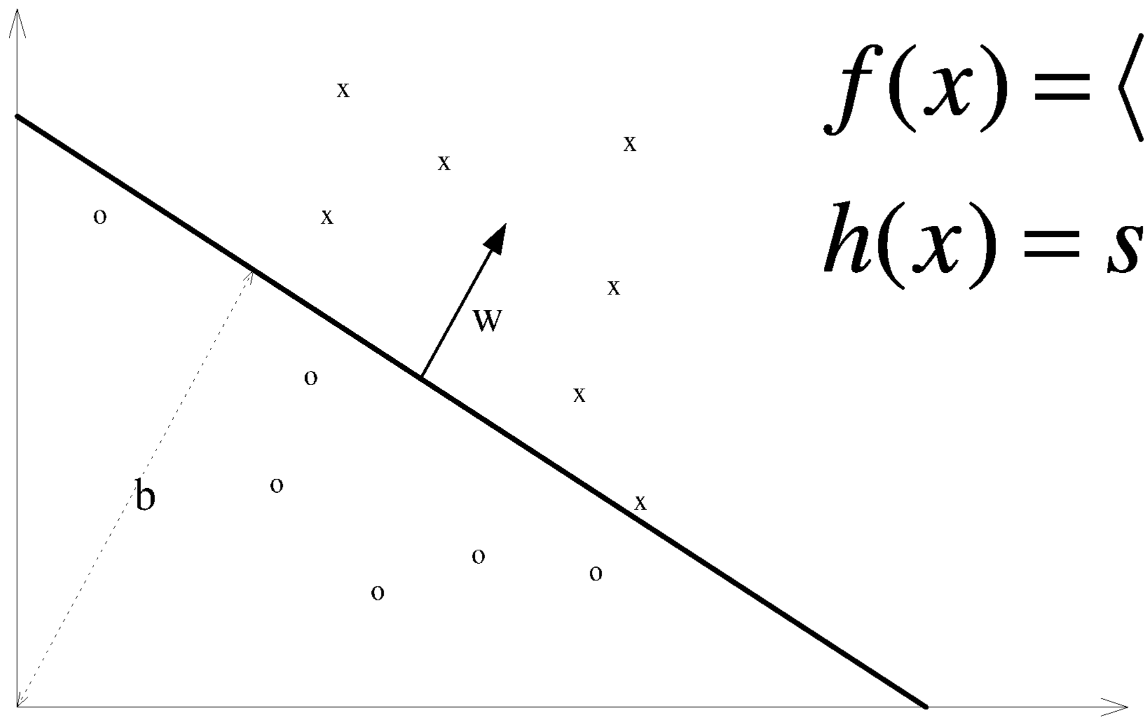
$$\langle \bar{x}, \bar{z} \rangle = \sum_i x_i z_i$$

- Hyperplane:

$$\langle w, x \rangle + b = 0$$



Linear Classifier



- Linear Separation of the input space

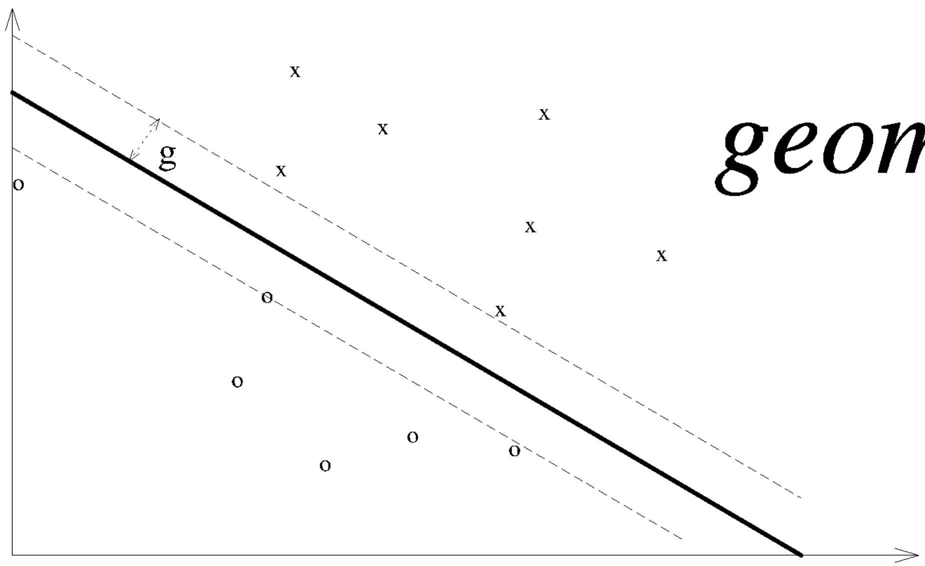
$$f(x) = \langle w, x \rangle + b$$

$$h(x) = sign(f(x))$$

Margin Definition

- Functional and geometric margin:

$$funct = y_i f(x_i)$$



$$geom = \frac{y_i f(x_i)}{\|f\|}$$

Definitions of Margin in the Papers

Margin Definition in MarginBoost Paper

The margin on labeled point i given classifier g_t is:

$$\rho(g_t(\mathbf{x}_i), y_i) = y_i g_t(\mathbf{x}_i)$$