ICML 2009 Tutorial
Survey of Boosting
from an Optimization Perspective

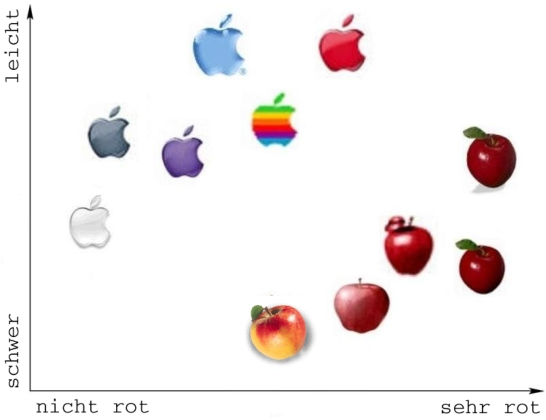
Part I: Entropy Regularized LPBoost
Part II: Boosting from an Optimization Perspective

Manfred K. Warmuth - UCSC
S.V.N. Vishwanathan - Purdue & Microsoft Research

Updated: August 15, 2009

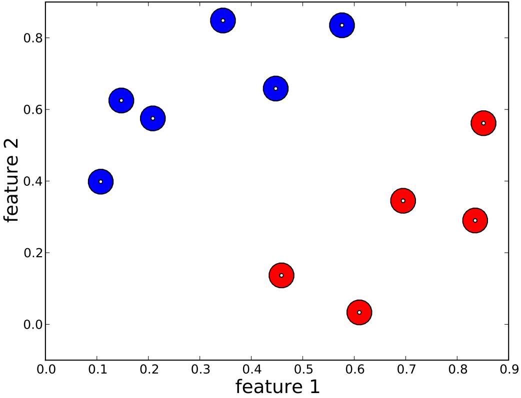
Outline

Setup for Boosting
[Giants of field: Schapire,Freund]



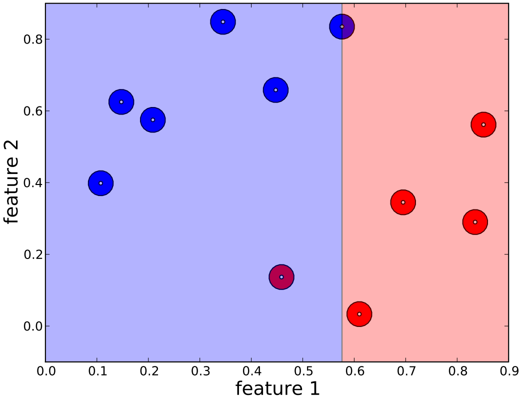
- examples: 11 apples
- -1 if artificial
  +1 if natural
- goal:
  classification

Setup for Boosting



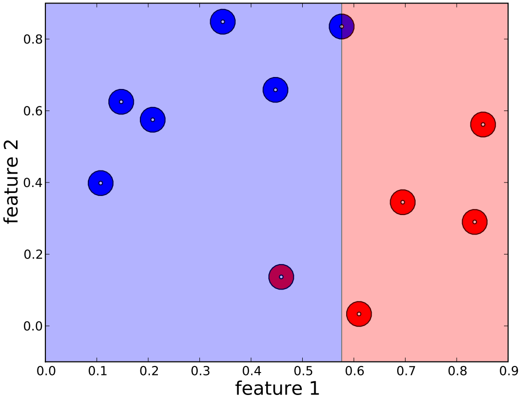
- -1/+1 examples
- weight $d_n \approx$ size

Weak hypotheses



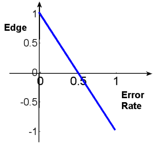
- weak hypotheses:
  decision stumps on two features
- goal:
  find convex combination of weak hypotheses that classifies all

Boosting: 1st iteration



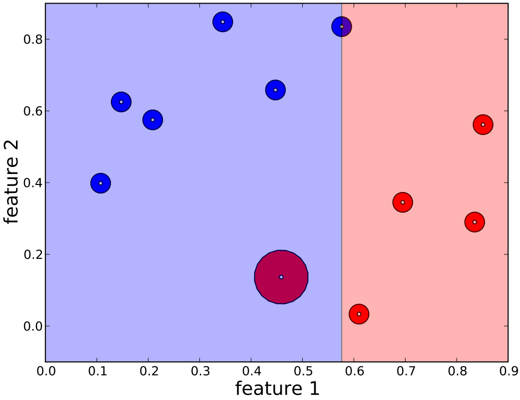
First hypothesis:

- error: $\frac{1}{11}$
- edge: $\frac{9}{11}$



edge $=$ $1 - 2$ error
low error $=$ high edge

Update after 1st



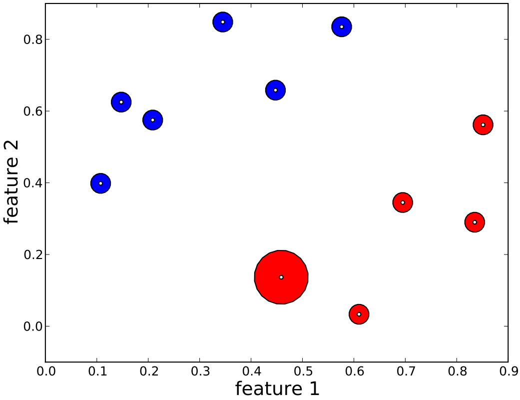
Misclassified examples

- increased weights

After update

- edge of hypothesis decreased

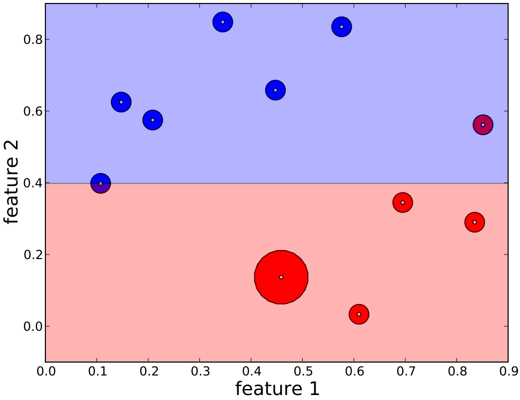
Before 2nd iteration



Hard examples

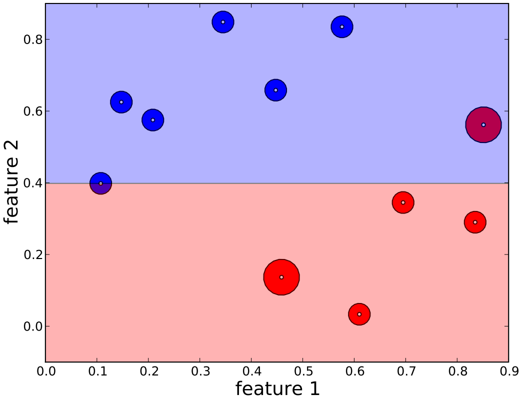
- high weight

Boosting: 2nd hypothesis



Pick hypotheses with high edge

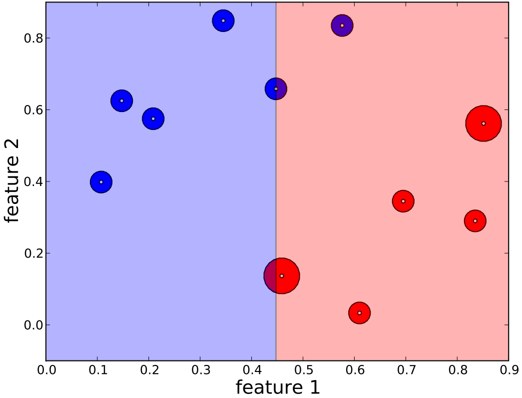
Update after 2nd



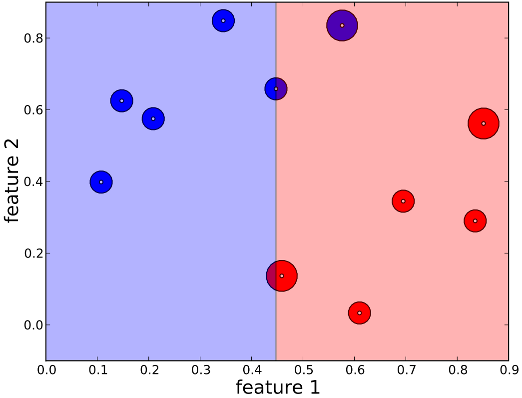
After update

- edges of all past hypotheses should be small

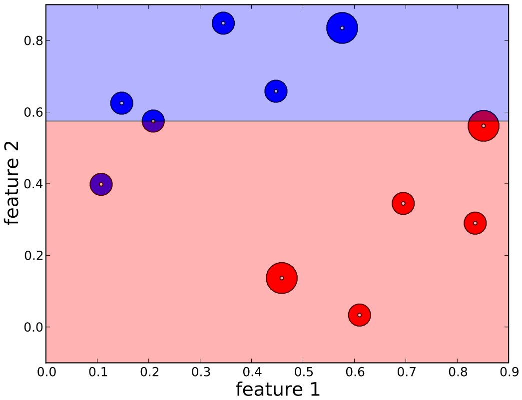
3rd hypothesis



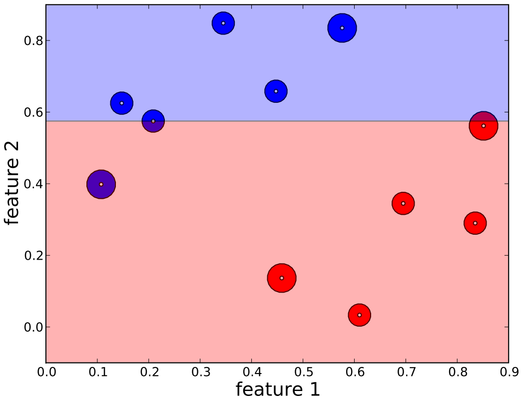
Update after 3rd



4th hypothesis

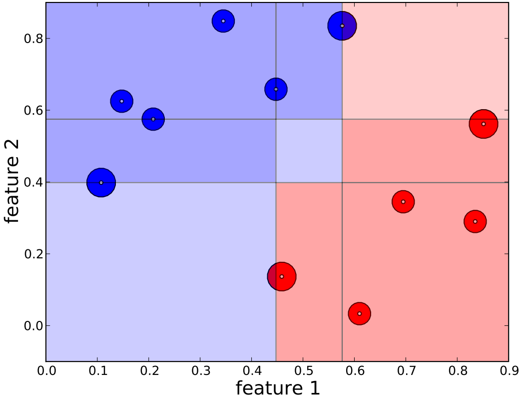


Update after 4th



Final convex combination of all hypotheses

Decision: $\sum_{t=1}^{T} w_t h^t(\mathbf{x}) \geq 0$?



Positive total weight - Negative total weight

Protocol of Boosting [FS97]

- Maintain distribution on N ± 1 labeled examples
- At iteration $t = 1, \ldots, T$:
  - Receive "weak" hypothesis h^t of high edge
  - Update $\mathbf{d}^{t-1}$ to $\mathbf{d}^t$ more weights on "hard" examples
- Output convex combination of the weak hypotheses $\sum_{t=1}^{T} w_t \, h^t(x)$

Two sets of weights:
- distribution on $\mathbf{d}$ on examples
- distribution on $\mathbf{w}$ on hypotheses

Edge vs. margin [Br99]

Edge of a hypothesis h for a distribution $\mathbf{d}$ on the examples

$$\underbrace{\sum_{n=1}^{N} \overbrace{y_n h(\mathbf{x}_n)}^{\text{goodness of example}} d_n}_{\text{average goodness of hypothesis}} \qquad \mathbf{d} \in \mathcal{P}^N$$

Margin of example n for current hypothesis weighting $\mathbf{w}$

$$\underbrace{\sum_{t=1}^{T} \overbrace{y_n h^t(\mathbf{x}_n)}^{\text{goodness of example}} w_t}_{\text{average goodness of example}} \qquad \mathbf{w} \in \mathcal{P}^T$$

Edge vs. margin [Br99]

Edge of a hypothesis h for a distribution **d** on the examples

$$\underbrace{\sum_{n=1}^{N} \overbrace{y_n h(\mathbf{x}_n)}^{\text{goodness of example}} \quad d_n}_{\text{average goodness of hypothesis}} \qquad \mathbf{d} \in \mathcal{P}^N$$

Margin of example n for current hypothesis weighting **w**

$$\underbrace{\sum_{t=1}^{T} \overbrace{y_n h^t(\mathbf{x}_n)}^{\text{goodness of example}} \quad w_t}_{\text{average goodness of example}} \qquad \mathbf{w} \in \mathcal{P}^T$$

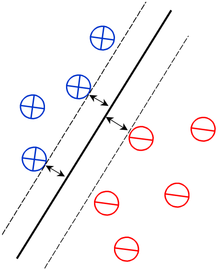
Objectives

Edge

- Edges of past hypotheses should be small after update
- Minimize maximum edge of past hypotheses

Margin

- Choose convex combination of weak hypotheses that maximizes the minimum margin



	Which margin?
SVM	2-norm (weights on examples)
Boosting	1-norm (weights on base hypotheses)

Connection between objectives?

Edge vs. margin

$$\text{min max edge} \quad = \quad \text{max min margin}$$

$$\min_{\mathbf{d}\in\mathcal{S}^N} \max_{q=1,2,\ldots,t-1} \underbrace{\sum_{n=1}^{N} y_n h^q(x_n) d_n}_{\text{edge of hypothesis q}} \quad = \quad \max_{\mathbf{w}\in\mathcal{S}^{t-1}} \min_{n=1,2,\ldots,N} \underbrace{\sum_{q=1}^{t-1} y_n h^q(x_n)\, w_q}_{\text{margin of example } n}$$

Linear Programming duality

Boosting as zero-sum-game [FS97]

Rock, Paper, Scissors game

column player

			R	P	S
			w_1	w_2	w_3
row player	R	d_1	0	1	-1
	P	d_2	-1	0	1
	S	d_3	1	-1	0

gain matrix

Single row is pure strategy of row player and $\mathbf{d}$ is mixed strategy

Single column is pure strategy of column player and $\mathbf{w}$ is mixed strategy

Row player minimizes
Column player maximizes

$$\text{payoff} = \mathbf{d}^T \mathbf{U}\, \mathbf{w} = \sum_{i,j} d_i U_{i,j} \mathbf{w}_j$$

Optimum strategy

$$
\begin{array}{ccc|ccc}
 & & & R & P & S \\
 & & & w_1 & w_2 & w_3 \\
 & & & .33 & .33 & .33 \\
R & d_1 & .33 & 0 & 1 & -1 \\
P & d_2 & .33 & -1 & 0 & 1 \\
S & d_3 & .33 & 1 & -1 & 0
\end{array}
$$

- Min-max theorem:

$$
\begin{aligned}
& \min_{d} \max_{\mathbf{w}} \mathbf{d}^T \mathbf{U}\mathbf{w} = \min_{d} \max_{j} \mathbf{d}^T \ \mathbf{U} e_j \\
= \quad & \max_{\mathbf{w}} \min_{d} \ \mathbf{d}^T \mathbf{U}\mathbf{w} = \max_{\mathbf{w}} \min_{i} \ e_i^\top \mathbf{U}\mathbf{w} \\
= \quad & \text{value of the game (0 in example)}
\end{aligned}
$$

e_j is pure strategy

Connection to Boosting?

- Rows are the examples
- Columns the weak hypothesis
- $U_{i,j} = h^j(\mathbf{x}_i)y_i$
- Row sum: margin of example
- Column sum: edge of weak hypothesis
- Value of game:

$$\text{min max edge} = \text{max min margin}$$

Van Neumann's Minimax Theorem

Weak hypothesis = column of game matrix $\mathbf{U}$

examples x_n	labels y_n	1st stump $h^1(x_n)$	$U_{*,1} = \mathbf{u}_1$
	-1	-1	$\mathbf{1}$
	-1	-1	$\mathbf{1}$
	-1	-1	$\mathbf{1}$
	-1	1	$-\mathbf{1}$
	1	1	$\mathbf{1}$
	1	1	$\mathbf{1}$
	1	1	$\mathbf{1}$
	1	-1	$-\mathbf{1}$

Edges/margins

			R	P	S		
			w_1	w_2	w_3	margin	
			.33	.33	.33		
R	d_1	.33	0	1	1	0	
P	d_2	.33	-1	0	1	0	min
S	d_3	.33	1	-1	-1	0	
	edge		0	0	0		
				max			

Value of game 0

New column added: boosting

			R	P	S		
			w_1	w_2	w_3	w_4	margin
			.44	0	.22	.33	
R	d_1	.22	0	1	-1	1	.11
P	d_2	.33	-1	0	1	1	.11 min
S	d_3	.44	1	-1	0	-1	.11
	edge		.11	-.22	.11	.11	
				max			

Value of game **increases** from 0 to .11

Row added: on-line learning

			R	P	S		
			w_1	w_2	w_3	margin	
			.33	.44	.22		
R	d_1	0	0	1	-1	.22	
P	d_2	.22	-1	0	1	-.11	min
S	d_3	.44	1	-1	0	-.11	
	d_4	.33	-1	1	-1	-.11	
	edge		-.11	-.11	-.11		
				max			

Value of game **decreases** from 0 to -.11

Boosting: maximize margin incrementally

$$\begin{array}{c|c} & w_1^1 \\ d_1^1 & 0 \\ d_2^1 & 1 \\ d_3^1 & -1 \end{array} \qquad \begin{array}{c|cc} & w_1^2 & w_2^2 \\ d_1^2 & 0 & -1 \\ d_2^2 & 1 & 0 \\ d_3^2 & -1 & 1 \end{array} \qquad \begin{array}{c|ccc} & w_1^3 & w_2^3 & w_3^3 \\ d_1^3 & 0 & -1 & 1 \\ d_2^3 & 1 & 0 & -1 \\ d_3^3 & -1 & 1 & 0 \end{array}$$

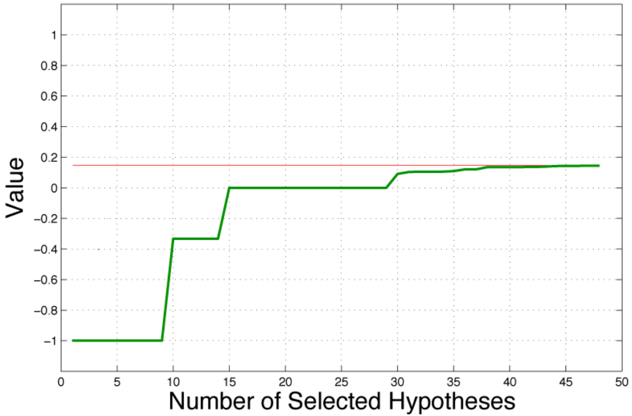
iteration 1   iteration 2   iteration 3

- In each iteration solve optimization problem to update **d**
- Column player / oracle provides new hypothesis
- Boosting is column generation method in **d** domain
  and coordinate/gradient descent in **w** domain

Outline

Boosting = greedy method for increasing margin

Converges to optimum margin w.r.t. all hypotheses



Want small number of iterations

Assumption on next weak hypothesis

For current weighting of examples,
oracle returns hypothesis of edge $\geq g$

Goal

- For given ϵ, produce convex combination of weak hypotheses with margin $\geq g - \epsilon$
- Number of iterations $O(\frac{\log N}{\epsilon^2})$

Min max thm for the inseparable case

Slack variables in **w** domain = capping in **d** domain

$$\max_{\mathbf{w}\in\mathcal{S}^t,\boldsymbol{\psi}\geq\mathbf{0}} \min_{n=1,2,\ldots,N} \underbrace{\left(\sum_{q=1}^{t} u_n^q\, w_q + \psi_n\right)}_{\text{margin of example } n} - \frac{1}{\nu}\sum_{n=1}^{N}\psi_n$$

$$= \min_{\mathbf{d}\in\mathcal{S}^N,\mathbf{d}\leq\frac{1}{\nu}\mathbf{1}} \max_{q=1,2,\ldots,t} \underbrace{\mathbf{u}^q\cdot\mathbf{d}}_{\text{edge of hypothesis q}}$$

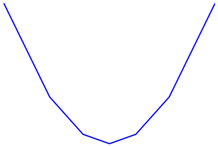
Notation: $u_n^q = y_n h^q(x_n)$

Outline

LPBoost [GS98,RSS+00,DBST02]

Choose distribution that minimizes the maximum edge via LP

$$\min_{\sum_n d_n=1, \mathbf{d} \leq \frac{1}{\nu}\mathbf{1}} \underbrace{\max_{q=1,2,\ldots,t} \mathbf{u}^q \cdot \mathbf{d}}_{f(\mathbf{d})}$$



- All weight is put on examples with minimum soft margin
- Brittle: iteration bound can be linear in N on carefully constructed artificial data sets [WGR07]

LPBoost may require $\Omega(N)$ iterations

		α_1	α_2	α_3	α_4	α_5	margin
		0	0	0	0	0	
d_1	.125	+1	-.95	-.93	-.91	-.99	–
d_2	.125	+1	-.95	-.93	-.91	-.99	–
d_3	.125	+1	-.95	-.93	-.91	-.99	–
d_4	.125	+1	-.95	-.93	-.91	-.99	–
d_5	.125	-.98	+1	-.93	-.91	+.99	–
d_6	.125	-.97	-.96	+1	-.91	+.99	–
d_7	.125	-.97	-.95	-.94	+1	+.99	–
d_8	.125	-.97	-.95	-.93	-.92	+.99	–
edge		.0137	-.7075	-.6900	-.6725	.0000	
value	**-1**						

LPBoost may require $\Omega(N)$ iterations

		α_1	α_2	α_3	α_4	α_5	margin
		1	0	0	0	0	
d_1	0	+1	-.95	-.93	-.91	-.99	1
d_2	0	+1	-.95	-.93	-.91	-.99	1
d_3	0	+1	-.95	-.93	-.91	-.99	1
d_4	0	+1	-.95	-.93	-.91	-.99	1
d_5	1	-.98	+1	-.93	-.91	+.99	-.98
d_6	0	-.97	-.96	+1	-.91	+.99	-.97
d_7	0	-.97	-.95	-.94	+1	+.99	-.97
d_8	0	-.97	-.95	-.93	-.92	+.99	-.97
edge		-.98	1	-.93	-.91	.99	
value	-1	-.98					

LPBoost may require $\Omega(N)$ iterations

		α_1	α_2	α_3	α_4	α_5	margin
		0	1	0	0	0	
d_1	0	+1	-.95	-.93	-.91	-.99	-.95
d_2	0	+1	-.95	-.93	-.91	-.99	-.95
d_3	0	+1	-.95	-.93	-.91	-.99	-.95
d_4	0	+1	-.95	-.93	-.91	-.99	-.95
d_5	0	-.98	+1	-.93	-.91	+.99	1
d_6	1	-.97	-.96	+1	-.91	+.99	-.96
d_7	0	-.97	-.95	-.94	+1	+.99	-.95
d_8	0	-.97	-.95	-.93	-.92	+.99	-.95
edge		-.97	-.96	1	-.91	.99	
value	-1	-.98	-.96				

LPBoost may require $\Omega(N)$ iterations

		α_1	α_2	α_3	α_4	α_5	margin
		0	0	1	0	0	
d_1	0	+1	-.95	-.93	-.91	-.99	-.93
d_2	0	+1	-.95	-.93	-.91	-.99	-.93
d_3	0	+1	-.95	-.93	-.91	-.99	-.93
d_4	0	+1	-.95	-.93	-.91	-.99	-.93
d_5	0	-.98	+1	-.93	-.91	+.99	-.93
d_6	0	-.97	-.96	+1	-.91	+.99	1
d_7	1	-.97	-.95	-.94	+1	+.99	-.94
d_8	0	-.97	-.95	-.93	-.92	+.99	-.93
edge		-.97	-.95	-.94	1	.99	
value	-1	-.98	-.96	-.94			

LPBoost may require $\Omega(N)$ iterations

		α_1	α_2	α_3	α_4	α_5	margin
		0	0	0	1	0	
d_1	0	+1	-.95	-.93	-.91	-.99	-.91
d_2	0	+1	-.95	-.93	-.91	-.99	-.91
d_3	0	+1	-.95	-.93	-.91	-.99	-.91
d_4	0	+1	-.95	-.93	-.91	-.99	-.91
d_5	0	-.98	+1	-.93	-.91	+.99	-.91
d_6	0	-.97	-.96	+1	-.91	+.99	-.91
d_7	0	-.97	-.95	-.94	+1	+.99	1
d_8	1	-.97	-.95	-.93	-.92	+.99	-.92
edge		-.97	-.95	-.94	-.92	.99	
value	-1	-.98	-.96	-.94	-.92		

LPBoost may require $\Omega(N)$ iterations

		α_1	α_2	α_3	α_4	α_5	margin
		.5	.0026	0	0	.4975	
d_1	0.4974	+1	-.95	-.93	-.91	-.99	.0051
d_2	0	+1	-.95	-.93	-.91	-.99	.0051
d_3	0	+1	-.95	-.93	-.91	-.99	.0051
d_4	0	+1	-.95	-.93	-.91	-.99	.0051
d_5	0	-.98	+1	-.93	-.91	+.99	.0051
d_6	.4898	-.97	-.96	+1	-.91	+.99	.0051
d_7	0	-.97	-.95	-.94	+1	+.99	.0051
d_8	.0127	-.97	-.95	-.93	-.92	+.99	.0051
edge		.0051	.0051	.9055	.9100	.0051	
value	-1	-.98	-.96	-.94	-.92	.0051	

Outline

Entropy Regularized LPBoost

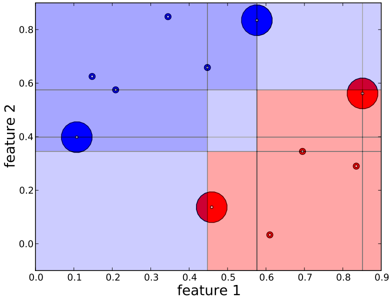
$$\min_{\sum_n d_n = 1, \mathbf{d} \le \frac{1}{\nu}\mathbf{1}} \max_{q=1,2,\ldots,t} \mathbf{u}^q \cdot \mathbf{d} + \frac{1}{\eta}\Delta(\mathbf{d}, \mathbf{d}^0)$$

- $$\mathbf{d}_n = \frac{\exp^{-\eta \text{ soft margin of example } n}}{Z} \qquad \text{"soft min"}$$

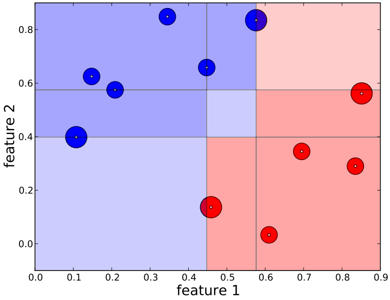
- Form of weights first in ν-Arc algorithm [RSS+00]
- Regularization in $\mathbf{d}$ domain makes problem strongly convex
- Gradient of dual Lipschitz continuous in $\mathbf{w}$ [e.g. HL93,RW97]

The effect of entropy regularization

Different distribution on the examples



LPBoost: lots of zeros / brittle



ERLPBoost: smoother

Outline

AdaBoost [FS97]

$$d_n^t := \frac{d_n^{t-1} \exp(-w_t u_n^t)}{\sum_{n'} d_{n'}^{t-1} \exp(-w_t u_{n'}^t)},$$

where w_t s.t. $\sum_{n'} d_{n'}^{t-1} \exp(-w\ u_{n'}^t)$ is minimized

- Easy to implement
- Adjusts distribution so that edge of last hypothesis is zero
- Gets within half of the optimal hard margin [RSD07]
  but only in the limit

Corrective versus totally corrective

Processing last hypothesis versus all past hypotheses

Corrective	Totally Corrective
AdaBoost	LPBoost
LogitBoost	TotalBoost
AdaBoost*	SoftBoost
SS,Colt08	ERLPBoost

From AdaBoost to ERLPBoost

AdaBoost (as interpreted in [KW99,La99])

Primal:

$$\min_{\mathbf{d}} \quad \Delta(\mathbf{d}, \mathbf{d}^{t-1})$$
$$\text{s.t.} \quad \mathbf{d} \cdot \mathbf{u}^{t-1} = 0, \; \|\mathbf{d}\|_1 = 1$$

Dual:

$$\max_{\mathbf{w}} \quad -ln \sum_n d_n^{t-1} \exp(u_n^{t-1} w_{t-1})$$
$$\text{s.t.} \quad \mathbf{w} \geq 0$$

Achieves half of optimum hard margin in the limit

AdaBoost* [RW05]

Primal:

$$\min_{\mathbf{d}} \quad \Delta(\mathbf{d}, \mathbf{d}^{t-1})$$
$$\text{s.t.} \quad \mathbf{d} \cdot \mathbf{u}^{t-1} \leq \gamma_{t-1},$$
$$\|\mathbf{d}\|_1 = 1$$

Dual:

$$\max_{\mathbf{w}} \; -ln \sum_n d_n^{t-1} \exp(u_n^{t-1} w_{t-1})$$
$$-\gamma_{t-1} ||\mathbf{w}||_1$$
$$\text{s.t.} \quad \mathbf{w} \geq 0$$

where edge bound γ_t is adjusted downward by a heuristic

Good iteration bound for reaching optimum hard margin

SoftBoost [WGR07]

Primal:

$$\min_{\mathbf{d}} \quad \Delta(\mathbf{d}, \mathbf{d}^0)$$
$$\text{s.t.} \quad \|\mathbf{d}\|_1 = 1,\ \mathbf{d} \le \tfrac{1}{\nu}\mathbf{1}$$
$$\mathbf{d} \cdot \mathbf{u}^q \le \gamma_{t-1},$$
$$1 \le q \le t-1$$

Dual:

$$\min_{\mathbf{w},\boldsymbol{\psi}} \quad -\ln \sum_n \mathbf{d}_n^0 \exp(-\eta \sum_{q=1}^{t-1} u_n^q w_q - \eta\psi_n) - \tfrac{1}{\nu}\|\boldsymbol{\psi}\|_1 - \gamma_{t-1}\|\mathbf{w}\|_1$$
$$\text{s.t.} \quad \mathbf{w} \ge 0,\ \boldsymbol{\psi} \ge 0$$

where edge bound γ_{t-1} is adjusted downward by a heuristic

Good iteration bound for reaching soft margin

ERLPBoost [WGV08]

Primal:

$$\min_{\mathbf{d},\gamma} \quad \gamma + \tfrac{1}{\eta}\Delta(\mathbf{d}, \mathbf{d}^0)$$
$$\text{s.t.} \quad \|\mathbf{d}\|_1 = 1,\ \mathbf{d} \le \tfrac{1}{\nu}\mathbf{1}$$
$$\mathbf{d} \cdot \mathbf{u}^q \le \gamma,$$
$$1 \le q \le t-1$$

Dual:

$$\min_{\mathbf{w},\boldsymbol{\psi}} \quad -\tfrac{1}{\eta} \ln \sum_n \mathbf{d}_n^0 \exp(-\eta \sum_{q=1}^{t-1} u_n^q w_q - \eta\psi_n) - \tfrac{1}{\nu}\|\boldsymbol{\psi}\|_1$$
$$\text{s.t.} \quad \mathbf{w} \ge 0,\ \|\mathbf{w}\|_1 = 1,\ \boldsymbol{\psi} \ge 0$$

where for the iteration bound η is fixed to $\max(\frac{2}{\epsilon} \ln \frac{N}{\nu}, \frac{1}{2})$

Good iteration bound for reaching soft margin

Iteration bounds

Corrective	Totally Corrective
AdaBoost	LPBoost
LogitBoost	TotalBoost
AdaBoost*	SoftBoost
SS,Colt08	ERLPBoost

- Strong oracle: returns hypothesis with maximum edge
- Weak oracle: returns hypothesis with edge $\geq g$

- In $O(\frac{\log \frac{N}{\nu}}{\epsilon^2})$ iterations within ϵ of maximum soft margin for strong oracle or within ϵ of g for weak oracle
- Ditto for hard margin case
- In $O(\frac{\log N}{g^2})$ iterations consistency with weak oracle

Synopsis

- LPBoost often unstable
- For safety, add relative entropy regularization
- Corrective algs
  - Sometimes easy to code
  - Fast per iteration
- Totally corrective algs
  - Smaller number of iterations
  - Nevertheless faster overall time
- Weak versus strong oracle makes a big difference in practice

$O(\frac{\log N}{\epsilon^2})$ iteration bounds

Good

- Bound is major design tool
- Any reasonable Boosting algorithm should have this bound

Bad

- Bound is weak

	$\frac{\ln N}{\epsilon^2} \geq N$
$\epsilon = .01$	$N \leq 1.2 \times 10^5$
$\epsilon = .001$	$N \leq 1.7 \times 10^7$

- Why are totally corrective algorithms much better in practice?

Lower bounds on the number of iterations

- Majority of $\Omega(\frac{\log N}{g^2})$ hypotheses for achieving consistency with weak oracle of guarantee g [Fr95]
- Later: $\Omega(\frac{1}{\epsilon^2})$ iteration bound for getting within ϵ of hard margin with strong oracle

Outline

Conclusion

- Adding relative entropy regularization of LPBoost
  leads to good boosting alg.
- Boosting is instantiation of MaxEnt and MinxEnt principles
  [Jaines 57,Kullback 59]
- Relative entropy regularization
  smoothens one-norm regularization

Open

- When hypotheses have one-sided error then
  $O(\frac{\log N}{\epsilon})$ iterations suffice [As00,HW03]
  Does ERLPBoost have $O(\frac{\log N}{\epsilon})$ bound when hypotheses one-sided?
- Strengthen general lower bound to $\Omega(\frac{\log N}{\epsilon^2})$
- Compare ours with Freund's algorithms that don't just cap, but forget examples

Acknowledgement

- Rob Schapire and Yoav Freund for pioneering Boosting
- Gunnar Rätsch for bringing in optimization
- Karen Glocer for helping with figures and plots