COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Pannaga Shivaswamy

Tony Jebara

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Empirical Bernstein Boosting

Pannaga Shivaswamy
Tony Jebara

empirical risk minimization

empirical risk minimization

- at the core of most machine learning algorithms

empirical risk minimization

- at the core of most machine learning algorithms
- examples
  - exponential loss : AdaBoost
  - hinge loss : SVM
  - squared loss, absolute loss: regression

empirical risk minimization

- at the core of most machine learning algorithms
- examples
  - exponential loss : AdaBoost
  - hinge loss : SVM
  - squared loss, absolute loss: regression
- minimize *mean loss* on training examples

empirical risk minimization

- at the core of most machine learning algorithms
- examples
  - exponential loss : AdaBoost
  - hinge loss : SVM
  - squared loss, absolute loss: regression
- minimize *mean loss* on training examples
- what about second order moment of the loss?

background

background

- Fisher linear discriminant
  - interclass distance, intraclass variance

background

- Fisher linear discriminant
  - interclass distance, intraclass variance
- second order perceptron (Cesa-Bianchi et al. ‘05)
  - update rule with whitening

background

- Fisher linear discriminant
  - interclass distance, intraclass variance
- second order perceptron (Cesa-Bianchi et al. ‘05)
  - update rule with whitening
- relative margin machines (Shivaswamy, Jebara ‘08)
  - margin with respect to spread

background

- Fisher linear discriminant
  - interclass distance, intraclass variance
- second order perceptron (Cesa-Bianchi et al. ‘05)
  - update rule with whitening
- relative margin machines (Shivaswamy, Jebara ‘08)
  - margin with respect to spread
- Gaussian margin machines (Crammer et al. ‘09)
  - PAC-Bayes bound minimization

background

- Fisher linear discriminant
  - interclass distance, intraclass variance
- second order perceptron (Cesa-Bianchi et al. ‘05)
  - update rule with whitening
- relative margin machines (Shivaswamy, Jebara ‘08)
  - margin with respect to spread
- Gaussian margin machines (Crammer et al. ‘09)
  - PAC-Bayes bound minimization
- confidence weighted learning (Crammer et al. ‘09)
  - online learning with first & second moments

Hoeffding's inequality

Hoeffding's inequality

- on a bounded random variable

$Z_1, \ldots, Z_n$ i.i.d. $\quad Z \in [0,1]$

with probability at least $1 - \delta$

$$\mathbf{E}[\quad Z \quad] \leq \frac{1}{n}\sum_{i=1}^{n} \quad Z_i \qquad + \sqrt{\frac{1}{2n}\ln(\ 1\ /\delta)}$$

Hoeffding's inequality

- on a bounded random variable
- on a bounded loss

$$(X_1, y_1), \ldots, (X_n, y_n) \text{ i.i.d. } \; l(f(X), y) \in [0, 1]$$

$$f : \mathcal{X} \rightarrow \mathbf{R}$$

with probability at least $1 - \delta$

$$\mathbf{E}[\; l(f(X), y) \;] \leq \frac{1}{n} \sum_{i=1}^{n} \; l(f(X_i), y_i) \; + \sqrt{\frac{1}{2n} \ln(\; 1 \; / \delta)}$$

Hoeffding's inequality

- on a bounded random variable
- on a bounded loss
- uniform convergence

$$(X_1, y_1), \ldots, (X_n, y_n) \text{ i.i.d. } \; l(f(X), y) \in [0,1]$$

$$f : \mathcal{X} \rightarrow \mathbf{R}$$

$$\text{with probability at least } 1 - \delta \quad \forall f \in \mathcal{F}$$

$$\mathbf{E}[\; l(f(X), y) \;] \leq \frac{1}{n} \sum_{i=1}^{n} \; l(f(X_i), y_i) \; + \sqrt{\frac{1}{2n} \ln(|\mathcal{F}|/\delta)}$$

Hoeffding's inequality

- on a bounded random variable
- on a bounded loss
- uniform convergence
- suggests ERM $\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} l(f(X_i), y_i)$

$(X_1, y_1), \ldots, (X_n, y_n)$ i.i.d. $\; l(f(X), y) \in [0, 1]$

$f : \mathcal{X} \to \mathbf{R}$

with probability at least $1 - \delta \quad \forall f \in \mathcal{F}$

$$\mathbf{E}[\, l(f(X), y) \,] \leq \frac{1}{n} \sum_{i=1}^{n} \; l(f(X_i), y_i) \; + \sqrt{\frac{1}{2n} \ln(|\mathcal{F}|/\delta)}$$

incorporating variance

incorporating variance

- Hoeffding's inequality

$$\mathbf{E}[Z] \leq \frac{1}{n}\sum_{i=1}^{n} Z_i + \sqrt{\frac{\ln(1/\delta)}{2n}}$$

incorporating variance

- Hoeffding's inequality
- Bernstein's inequality

$$\mathbf{E}[Z] \leq \frac{1}{n}\sum_{i=1}^{n} Z_i + \sqrt{\frac{4\mathbf{V}[Z]\ln(1/\delta)}{2n}} + \frac{\ln(1/\delta)}{3n}$$

$$\mathbf{V}[Z] = \mathbf{E}[Z - \mathbf{E}[Z]]^2$$

  - much tighter compared to Hoeffding's
  - limitation: true variance required

empirical Bernstein bound

empirical Bernstein bound

- Bernstein's inequality

$$\mathbf{E}[Z] \leq \frac{1}{n}\sum_{i=1}^{n} Z_i + \sqrt{\frac{2\mathbf{V}[Z]\ln(1/\delta)}{n}} + \frac{\ln(1/\delta)}{3n}$$

$$\mathbf{V}[Z] = \mathbf{E}[Z - \mathbf{E}[Z]]^2$$

- much tighter compared to Hoeffding's
- limitation: true variance required in equation

empirical Bernstein bound

- empirical Bernstein's inequality (Maurer & Pontil '09)

$$\mathbf{E}[Z] \leq \frac{1}{n}\sum_{i=1}^{n} Z_i + \sqrt{\frac{2\hat{\mathbf{V}}[\mathbf{Z}]\ln(2/\delta)}{n}} + \frac{7\ln(2/\delta)}{3(n-1)}$$

$$\hat{\mathbf{V}}[Z] = \frac{1}{n(n-1)} \sum_{1\leq i<j\leq n} (Z_i - Z_j)^2$$

- much tighter compared to Hoeffding's
- ~~limitation: true variance required in equation~~

empirical Bernstein bound

- empirical Bernstein's inequality (Maurer & Pontil '09)

$$\mathbf{E}[Z] \leq \frac{1}{n}\sum_{i=1}^{n} Z_i + \sqrt{\frac{2\hat{\mathbf{V}}[\mathbf{Z}]\ln(2/\delta)}{n}} + \frac{7\ln(2/\delta)}{3(n-1)}$$

$$\hat{\mathbf{V}}[Z] = \frac{1}{n(n-1)} \sum_{1\leq i<j\leq n} (Z_i - Z_j)^2$$

  - much tighter compared to Hoeffding's
  - ~~limitation: true variance required in equation~~

- suggests Sample Variance Penalization (SVP)

$$\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} l(f(X_i), y_i) + \lambda\sqrt{\hat{\mathbf{V}}[l(f(X), y)]}$$

SVP on 0-1 loss?

SVP on 0-1 loss?

- is SVP qualitatively different?

$$\hat{p} := \frac{1}{n}\sum_{i=1}^{n} l_1(y_i, f(X_i))$$

$$\hat{\mathbf{V}}\left[l_1(f(X), y)\right] = \frac{n}{n-1}\hat{p}(1-\hat{p})$$

SVP on 0-1 loss?

- is SVP qualitatively different?

$$\hat{p} := \frac{1}{n}\sum_{i=1}^{n} l_1(y_i, f(X_i))$$

$$\hat{\mathbf{V}}\left[l_1(f(X), y)\right] = \frac{n}{n-1}\hat{p}(1-\hat{p})$$

- ERM $\longrightarrow \hat{p}$

SVP on 0-1 loss?

- is SVP qualitatively different?

$$\hat{p} := \frac{1}{n}\sum_{i=1}^{n} l_1(y_i, f(X_i))$$

$$\hat{\mathbf{V}}\left[l_1(f(X), y)\right] = \frac{n}{n-1}\hat{p}(1-\hat{p})$$

  - ERM $\rightarrow \hat{p}$
  - SVP $\rightarrow \hat{p} + \lambda\sqrt{\hat{p}(1-\hat{p})}$

SVP on 0-1 loss?

- is SVP qualitatively different?

$$\hat{p} := \frac{1}{n} \sum_{i=1}^{n} l_1(y_i, f(X_i))$$

$$\hat{\mathbf{V}}\left[l_1(f(X), y)\right] = \frac{n}{n-1} \hat{p}(1 - \hat{p})$$

  - ERM $\rightarrow \hat{p}$
  - SVP $\rightarrow \hat{p} + \lambda \sqrt{\hat{p}(1 - \hat{p})}$
  - monotonic in $\hat{p} \in [0, 0.5)$

SVP on 0-1 loss?

- is SVP qualitatively different?

$$\hat{p} := \frac{1}{n}\sum_{i=1}^{n} l_1(y_i, f(X_i))$$

$$\hat{\mathbf{V}}\left[l_1(f(X), y)\right] = \frac{n}{n-1}\hat{p}(1-\hat{p})$$

  - ERM $\rightarrow \hat{p}$
  - SVP $\rightarrow \hat{p} + \lambda\sqrt{\hat{p}(1-\hat{p})}$
  - monotonic in $\hat{p} \in [0, 0.5)$
  - SVP on 0-1 loss gives back ERM for any λ !

SVP with exponential loss

SVP with exponential loss

- minimize

$$\sum_{i=1}^{n} e^{-y_i f(X_i)} + \tau \sqrt{\sum_{i>j} \left(e^{-y_i f(X_i)} - e^{-y_j f(X_j)} \right)^2}$$

SVP with exponential loss

- minimize

$$\sum_{i=1}^{n} e^{-y_i f(X_i)} + \tau \sqrt{\sum_{i>j} \left(e^{-y_i f(X_i)} - e^{-y_j f(X_j)}\right)^2}$$

- equivalently

$$\min_{f \in \mathcal{F}} \quad \sum_{i=1}^{n} e^{-y_i f(X_i)}$$

$$\text{s.t.} \quad \sqrt{\sum_{i>j} (e^{-y_i f(X_i)} - e^{-y_j f(X_j)})^2} \leq B$$

SVP with exponential loss

- minimize

$$\sum_{i=1}^{n} e^{-y_i f(X_i)} + \tau \sqrt{\sum_{i>j} \left(e^{-y_i f(X_i)} - e^{-y_j f(X_j)}\right)^2}$$

- equivalently

$$\min_{f \in \mathcal{F}} \quad \left(\sum_{i=1}^{n} e^{-y_i f(X_i)}\right)^2$$

$$\text{s.t.} \quad \sum_{i>j} (e^{-y_i f(X_i)} - e^{-y_j f(X_j)})^2 \leq B^2$$

SVP with exponential loss

- minimize

$$\sum_{i=1}^{n} e^{-y_i f(X_i)} + \tau \sqrt{\sum_{i>j} \left(e^{-y_i f(X_i)} - e^{-y_j f(X_j)}\right)^2}$$

- equivalently

$$\min_{f \in \mathcal{F}} \left(\sum_{i=1}^{n} e^{-y_i f(X_i)} \right)^2 + \lambda \left(\sum_{i>j} (e^{-y_i f(X_i)} - e^{-y_j f(X_j)})^2 - B^2 \right)$$

SVP with exponential loss

- minimize

$$\sum_{i=1}^{n} e^{-y_i f(X_i)} + \tau \sqrt{\sum_{i>j} \left(e^{-y_i f(X_i)} - e^{-y_j f(X_j)}\right)^2}$$

- equivalently

$$\min_{f \in \mathcal{F}} \left(\sum_{i=1}^{n} e^{-y_i f(X_i)} \right)^2 + \lambda \sum_{i>j} \left(e^{-y_i f(X_i)} - e^{-y_j f(X_j)}\right)^2$$

deriving an update rule

deriving an update rule

- start with

$$\min_{f\in\mathcal{F}} \left(\sum_{i=1}^{n} e^{-y_i f(X_i)}\right)^2 + \lambda \sum_{i>j} \left(e^{-y_i f(X_i)} - e^{-y_j f(X_j)}\right)^2$$

deriving an update rule

- start with

$$\min_{f \in \mathcal{F}} \left(\sum_{i=1}^{n} e^{-y_i f(X_i)} \right)^2 + \lambda \sum_{i>j} \left(e^{-y_i f(X_i)} - e^{-y_j f(X_j)} \right)^2$$

- build an additive model greedily

$$f(X) = \sum_{s=1}^{S} \alpha_s G^s(X)$$

deriving an update rule

- start with

$$\min_{f \in \mathcal{F}} \left(\sum_{i=1}^{n} e^{-y_i f(X_i)} \right)^2 + \lambda \sum_{i>j} \left(e^{-y_i f(X_i)} - e^{-y_j f(X_j)} \right)^2$$

- build an additive model greedily

$$f(X) = \sum_{s=1}^{S} \alpha_s G^s(X)$$

- choose a $G^s(X)$ and find α_s to minimize the above convex cost

AdaBoost (Freund & Schapire '97)

AdaBoost (Freund & Schapire '97)

- greedily minimizes

$$\sum_{i=1}^{n} e^{-y_i f(X_i)}$$

AdaBoost (Freund & Schapire '97)

- greedily minimizes

$$\sum_{i=1}^{n} e^{-y_i f(X_i)}$$

Initialize: $w_i \leftarrow \frac{1}{n}$

for s=1:S do

Get a weak learner $G^s(\cdot)$

$$\alpha_s = \frac{1}{4} \log \left(\frac{(\sum_{y_i = G^s(X_i)} w_i)^2}{(\sum_{y_i \neq G^s(X_i)} w_i)^2} \right)$$

if $\alpha_s < 0$ then *break*;

$w_i \leftarrow w_i e^{-y_i \alpha_s G^s(X_i)}$, normalize w

EBBoost

- greedily minimizes

$$\sum_{i=1}^{n} e^{-y_i f(X_i)}$$

Initialize: $w_i \leftarrow \frac{1}{n}$

for s=1:S do

Get a weak learner $G^s(\cdot)$

$$\alpha_s = \frac{1}{4} \log \left(\frac{(\sum_{y_i = G^s(X_i)} w_i)^2}{(\sum_{y_i \neq G^s(X_i)} w_i)^2} \right)$$

if $\alpha_s < 0$ then *break*;

$w_i \leftarrow w_i e^{-y_i \alpha_s G^s(X_i)}$, normalize w

EBBoost

- greedily minimizes

$$\left(\sum_{i=1}^{n} e^{-y_i f(X_i)}\right)^2 + \lambda \sum_{i>j} \left(e^{-y_i f(X_i)} - e^{-y_j f(X_j)}\right)^2$$

Initialize: $w_i \leftarrow \frac{1}{n}$

for s=1:S do

Get a weak learner $G^s(\cdot)$

$$\alpha_s = \frac{1}{4} \log \left(\frac{(\sum_{y_i = G^s(X_i)} w_i)^2}{(\sum_{y_i \neq G^s(X_i)} w_i)^2} \right)$$

if $\alpha_s < 0$ then *break*;

$w_i \leftarrow w_i e^{-y_i \alpha_s G^s(X_i)}$, normalize w

EBBoost

- greedily minimizes

$$\left(\sum_{i=1}^{n} e^{-y_i f(X_i)}\right)^2 + \lambda \sum_{i>j} \left(e^{-y_i f(X_i)} - e^{-y_j f(X_j)}\right)^2$$

Initialize: $w_i \leftarrow \frac{1}{n}$

for s=1:S do

Get a weak learner $G^s(\cdot)$

$$\alpha_s = \frac{1}{4} \log \left(\frac{(\sum_{y_i = G^s(X_i)} w_i)^2 + \lambda n \sum_{y_i = G^s(X_i)} w_i^2/(1-\lambda)}{(\sum_{y_i \neq G^s(X_i)} w_i)^2 + \lambda n \sum_{y_i \neq G^s(X_i)} w_i^2/(1-\lambda)} \right)$$

if $\alpha_s < 0$ then *break*;

$w_i \leftarrow w_i e^{-y_i \alpha_s G^s(X_i)}$, normalize w

experiments

experiments

- several benchmark datasets
- weak learner: decision stump
- parameters via a validation set
- boosting until no drop in validation error in 50 steps
- competing methods
  - AdaBoost
  - RLP-Boost
  - RQP-Boost
  - Soft-margin : relaxed boosting

results

results

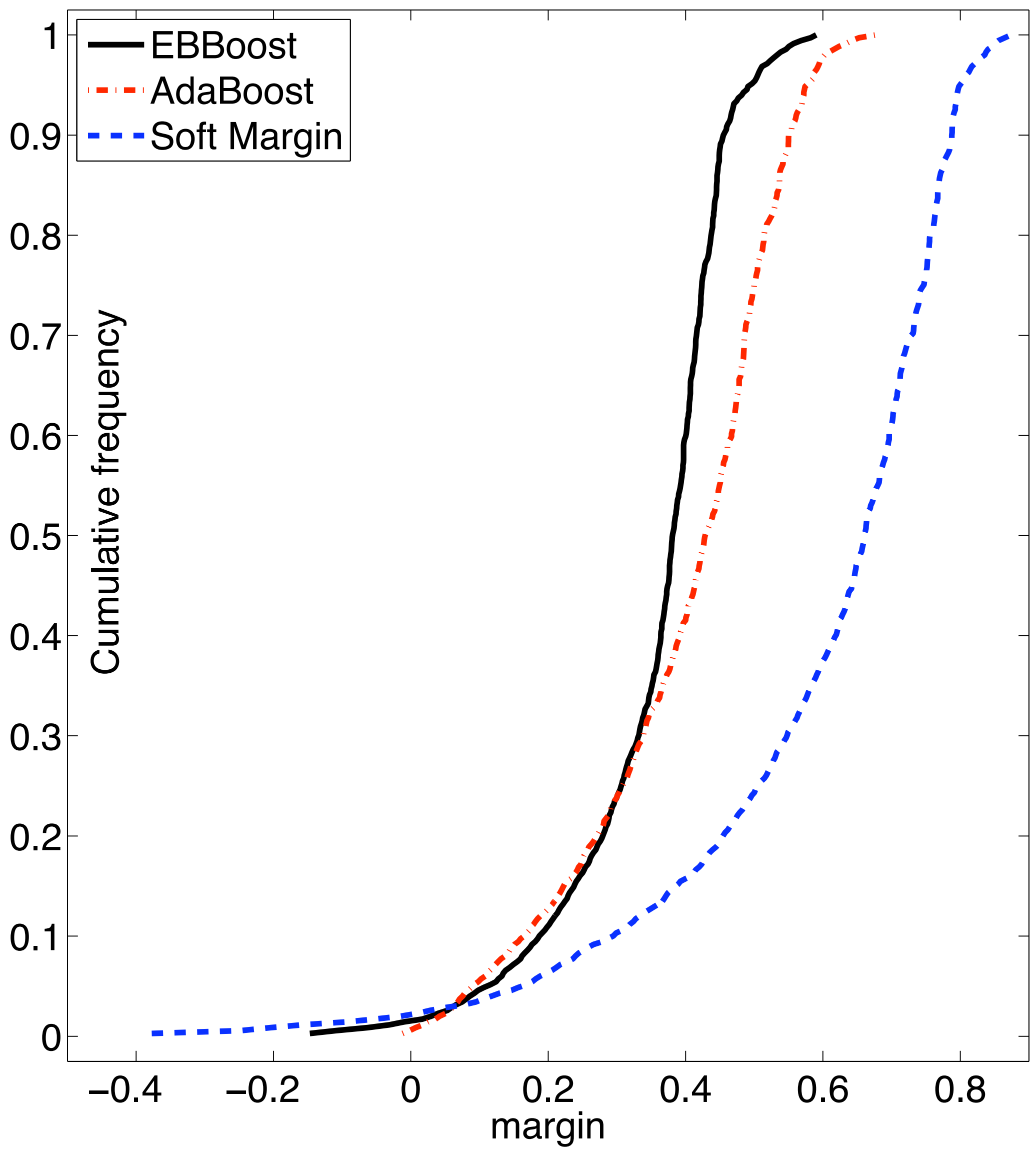
Dataset	AdaBoost	EBBoost
a5a	18.07 ± 0.6	17.82 ± 0.6
abalone	22.53 ± 0.8	22.38 ± 0.9
image	4.28 ± 0.8	4.04 ± 0.8
nist09	1.28 ± 0.2	1.17 ± 0.1
nist14	0.80 ± 0.2	0.70 ± 0.1
nist27	2.56 ± 0.3	2.41 ± 0.3
nist38	5.68 ± 0.6	5.34 ± 0.4
nist56	3.64 ± 0.5	3.38 ± 0.4
mushrooms	0.35 ± 0.3	0.28 ± 0.3
musklarge	7.80 ± 1.0	6.89 ± 0.6
ringnorm	15.05 ± 3.1	13.45 ± 2.4
spambase	7.74 ± 0.7	7.18 ± 0.8
splice	10.57 ± 1.1	10.27 ± 0.9
twonorm	4.30 ± 0.4	4.00 ± 0.2
w4a	2.80 ± 0.2	2.75 ± 0.2
waveform	12.96 ± 0.8	12.90 ± 0.8
wine	26.03 ± 1.2	25.66 ± 1.0
wisc	5.00 ± 1.5	4.00 ± 1.3

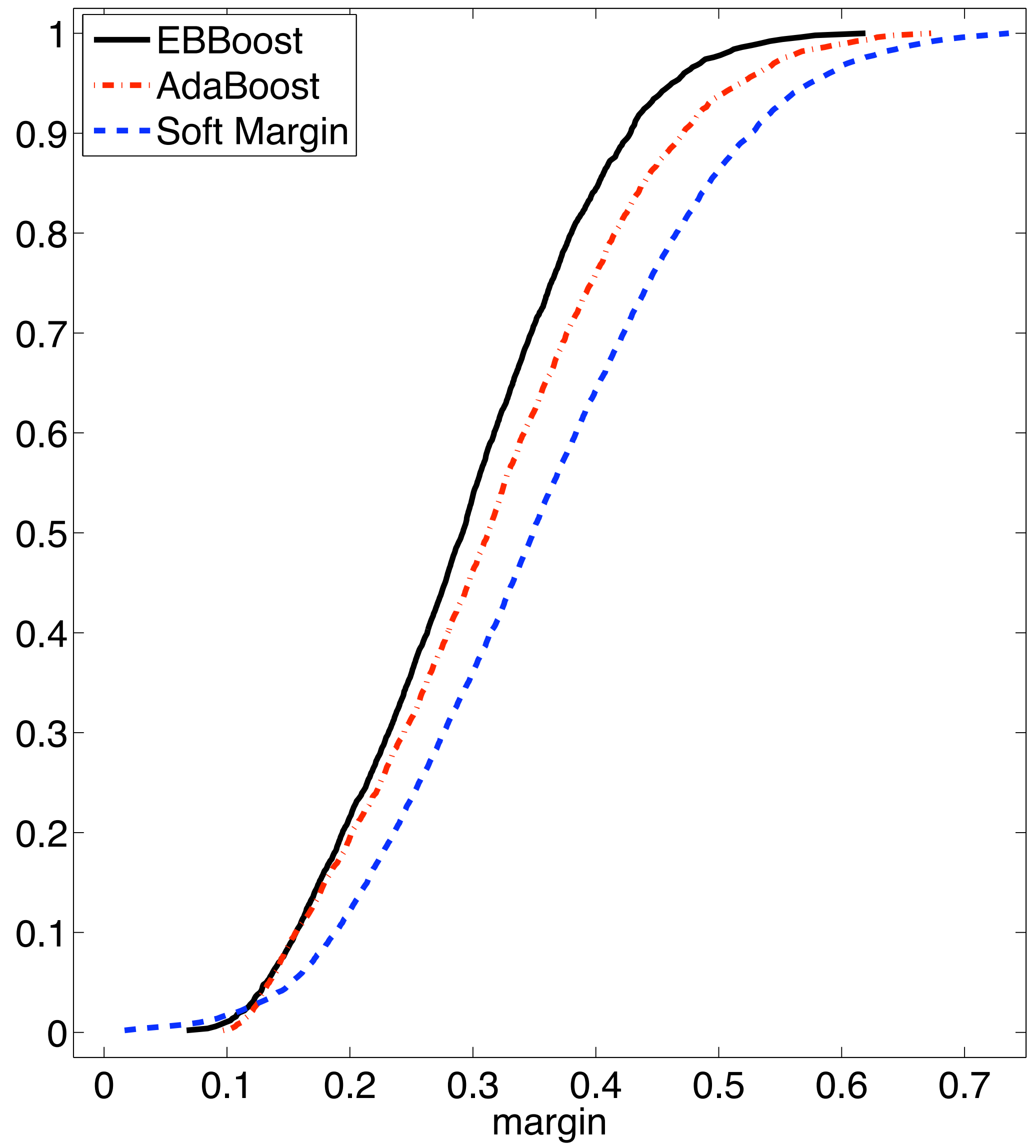
results

Dataset	AdaBoost	EBBoost	RLP-Boost	RQP-Boost	ABR
a5a	18.07 ± 0.6	17.82 ± 0.6	17.90 ± 0.8	18.06 ± 0.9	17.80 ± 0.5
abalone	22.53 ± 0.8	22.38 ± 0.9	23.68 ± 1.3	23.01 ± 1.3	22.40 ± 0.7
image	4.28 ± 0.8	4.04 ± 0.8	4.19 ± 0.8	3.79 ± 0.7	4.27 ± 0.8
nist09	1.28 ± 0.2	1.17 ± 0.1	1.43 ± 0.2	1.25 ± 0.2	1.18 ± 0.2
nist14	0.80 ± 0.2	0.70 ± 0.1	0.89 ± 0.2	0.78 ± 0.2	0.74 ± 0.1
nist27	2.56 ± 0.3	2.41 ± 0.3	2.72 ± 0.3	2.49 ± 0.3	2.32 ± 0.3
nist38	5.68 ± 0.6	5.34 ± 0.4	6.04 ± 0.4	5.48 ± 0.5	5.24 ± 0.5
nist56	3.64 ± 0.5	3.38 ± 0.4	3.97 ± 0.5	3.61 ± 0.4	3.42 ± 0.3
mushrooms	0.35 ± 0.3	0.28 ± 0.3	0.30 ± 0.3	0.30 ± 0.3	0.29 ± 0.4
musklarge	7.80 ± 1.0	6.89 ± 0.6	7.83 ± 1.0	7.29 ± 1.0	7.22 ± 0.7
ringnorm	15.05 ± 3.1	13.45 ± 2.4	15.25 ± 4.2	14.55 ± 3.0	14.35 ± 3.1
spambase	7.74 ± 0.7	7.18 ± 0.8	7.45 ± 0.6	7.25 ± 0.7	6.99 ± 0.6
splice	10.57 ± 1.1	10.27 ± 0.9	10.28 ± 0.8	10.18 ± 1.0	10.02 ± 0.9
twonorm	4.30 ± 0.4	4.00 ± 0.2	4.87 ± 0.5	4.19 ± 0.4	4.16 ± 0.4
w4a	2.80 ± 0.2	2.75 ± 0.2	2.76 ± 0.1	2.77 ± 0.2	2.75 ± 0.2
waveform	12.96 ± 0.8	12.90 ± 0.8	12.75 ± 0.9	12.22 ± 0.9	12.47 ± 0.7
wine	26.03 ± 1.2	25.66 ± 1.0	25.00 ± 1.2	25.20 ± 1.0	25.09 ± 1.2
wisc	5.00 ± 1.5	4.00 ± 1.3	4.14 ± 1.5	4.71 ± 1.5	4.46 ± 1.6

margin distribution

margin distribution





Margin statistics

Margin statistics

	AdaBoost	EBBoost	ABR
a5a	0.21 ± 0.20	0.19 ± 0.17	0.20 ± 0.19
abal	0.12 ± 0.12	0.12 ± 0.12	0.13 ± 0.13
image	0.14 ± 0.08	0.13 ± 0.06	0.14 ± 0.08
nist09	0.45 ± 0.13	0.44 ± 0.12	0.48 ± 0.13
nist14	0.47 ± 0.12	0.38 ± 0.07	0.51 ± 0.12
nist27	0.32 ± 0.12	0.29 ± 0.10	0.35 ± 0.13
nist38	0.22 ± 0.10	0.20 ± 0.08	0.24 ± 0.10
nist56	0.30 ± 0.12	0.29 ± 0.11	0.32 ± 0.13
mush	0.26 ± 0.06	0.26 ± 0.05	0.28 ± 0.07
musk	0.18 ± 0.09	0.15 ± 0.06	0.18 ± 0.09
ring	0.15 ± 0.07	0.14 ± 0.06	0.15 ± 0.07
spam	0.21 ± 0.13	0.19 ± 0.10	0.23 ± 0.13
splice	0.19 ± 0.12	0.18 ± 0.10	0.22 ± 0.14
twon	0.29 ± 0.14	0.26 ± 0.11	0.30 ± 0.14
w4a	0.27 ± 0.11	0.23 ± 0.07	0.38 ± 0.12
wave	0.25 ± 0.17	0.22 ± 0.14	0.28 ± 0.19
wine	0.13 ± 0.15	0.13 ± 0.14	0.12 ± 0.14
wisc	0.39 ± 0.15	0.35 ± 0.12	0.59 ± 0.21

conclusions

conclusions

- proposed a novel boosting algorithm
  - well motivated
  - easy to implement
  - superior performance

conclusions

- proposed a novel boosting algorithm
  - well motivated
  - easy to implement
  - superior performance
- SVP is viable

conclusions

- proposed a novel boosting algorithm
  - well motivated
  - easy to implement
  - superior performance
- SVP is viable

- extending to other losses

conclusions

- proposed a novel boosting algorithm
  - well motivated
  - easy to implement
  - superior performance
- SVP is viable

- extending to other losses
- sample variance in margin distribution bounds

conclusions

- proposed a novel boosting algorithm
  - well motivated
  - easy to implement
  - superior performance
- SVP is viable

- extending to other losses
- sample variance in margin distribution bounds
- is it possible to estimate λ?