APPLICATIONS OF REGULARIZED LEAST SQUARES TO CLASSIFICATION PROBLEMS

Nicolò Cesa-Bianchi
Università di Milano, Italy

Contributors: Alex Conconi, Claudio Gentile, Luca Zaniboni

DESIDERATA FOR LEARNING ALGORITHMS

- Good empirical performance
- Scalability
- Versatility
- Theoretical guarantees

CONTENTS

BINARY PATTERN CLASSIFICATION

Each data element is encoded as an attribute vector

$$\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d \qquad \text{(instance)}$$

A binary label $y \in \{-1, +1\}$ classifies each instance $\boldsymbol{x}$ according to a given semantic property

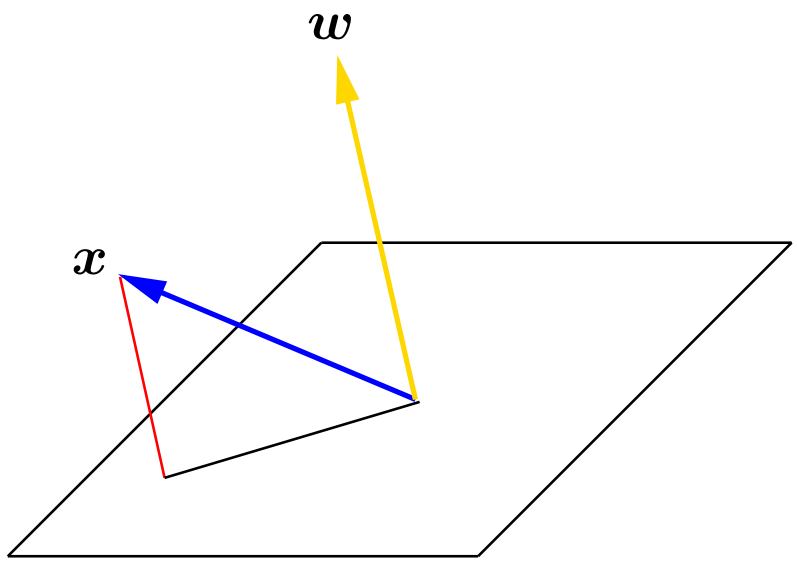
Design a rule $h : \mathbb{R}^d \to \{-1, +1\}$ for predicting labels given instances

No assumptions on the data-generating mechanism

LINEAR-THRESHOLD CLASSIFIERS

Classifier parametrized by coefficients (weights) $\boldsymbol{w} \in \mathbb{R}^d$

$h_{\boldsymbol{w}}(\boldsymbol{x}) = \text{SGN}(\boldsymbol{w}^\top \boldsymbol{x})$ $\quad$ $h_{\boldsymbol{w}}$ errs on $(\boldsymbol{x}, y)$ $\quad$ iff $\quad$ $y\,\boldsymbol{w}^\top \boldsymbol{x} < 0$



RLSC vs. SVM

Training set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ $\quad \boldsymbol{x}_t \in \mathbb{R}^d \quad y_t \in \{-1, 1\}$

Margin on t-th example: $m_t(\boldsymbol{v}) = y_t \, \boldsymbol{v}^\top \boldsymbol{x}_t$

1-norm SVM classifier (hinge loss)

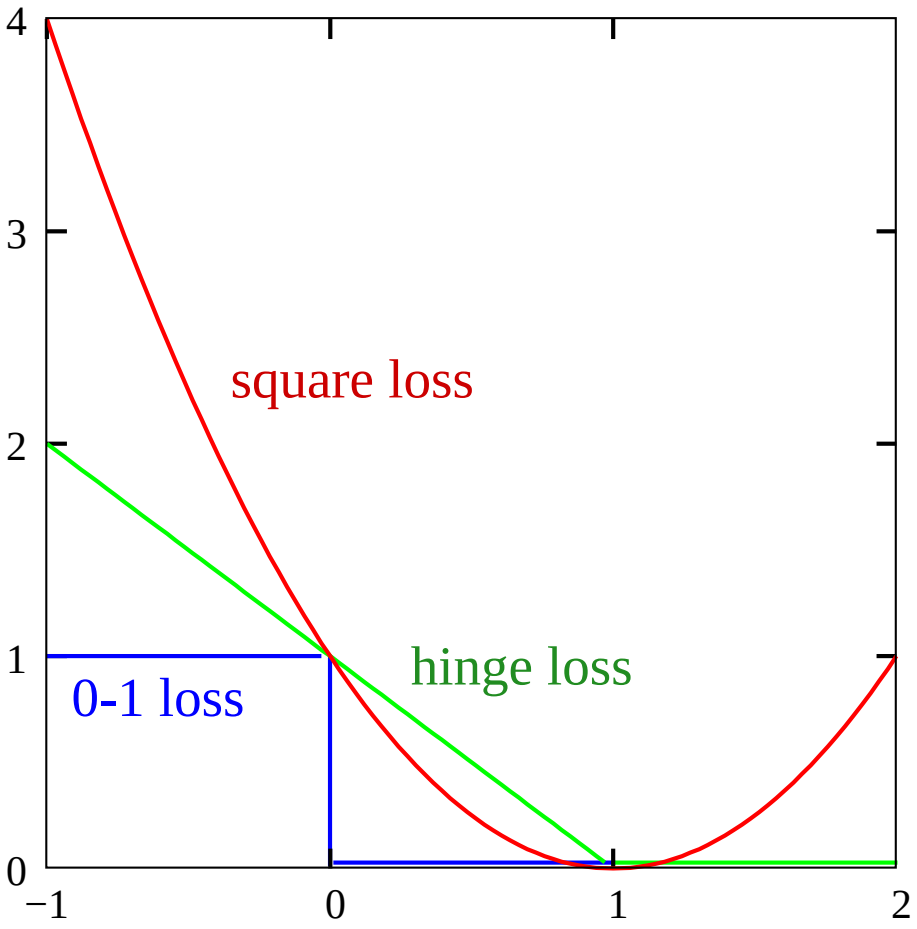
$$\boldsymbol{w} = \operatorname*{argmin}_{\boldsymbol{v} \in \mathbb{R}^d} \left(\sum_{t=1}^{n} \left[1 - m_t(\boldsymbol{v}) \right]_+ + a \, \|\boldsymbol{v}\|^2 \right)$$

Regularized Least Squares for Classification (square loss)

$$\boldsymbol{w} = \operatorname*{argmin}_{\boldsymbol{v} \in \mathbb{R}^d} \left(\sum_{t=1}^{n} \left(1 - m_t(\boldsymbol{v}) \right)^2 + a \, \|\boldsymbol{v}\|^2 \right)$$

2-norm SVM classifier $\left(\left[1 - m_t(\boldsymbol{v}) \right]_+ \right)^2$

LOSSES AS FUNCTIONS OF MARGIN



RLSC IS SOLVED ANALYTICALLY

$$\boldsymbol{w} = \operatorname*{argmin}_{\boldsymbol{v} \in \mathbb{R}^d} \left(\sum_{t=1}^{n} \left(1 - m_t(\boldsymbol{v})\right)^2 + a \left\|\boldsymbol{v}\right\|^2 \right)$$

Closed form solution $\boldsymbol{w} = \left(aI + S\,S^\top\right)^{-1} S\,\boldsymbol{y}$

- $I =$ identity matrix
- $S = [\, \boldsymbol{x}_1 \ \ldots \ \boldsymbol{x}_n \,]$ matrix of instances
- $\boldsymbol{y} = (\, y_1, \ldots, y_n \,)$ vector of labels

Solution is not sparse...

ENFORCING SPARSITY IN RLSC

Second-order Perceptron

Initial classifier $\boldsymbol{w} = (0, \ldots, 0)$

For $t = 1, 2, \ldots$

1. get next example $(\boldsymbol{x}_t, y_t)$
2. if $\text{SGN}(\boldsymbol{w}^\top \boldsymbol{x}_t) \neq y_t$ then perform RLSC-update of $\boldsymbol{w}$:

$$S = [\boldsymbol{x}_{s_1}, \ldots, \boldsymbol{x}_{s_m}, \boldsymbol{x}_t] \quad \text{matrix of mistaken instances}$$

$$\boldsymbol{y} = (y_{s_1}, \ldots, y_{s_m}, y_t) \quad \text{vector of mistaken labels}$$

$$\boldsymbol{w} = \left(a\, I + S\, S^\top\right)^{-1} S\, \boldsymbol{y}$$

REMARKS

Algorithms works in dual variables (kernels OK!)

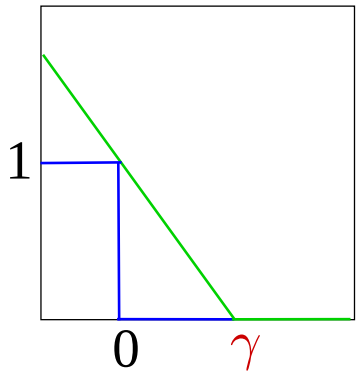
Final classifier only depends on mistaken examples:

$$\text{mistakes} \quad \approx \quad \text{support vectors}$$

Incremental update of $\left(I + S\,S^\top\right)^{-1}$ in time $\Theta(\text{mistakes}^2)$ when working with dual variables

Strong theoretical guarantees...

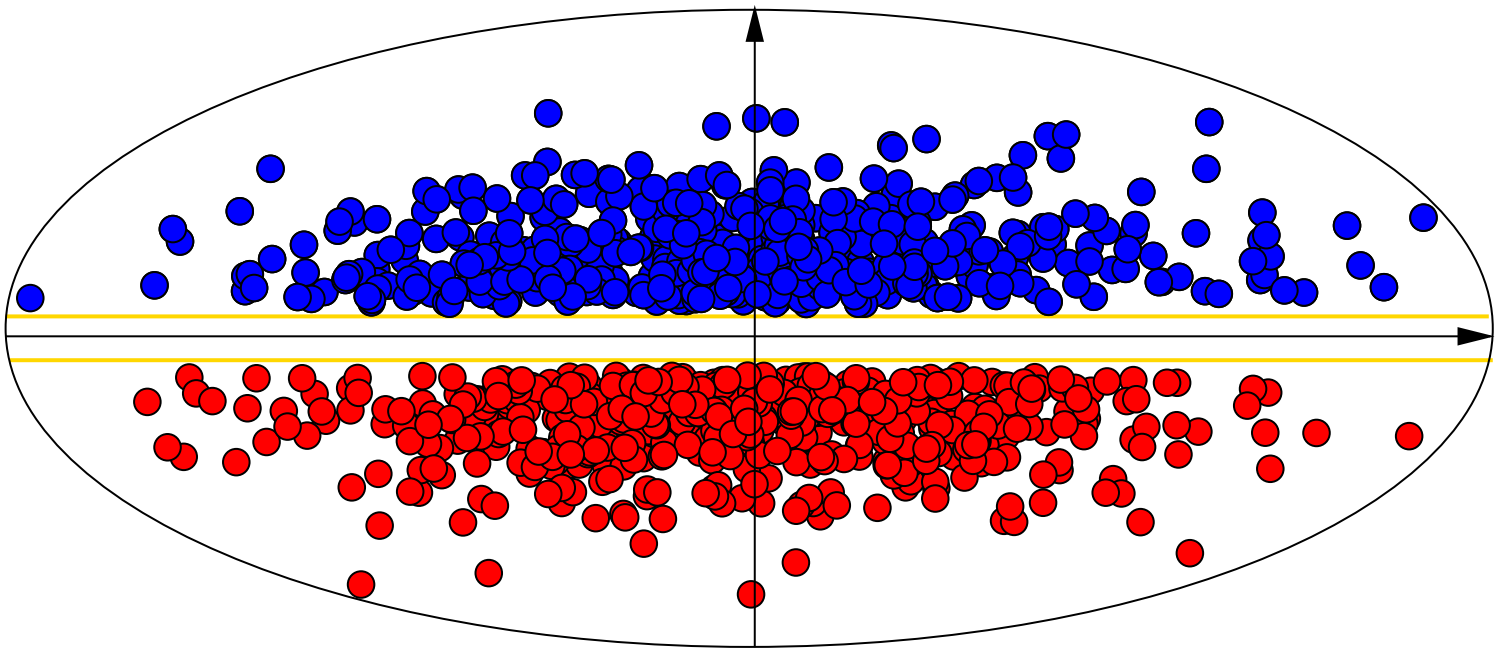
BOUND ON UPDATES (MISTAKES)



$$\inf_{\gamma>0} \min_{\|\boldsymbol{u}\|=1} \frac{1}{\gamma} \left(\sum_{t=1}^{n} \left[\gamma - m_t(\boldsymbol{u})\right]_+ + \sqrt{(1+\lambda(\boldsymbol{u})) \sum_{i=1}^{d} \ln(1+\lambda_i)} \right)$$

- $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $S\,S^\top$ $\qquad \lambda_{\min} \le \lambda(\boldsymbol{u}) \le \lambda_{\max}$
- Bound holds on any individual data sequence!

WHERE SECOND ORDER PROVABLY WINS



Mistakes of first-order Perceptron: $\leq \lambda_{\max}/\text{margin}^2$

Mistakes of second-order Perceptron: $\leq \lambda_{\min}/\text{margin}^2$

FROM MISTAKE BOUNDS TO RISK BOUNDS

Assume data $(\boldsymbol{x}, y)$ are drawn i.i.d. from an unknown distribution on $\mathbb{R}^d \times \{-1, 1\}$ (statistical learning model)

Statistical risk of classifier h is $\text{RISK}(h) = \mathbb{P}\big(h(\boldsymbol{x}) \neq y\big)$

Risk bound for typical second-order classifier

risk at most $\dfrac{M_n}{n} + \sqrt{\dfrac{2}{n} \ln \dfrac{1}{\delta}}$ with probability at least $1 - \delta$

M_n is number of updates (mistakes) made by 2nd order Perceptron on training set

HOW TO PICK A GOOD CLASSIFIER

$h_0, h_1, \ldots, h_m$ classifiers generated during a run

1. test each h_t on $(\boldsymbol{x}_{t+1}, y_{t+1}), (\boldsymbol{x}_{t+2}, y_{t+2}), \ldots$
2. pick h_t minimizing a penalized risk estimate

Risk bound

risk at most $\dfrac{M_n}{n} + 5\sqrt{\dfrac{1}{n} \ln \dfrac{2(n+1)}{\delta}}$ with probability at least $1 - \delta$

PREDICTION, LEARNING, AND GAMES

Repeated game theory

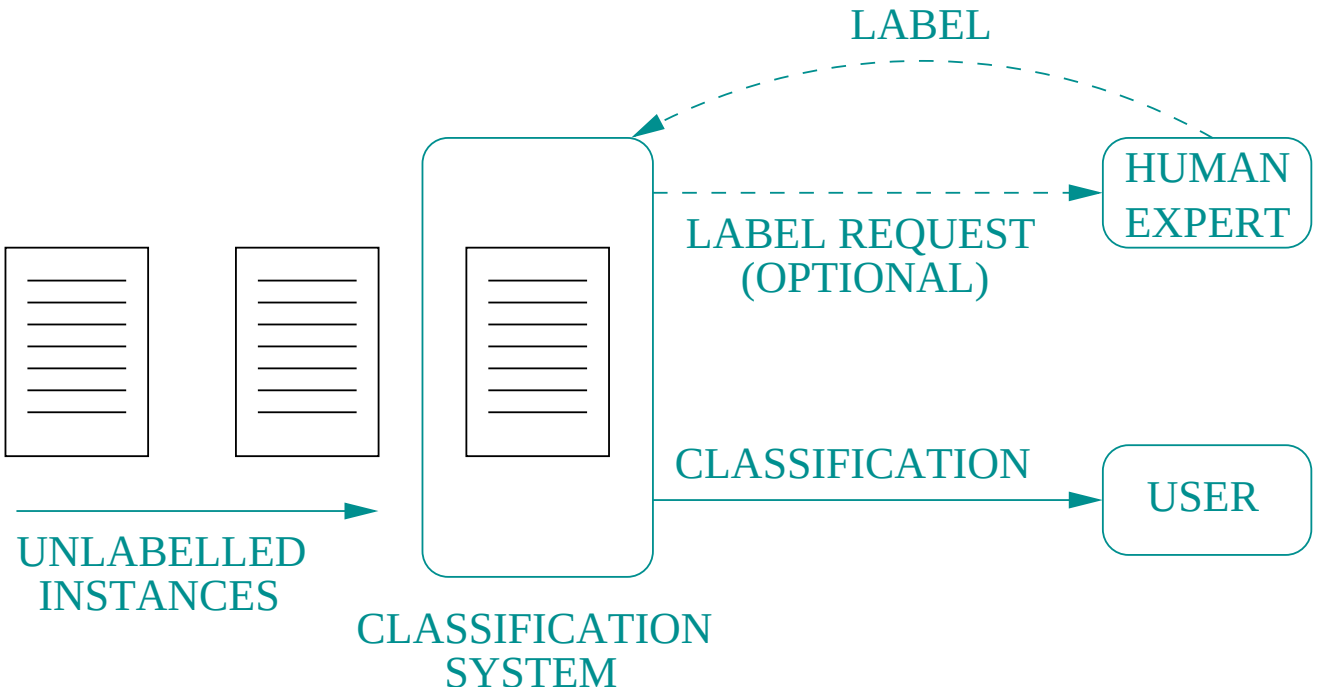
[Hannan, 1957]

Prediction with expert advice

[Littlestone and Warmuth, 1989; Vovk, 1990]

LEARNING	PREDICTION	GAMES
FEATURES	EXPERTS	PURE STRATEGIES
LINEAR CLASSIFIER	WEIGHTED FORECASTER	MIXED STRATEGY
PERCEPTRON TH.	EXPERTS TH.	HANNAN'S TH.
[1962]	[1993]	[1957]

SEMI-SUPERVISED LEARNING



Goal: Minimize classification mistakes and rate of requested labels

SUFFICIENT QUERY RATE

Measure performance against best classifier in a finite set (experts model)

Sampling data at random with rate $(\log n)(\log \log n)$ guarantees
convergence to best expert

A RLSC-BASED SEMI-SUPERVISED ALGORITHM

- Classify next instance $\boldsymbol{x}_t$ with $\text{SGN}(\boldsymbol{w}^\top \boldsymbol{x}_t)$
- Compute $\Delta_t = \boldsymbol{w}^\top \boldsymbol{x}_t / \|\boldsymbol{x}_t\|$
- Query label y_t with probability

$$\frac{1}{1 + |\Delta_t|/C + \Theta(\Delta_t^2)}$$

- If label queried and $\text{SGN}(\Delta_t) \neq y_t$,
  then perform RLSC-update of $\boldsymbol{w}$

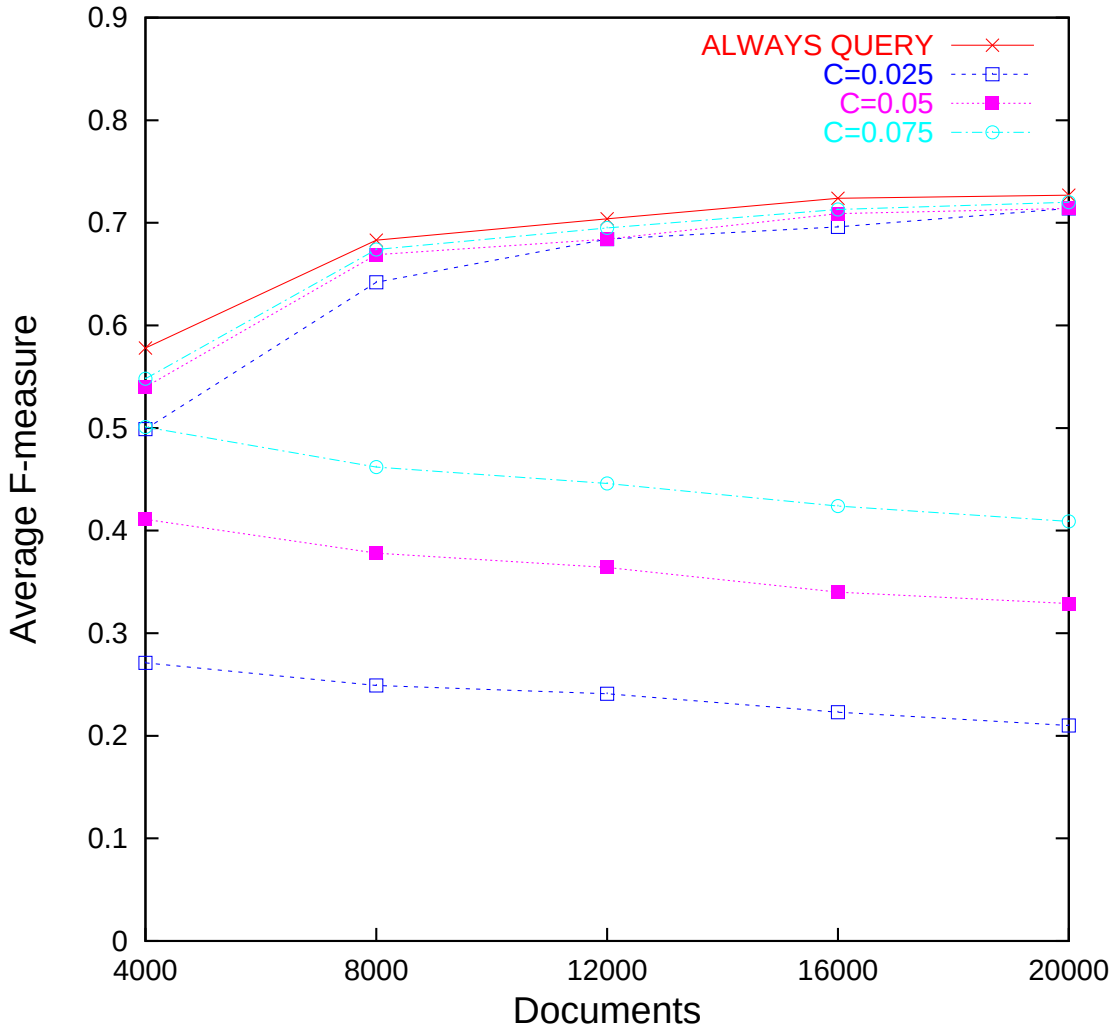
EXPECTED BOUNDS ON MISTAKES

Note: Only mistakes on queried labels trigger updates

$$\inf_{\gamma>0} \min_{\|\boldsymbol{u}\|=1} \frac{1}{\gamma} \left(\sum_{t=1}^{n} \left[\gamma - m_t(\boldsymbol{u})\right]_+ + \sqrt{(1+\mathbb{E}\Lambda(\boldsymbol{u})) \sum_{i=1}^{d} \mathbb{E} \ln\left(1+\Lambda_i\right)} \right)$$

- Bound holds on any data sequence (proper tuning of C needed)
- Apparently, no loss of performance due to semi-supervision!

REUTERS — 50 CATEGORIES



A PROBABILISTIC MODEL FOR LABELS

Labels $y_t \in \{-1, +1\}$ drawn from parametric distribution

$$\mathbb{P}(y_t = 1 \mid \boldsymbol{x}_t) = \frac{1 + \boldsymbol{u}^\top \boldsymbol{x}_t}{2} \qquad \|\boldsymbol{u}\| = 1$$

If $\boldsymbol{w}$ is solution of RLSC problem, then

$$\mathbb{E}\left[\boldsymbol{w}^\top \boldsymbol{x}_t\right] = \boldsymbol{u}^\top \boldsymbol{x}_t + \text{smaller order terms}$$

We conjecture a Chernoff bound for $|\boldsymbol{w}^\top \boldsymbol{x}_t - \boldsymbol{u}^\top \boldsymbol{x}_t|$

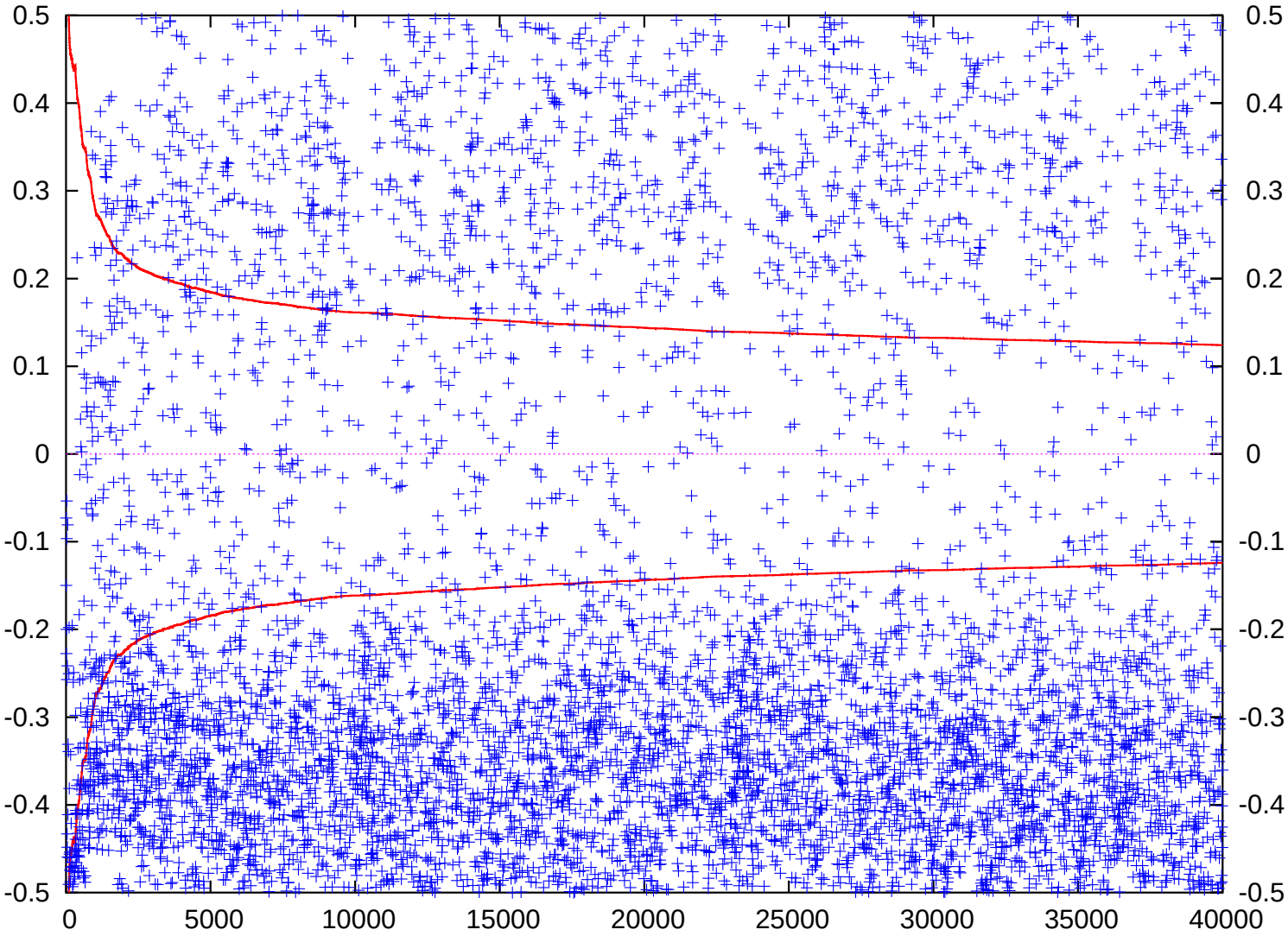
A SEMI-SUPERVISED DETERMINISTIC ALGORITHM

- Classify next instance $\boldsymbol{x}_t$ with $\text{SGN}(\boldsymbol{w}^\top \boldsymbol{x}_t)$
- Compute $\Delta_t = \boldsymbol{w}^\top \boldsymbol{x}_t / \|\boldsymbol{x}_t\|$
- If $|\Delta_t| \leq \sqrt{\frac{4 \ln t}{N_t}}$ then query label y_t of $\boldsymbol{x}_t$
- If label queried then perform RLSC-update of $\boldsymbol{w}$

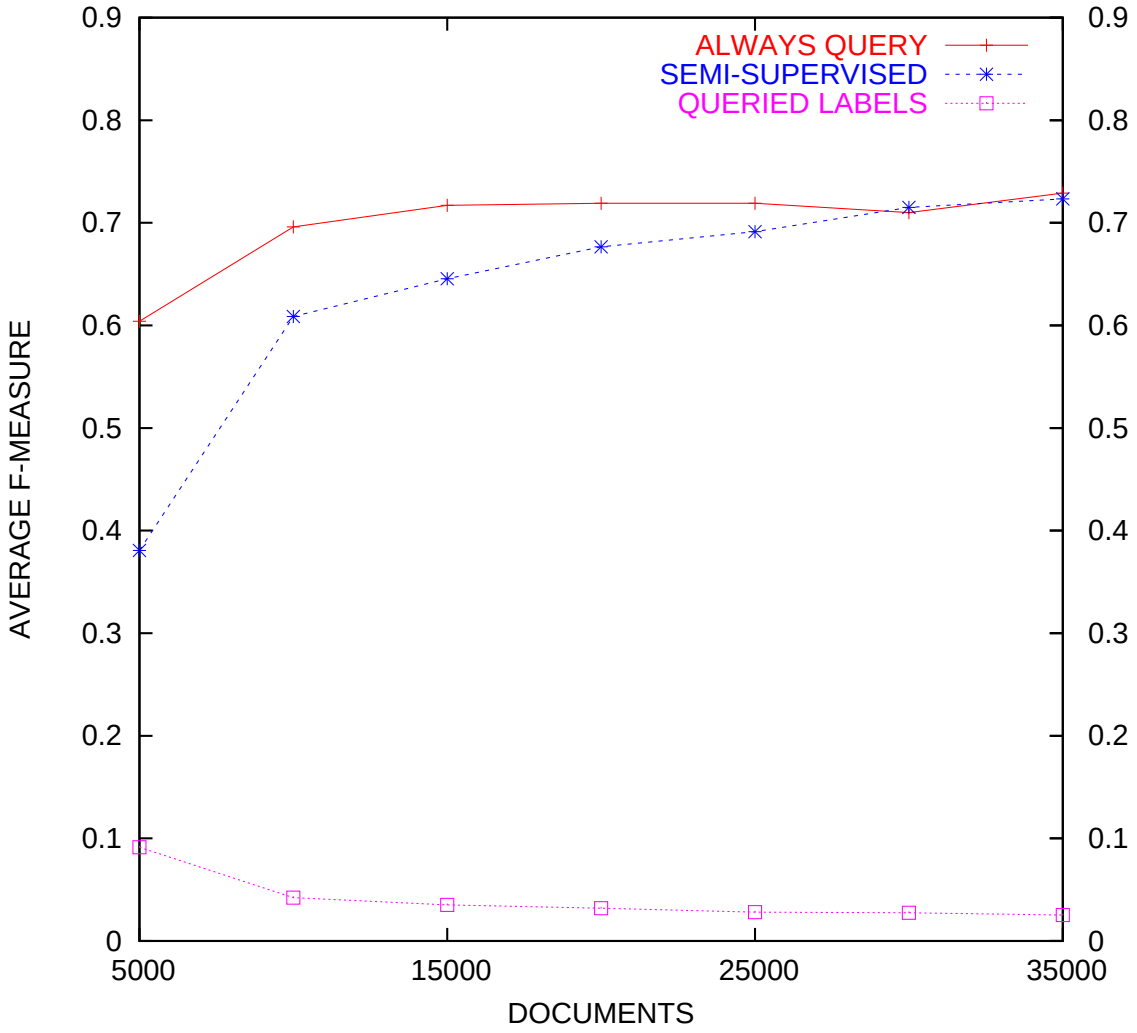
$N_t =$ number of labels sampled so far

Schedule $\sqrt{(4 \ln t)/N_t}$ derived from conjectured Chernoff bound

MARGIN PLOT


0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
0
5000
10000
15000
20000
25000
30000
35000
40000

REUTERS — 50 CATEGORIES



CONCLUSIONS

• Second-order Perceptron (incremental RLSC) is a practically reasonable algorithm with strong theoretical guarantees

• Trading margin with probability of mistake may save lots of labels in semi-supervised learning

• Applications of RLSC-based algorithms to filtering and hierarchical classification

REFERENCES

R. Rifkin, G. Yeo, and T. Poggio. Regularized least squares classification. In *Advances in Learning Theory: Methods, Model and Applications*. IOS Press, 2003.

N. Cesa-Bianchi, A. Conconi and C. Gentile. A second-order Perceptron algorithm. *SIAM J. on Computing*. To appear.

N. Cesa-Bianchi, A. Conconi and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*. To appear.

N. Cesa-Bianchi, A. Conconi and C. Gentile. Learning probabilistic linear-threshold classifiers via selective sampling. COLT 2003, pages 373–386. Springer, 2003.

REFERENCES (CONT.)

N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. Worst-case analysis of selective sampling for linear-threshold algorithms. NIPS 17. MIT Press, to appear.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. Margin-based algorithms for information filtering. NIPS 15. MIT Press, 2003.

N. Cesa-Bianchi, A. Conconi and C. Gentile. Regret bounds for hierarchical classification with linear-threshold functions. COLT 2004, pages 93–108. Springer, 2003.

N. Cesa-Bianchi, C. Gentile, A. Tironi, and L. Zaniboni. Incremental Algorithms for Hierarchical Classification. NIPS 17. MIT Press, to appear.