Learning And Mining from DatA
http://lamda.nju.edu.cn

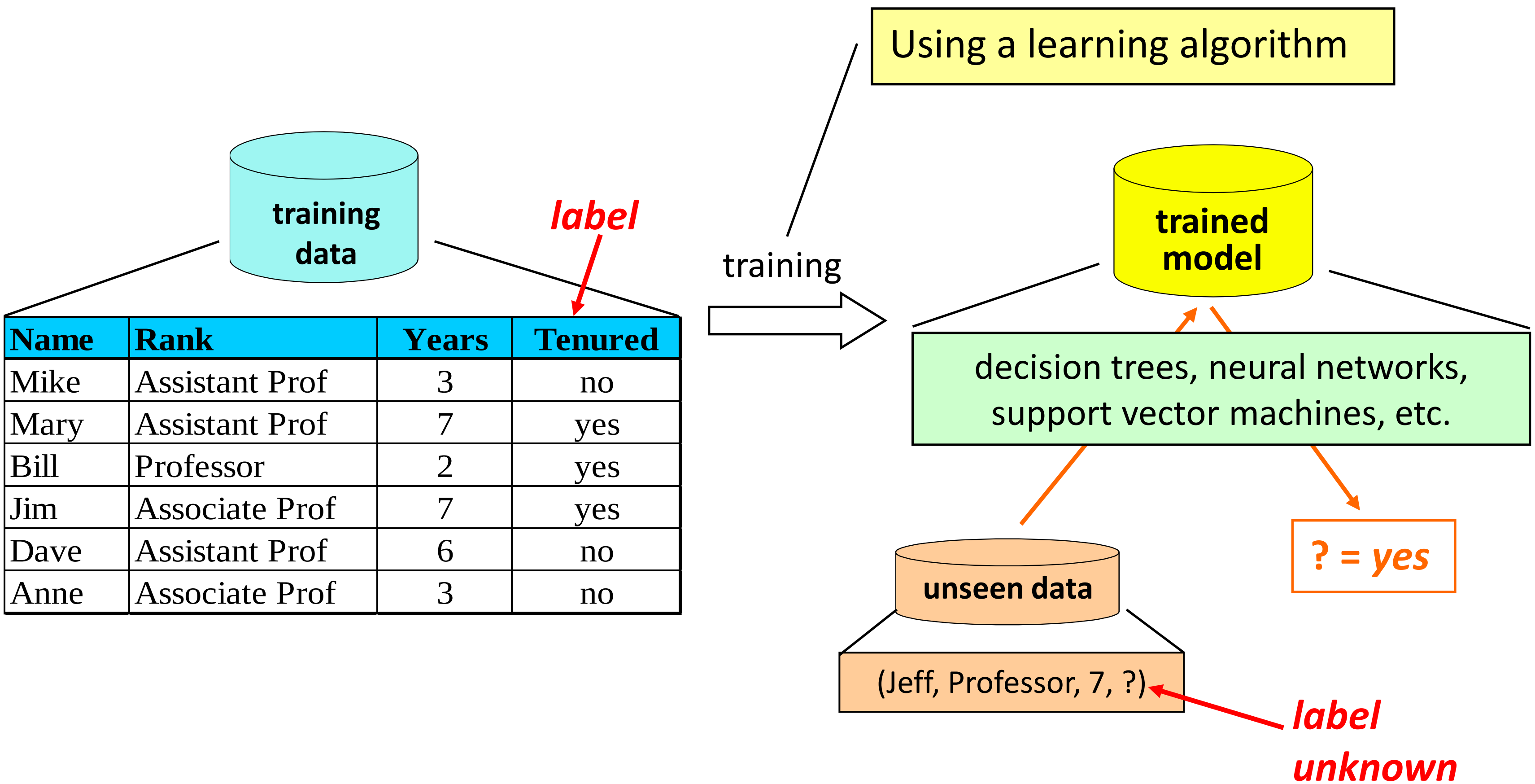
Boosting 25 Years

Zhi-Hua Zhou

http://cs.nju.edu.cn/zhouzh/
Email: zhouzh@nju.edu.cn

LAMDA Group
National Key Laboratory for Novel Software Technology,
Nanjing University, China

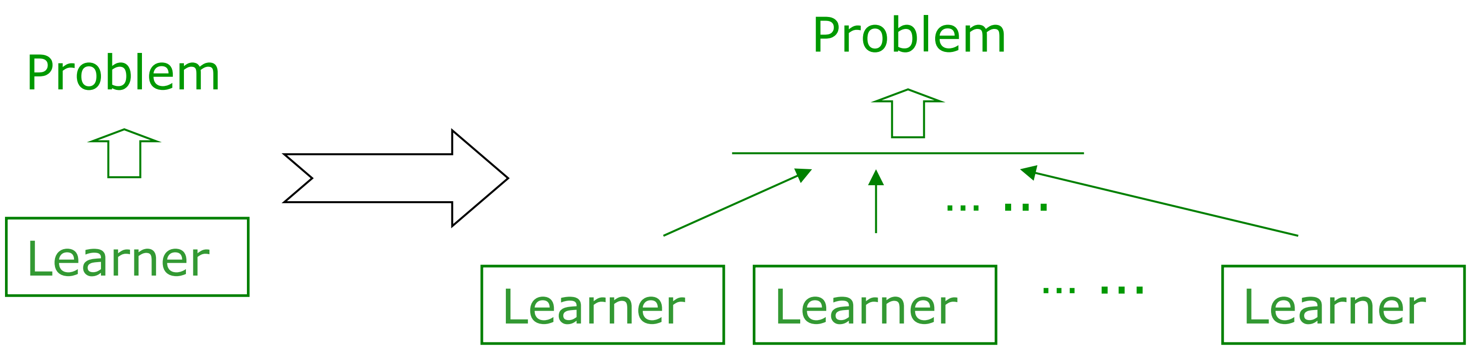
A typical machine learning process



Name	Rank	Years	Tenured
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Ensemble learning (集成学习)

"Ensemble methods" is a machine learning paradigm where multiple (homogenous/heterogeneous) individual learners are trained for the same problem
e.g. neural network ensemble, decision tree ensemble, etc.



The more **accurate** and **diverse** the component learners, the better the ensemble

Great success of ensemble methods

- KDDCup'07: 1st place for "… Decision Forests and …"
- KDDCup'08: 1st place of Challenge1 for a method using Bagging; 1st place of Challenge2 for "… Using an Ensemble Method "
- KDDCup'09: 1st place of Fast Track for "Ensemble … "; 2nd place of Fast Track for "… bagging … boosting tree models …", 1st place of Slow Track for "Boosting … "; 2nd place of Slow Track for "Stochastic Gradient Boosting"
- KDDCup'10: 1st place for "… Classifier ensembling"; 2nd place for "… Gradient Boosting machines … "

Great success of ensemble methods (cont')

- KDDCup'11: 1st place of Track 1 for "A Linear Ensemble … "; 2nd place of Track 1 for "Collaborative filtering Ensemble", 1st place of Track 2 for "Ensemble …"; 2nd place of Track 2 for "Linear combination of …"
- KDDCup'12: 1st place of Track 1 for "Combining… Additive Forest…"; 1st place of Track 2 for "A Two-stage Ensemble of…"
- KDDCup'13: 1st place of Track 1 for "Weighted Average Ensemble"; 2nd place of Track 1 for "Gradient Boosting Machine"; 1st place of Track 2 for "Ensemble the Predictions"

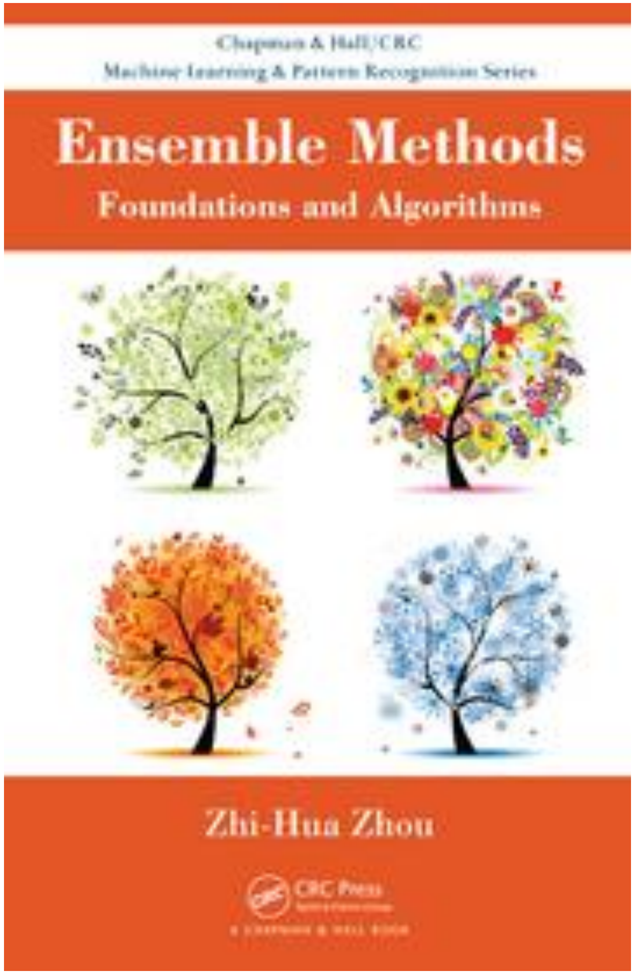
Great success of ensemble methods (cont')

- KDDCup'14: 1st place for "ensemble of GBM, ExtraTrees, Random Forest…" and "the weighted average"; 2nd place for "use both R and Python GBMs"; 3rd place for "gradient boosting machines... random forests" and "the weighted average of…"

- Netflix Prize:
  - ✓ 2007 Progress Prize Winner: Ensemble
  - ✓ 2008 Progress Prize Winner: Ensemble
  - ✓ 2009 $1 Million Grand Prize Winner:

Ensemble !!

More about ensemble methods



Z.-H. Zhou. Ensemble Methods: Foundations and Algorithms, Boca Raton, FL: Chapman & Hall/CRC, Jun. 2012. (ISBN 978-1-439-830031)

Many effective ensemble methods

- Sequential methods
  - **AdaBoost** [Freund & Schapire, JCSS97]
  - Arc-x4 [Breiman, AnnStat98]
  - LPBoost [Demiriz, Bennett, Shawe-Taylor, MLJ06]
  - … …
- Parallel methods
  - Bagging [Breiman, MLJ96]
  - Random Subspace [Ho, TPAMI98]
  - Random Forests [Breiman, MLJ01]
  - … …

Special focus of this talk: AdaBoost

Significant advantageous:

- Very accurate prediction
- Very simple ("*just 10 lines of code*" *as Schapire said*)
- Wide and successful applications
- Sound theoretical foundation
- ... ...

Gödel Prize (2003)

Freund & Schapire, A decision theoretic generalization of on-line learning and an application to Boosting. Journal of Computer and System Sciences, 1997, 55: 119-139.

The born of AdaBoost

An open problem [Kearns & Valiant, STOC'89]:

"weakly learnable" ?= "strongly learnable"

a problem is ***learnable*** or ***strongly learnable*** if there exists an algorithm that outputs a learner h in polynomial time such that for all $0 < \delta, \epsilon \leq 0.5,\ P(\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\mathbb{I}[h(\boldsymbol{x}) \neq f(\boldsymbol{x})]] < \epsilon) \geq 1 - \delta$

a problem is ***weakly learnable*** if there exists an algorithm that outputs a learner with error 0.5–1/p where p is a polynomial in problem size and other parameters

In other words, whether a "weak" learning algorithm that works just slightly better than random guess can be "boosted" into an arbitrarily accurate "strong" learning algorithm

The born of AdaBoost (con't)

- Amazingly, in 1990 Schapire proves that the answer is "yes". More importantly, the proof is a construction!

  This is the first Boosting algorithm

- In 1993, Freund presents a scheme of combining weak learners by majority voting in Phd thesis at UC Santa Cruz

However, these algorithms are not practical

- Later, at AT&T Bell Labs, Freund & Schapire published the 1997 journal paper (the work was reported in EuroCOLT'95), which proposed the AdaBoost algorithm, a practical algorithm

The AdaBoost algorithm

Learning And Mining from DatA
http://lamda.nju.edu.cn

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

$$D_{t+1}(i) = \frac{D_t(i) \boxed{\exp(-\alpha_t y_i h_t(x_i))}}{Z_t}$$

where Z_t is a normaliza tion).

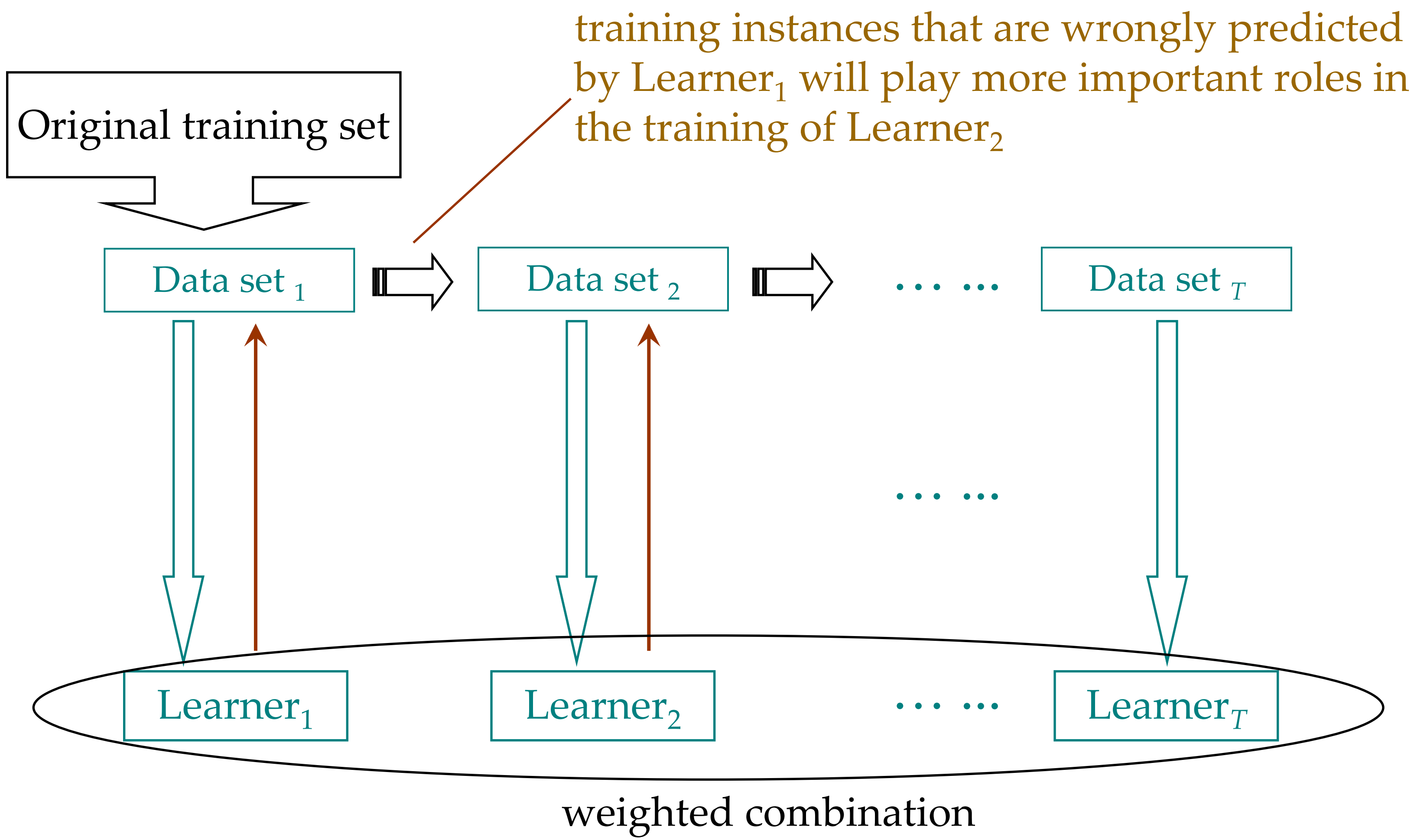
Output the final classifier:

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

typically $\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$
where $\epsilon_t = \Pr_{i \sim D_t}[h_t(x_i) \neq y_i]$

the weights of incorrectly classified examples are increased such that the base learner is forced to focus on the "hard" examples in the training set

A flowchart illustration



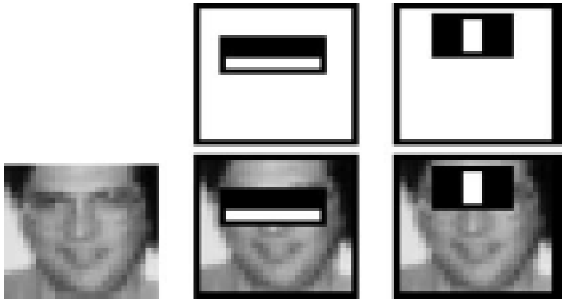
Why AdaBoost high impact?

First, it is simple yet effective

can be applied to almost all tasks where one wants to apply machine learning techniques

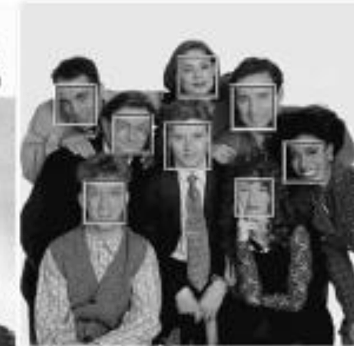
For example, in computer vision, the **Viola-Jones detector**

AdaBoost using harr-like features in a cascade structure



in average, only 8 features needed to be evaluated per image

The Viola-Jones detector



"the first real-time face detector"

Comparable accuracy, but **15 times faster** than state-of-the-art of face detectors (at that time)

Longuet-Higgins Prize (2011)

Viola & Jones, Rapid object detection using a Boosted cascade of simple features. CVPR, 2001.

Why AdaBoost high impact? (con't)

Learning And Mining from DatA
http://lamda.nju.edu.cn

Second, it generates the Boosting Family of algorithms

A general boosting procedure

Input: Sample distribution $\mathcal{D}$;
Base learning algorithm $\mathfrak{L}$;
Number of learning rounds T.

Process:
1. $\mathcal{D}_1 = \mathcal{D}$. % Initialize distribution
2. **for** $t = 1, \ldots, T$:
3. $h_t = \mathfrak{L}(\mathcal{D}_t)$; % Train a weak learner from distribution $\mathcal{D}_t$
4. $\epsilon_t = P_{\boldsymbol{x} \sim D_t}(h_t(\boldsymbol{x}) \neq f(\boldsymbol{x}))$; % Evaluate the error of h_t
5. $\mathcal{D}_{t+1} = Adjust\_Distribution(\mathcal{D}_t, \epsilon_t)$
6. **end**

Output: $H(\boldsymbol{x}) = Combine\_Outputs(\{h_1(\boldsymbol{x}), \ldots, h_t(\boldsymbol{x})\})$

A lot of Boosting algorithms:
AdaBoost.M1, AdaBoost.MR, FilterBoost, GentleBoost, GradientBoost, MadaBoost, LogitBoost, LPBoost, MultiBoost, RealBoost, RobustBoost, …

Why AdaBoost high impact? (con't)

Learning And Mining from DatA
http://lamda.nju.edu.cn

Third, there are sound theoretical results

Freund & Schapire [JCSS97] proved that the training error of AdaBoost is bounded by:

$$\epsilon = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\mathbb{I}[H(\boldsymbol{x}) \neq f(\boldsymbol{x})] \leq 2^T \prod_{t=1}^{T} \sqrt{\epsilon_t(1-\epsilon_t)} \leq e^{-2\sum_{t=1}^{T}\gamma_t^2}$$

where $\gamma_t = 0.5 - \epsilon_t$

Thus, if each base classifier is slightly better than random such that $\gamma_t \geq \gamma$ for some $\gamma > 0$, then the training error drops exponentially fast in T because the above bound is at most $e^{-2T\gamma^2}$

Generalization bound

Freund & Schapire [JCSS97] proved that the generalization error of AdaBoost is bounded by:

$$\epsilon_{\mathcal{D}} \le \epsilon_D + \tilde{O}\left(\sqrt{\frac{dT}{m}}\right)$$

with probability at least $1-\delta$, where d is the **VC-dimension** of base learners, m is the number of training instances, T is the number of learning rounds and $\tilde{O}(\cdot)$ is used instead of $O(\cdot)$ to hide logarithmic terms and constant factors.

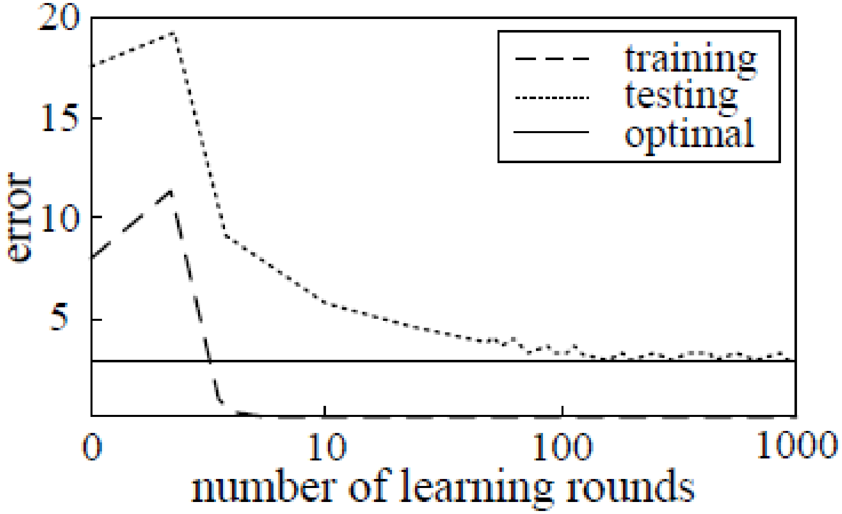
It implies that AdaBoost will **overfit** if T is large

Overfit (过拟合): The trained model fits the training data too much such that it can exaggerate minor fluctuations in the training data, leading to poor generalization performance

The Mystery

However, AdaBoost often does not overfit in real practice

A typical performance plot of AdaBoost on real data



Seems contradict with the **Occam's Razor**

Knowing the reason may inspire new methodology for algorithm design

Understanding why AdaBoost seems resistant to overfitting is the most fascinating fundamental theoretical issue

Figure replot based on [Schapire, Freund, Bartlett & Lee, Ann. Stat. 1998]

Major theoretical efforts

- **Margin Theory**

  Started from [Schapire, Freund, Bartlett & Lee, Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of Statistics, 26(5):1651–1686, 1998]

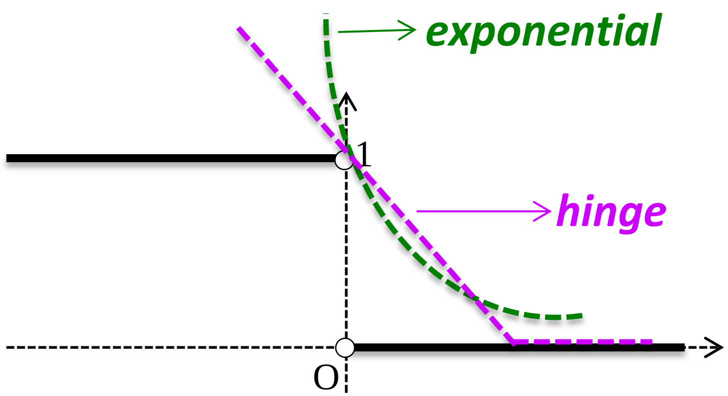
- **Statistical View**

  Started from [Friedman, Hastie & Tibshirani. Additive logistic regression: A statistical view of boosting (with discussions). Annals of Statistics, 28(2):337–407, 2000]

Intuition of the statistical view

In binary classification, we want to optimize the 0/1-loss

Because it is non-smooth, non-convex, …, in statistical learning usually we instead optimize a **surrogate loss**



The key step of the AdaBoost algorithm seems closely related to the exponential loss:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} e^{-\alpha_t y_i h_t(\boldsymbol{x}_i)} \qquad \alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

Statistical view of AdaBoost

Friedman, Hastie & Tibshirani [Ann. Stat. 2000] showed that if we consider the **additive model** $H(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t h_t(\boldsymbol{x}_t)$, take a logistic function and estimate probability via

$$P(f(\boldsymbol{x}) = 1 \mid \boldsymbol{x}) = \frac{e^{H(\boldsymbol{x})}}{e^{H(\boldsymbol{x})} + e^{-H(\boldsymbol{x})}}$$

then AdaBoost algorithm is a Newton-like procedure optimizing the exponential loss function and the log loss function (negative log-likelihood)

$$\ell_{log}(h \mid \mathcal{D}) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} \left[\ln \left(1 + e^{-2f(\boldsymbol{x})h(\boldsymbol{x})} \right) \right]$$

That is, AdaBoost can be viewed as a stage-wise estimation procedure for fitting an additive logistic regression model

Implications of the statistical view

As alternatives, one can fit the additive logistic regression model by optimizing the log loss function via other procedures, leading to many variants

e.g., LogitBoost [Friedman, Hastie & Tibshirani, Ann. Stat. 2000]
LPBoost [Demiriz, Bennett & Shawe-Taylor, MLJ 2002]
L2Boost [Bühlmann & Yu, JASA 2003]
RegBoost [Lugosi & Vayatis, Ann. Stat. 2004], etc.

The statistical view also encouraged the study of some specific statistical properties of AdaBoost

e.g., for **consistency**: Boosting with early stopping is consistent [Zhang & Yu, Ann. Stat. 2004], Exponential and logistic loss is consistent [Zhang, Ann. Stat. 2004, Bartlett, Jordana & McAuliffea, JASA 2006], etc.

LAMDA Learning And Mining from DatA http://lamda.nju.edu.cn

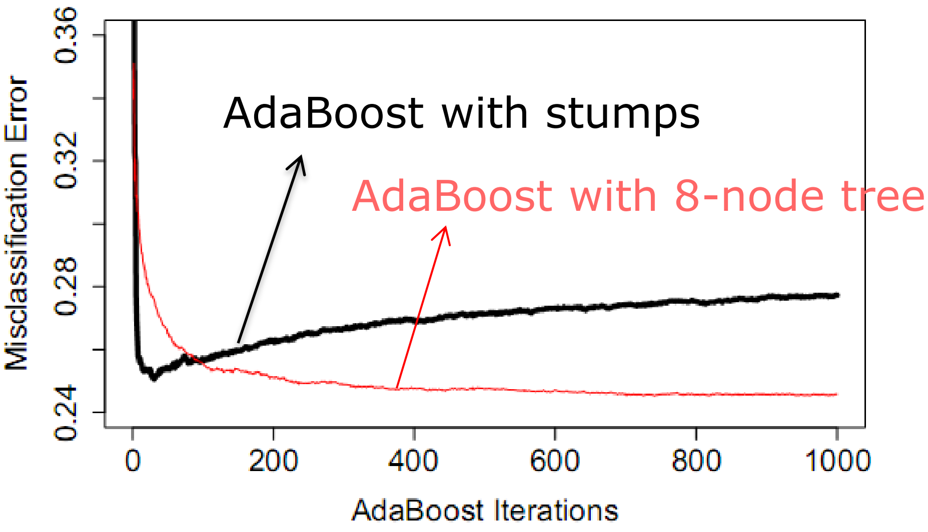
Concerns about the statistical view

However, many aspects of the statistical view have been questioned by empirical results

e.g., in a famous article [Mease & Wyner. Evidence contrary to the statistical view of boosting (with discussions). JMLR, 9:131–201, 2008] it was disclosed that:

Larger-size trees will lead to overfitting because of higher-level interaction [Friedman, Hastie & Tibshirani, Ann. Stat. 2000]

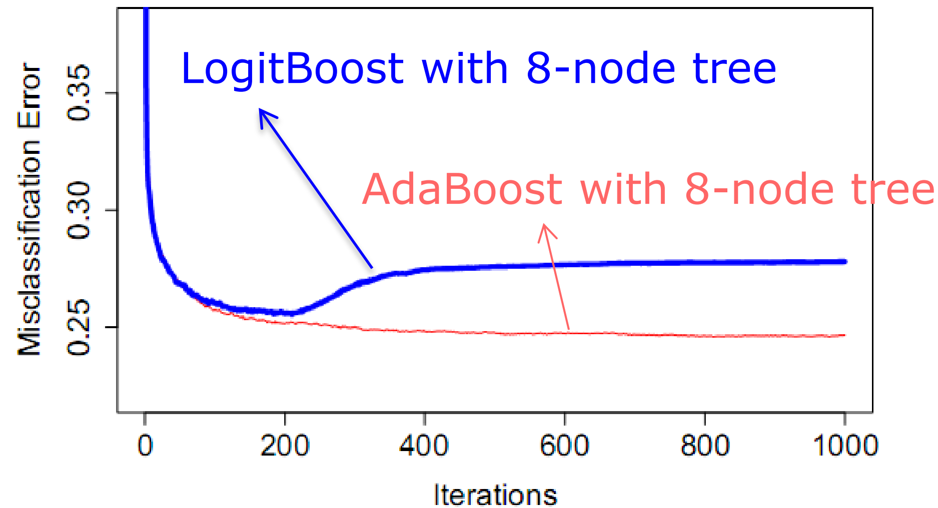
But in practice …



Concerns about the statistical view (con't)

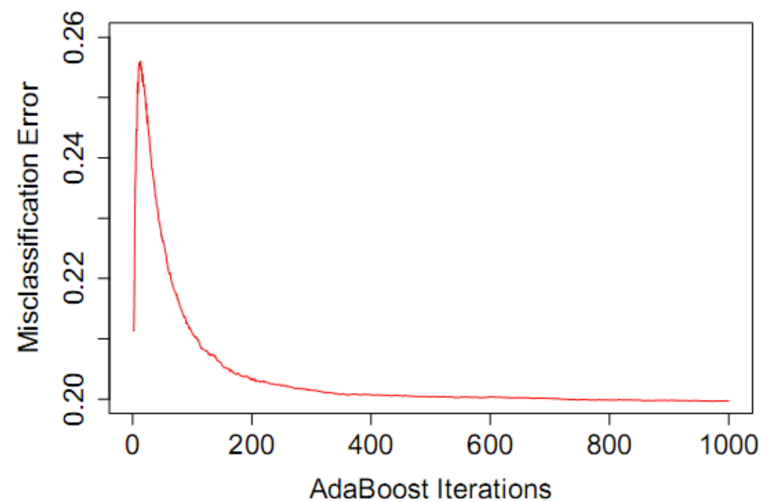
LogitBoost is better than AdaBoost for noisy data [Hastie, Tibshirani & Friedman, "The Elements of Statistical Learning", Springer 2001]

But in practice …



Early stopping can be used to prevent overfitting [Zhang & Yu, Ann. Stat. 2004]

But in practice …



[Mease & Wyner, JMLR 2008]

Major theoretical efforts

- **Margin Theory**

  Started from [Schapire, Freund, Bartlett & Lee, Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of Statistics, 26(5):1651–1686, 1998]

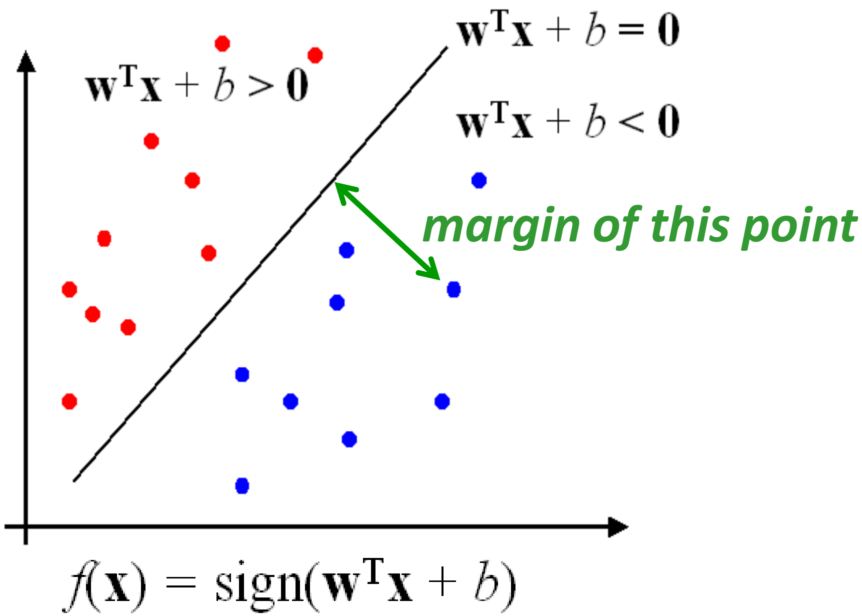
- Statistical View

  Started from [Friedman, Hastie & Tibshirani. Additive logistic regression: A statistical view of boosting (with discussions). Annals of Stati

The biggest issue:
The statistical view did not explain why AdaBoost is resistant to overfitting

The "margin" (间隔)

Binary classification can be viewed as the task of separating classes in a feature space



The bigger the margin,
the higher the predictive confidence

For binary classification, the ground-truth $f(\boldsymbol{x}) \in \{-1, +1\}$

The margin of a single classifier h: $f(\boldsymbol{x})h(\boldsymbol{x})$

For $H(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t h_t(\boldsymbol{x}_t)$

the margin is

$$f(\boldsymbol{x})H(\boldsymbol{x}) = \sum_{t-1}^{T} \alpha_t f(\boldsymbol{x}) h_t(\boldsymbol{x})$$

and the normalized margin:

$$\frac{\sum_{t=1}^{T} \alpha_t f(\boldsymbol{x}) h_t(\boldsymbol{x})}{\sum_{t=1}^{T} \alpha_t}$$

Margin explanation of AdaBoost

Based on the concept of margin, Schapire et al. [1998] proved that, given any threshold $\theta > 0$ of margin over the training data D, with probability at least $1-\delta$, the generalization error of the ensemble $\epsilon_{\mathcal{D}} = P_{\boldsymbol{x}\sim\mathcal{D}}(f(\boldsymbol{x}) \neq H(\boldsymbol{x}))$ is bounded by

$$\begin{aligned}\epsilon_{\mathcal{D}} &\le P_{\boldsymbol{x}\sim D}(f(\boldsymbol{x})H(\boldsymbol{x}) \le \theta) + \tilde{O}\left(\sqrt{\frac{d}{m\theta^2} + \ln\frac{1}{\delta}}\right) \\ &\le 2^T \prod_{t=1}^{T} \sqrt{\epsilon_t^{1-\theta}(1-\epsilon_t)^{1+\theta}} + \tilde{O}\left(\sqrt{\frac{d}{m\theta^2} + \ln\frac{1}{\delta}}\right)\end{aligned}$$

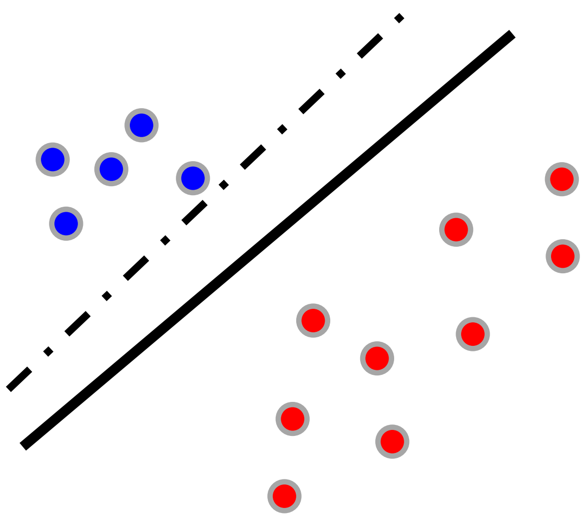
This bound implies that, when other variables are fixed, the larger the margin over the training data, the smaller the generalization error

Margin explanation of AdaBoost (con't)

Why AdaBoost tends to be resistant to overfitting?

the margin theory answers:

Because it is able to increase the ensemble margin even after the training error reaches zero



This explanation is quite intuitive

It receives good support in empirical study

The minimum margin bound

Schapire et al.'s bound depends heavily on the smallest margin, because $P_{\boldsymbol{x}\sim D}(f(\boldsymbol{x})H(\boldsymbol{x}) \leq \theta)$ will be small if the smallest margin is large

Thus, by considering the minimum margin:

$$\varrho = \min_{\boldsymbol{x}\in D} f(\boldsymbol{x})H(\boldsymbol{x})$$

Breiman [Neural Comp. 1999] proved a generalization bound, which is tighter than Schapire et al.'s bound

The two generalization bounds

Theorem 1. *(Schapire et al., 1998) For any $\delta > 0$ and $\theta > 0$, with probability at least $1-\delta$ over the random choice of sample S with size m, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound:*

$$\Pr_D[yf(x) < 0] \le \Pr_S[yf(x) \le \theta] + \underline{O\left(\frac{1}{\sqrt{m}}\left(\frac{\ln m \ln|\mathcal{H}|}{\theta^2} + \ln\frac{1}{\delta}\right)^{1/2}\right)}.$$

$O(\sqrt{\log m \,/m})$

Theorem 2. *(Breiman, 1999) If*

$$\theta = \hat{y}_1 f(\hat{x}_1) > 4\sqrt{\frac{2}{|\mathcal{H}|}} \text{ and } R = \frac{32\ln 2|\mathcal{H}|}{m\theta^2} \le 2m,$$

then, for any $\delta > 0$, with probability at least $1-\delta$ over the random choice of sample S with size m, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound:

$$\Pr_D[yf(x) < 0] \le \underline{R\left(\ln(2m) + \ln\frac{1}{R} + 1\right) + \frac{1}{m}\ln\frac{|\mathcal{H}|}{\delta}}.$$

$O(\log m \,/m)$

The doubt about margin theory

Breiman [Neural Comp. 1999] designed a variant of AdaBoost, the arc-gv algorithm, which directly maximizes the minimum margin

the margin theory would appear to predict that arc-gv should perform better than AdaBoost

However, experiments show that, comparing with AdaBoost:

- arc-gv does produce uniformly larger minimum margin
- the test error increases drastically in almost every case

Thus, Breiman convincingly concluded that **the margin theory was in serious doubt**. This almost sentenced the margin theory to death

7 years later …

Reyzin & Schapire [ICML'06 best paper] found that, amazingly, Breiman had not controlled model complexity well in exps

Breiman controlled the model complexity by using decision trees with a fixed number of leaves

Reyzin & Schapire found that, the trees of arc-gv are generally "deeper" than the trees of AdaBoost

Reyzin & Schapire repeated Brieman's exps using decision stumps with two leaves: arc-gv is with larger minimum margin, but worse margin distribution

R&S claimed that the minimum margin is not crucial, and the *average* or *median margin* is crucial

Experimental results

Tree depth using fixed number of leaves

Margin distribution using decision stumps

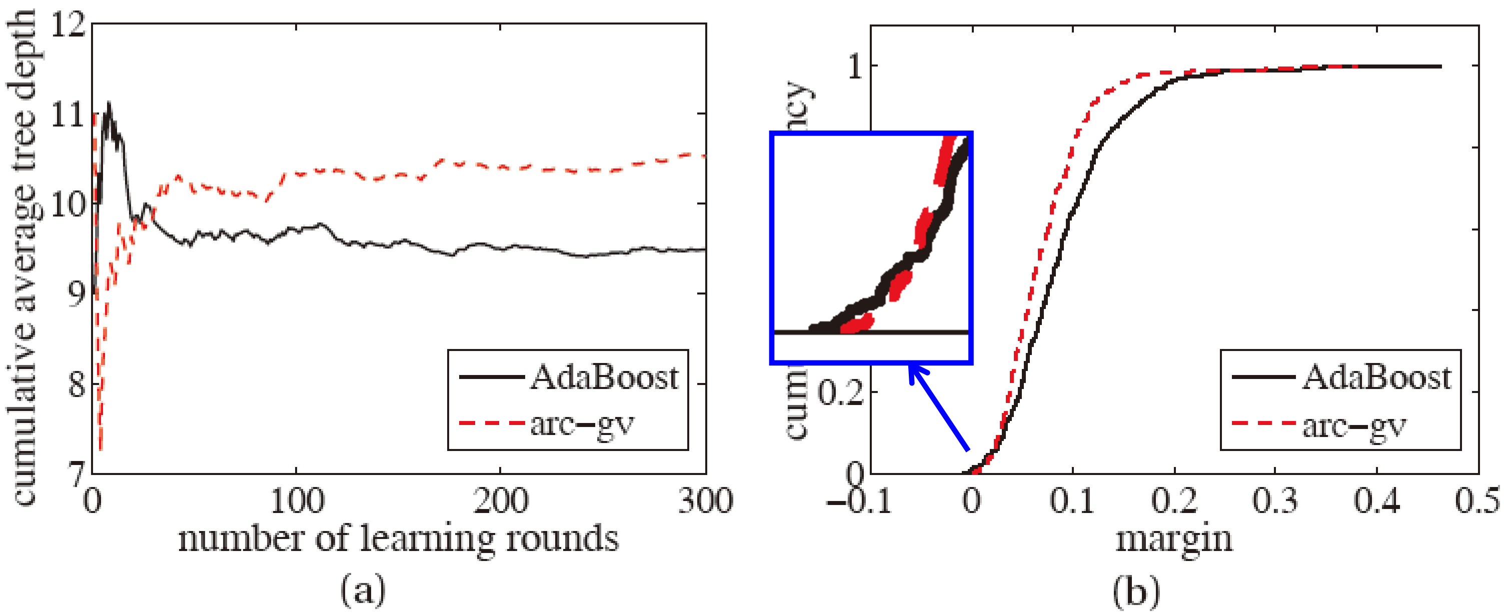


FIGURE 2.8: (a) Tree depth and (b) margin distribution of AdaBoost against arc-gv on the UCI *clean1* data set.

Figure from [Zhou's book, 2012]

Margin theory survive?

Not necessarily …

Breiman's minimum margin bound is tighter

To claim margin distribution is more crucial, we need a margin distribution bound which is even tighter

Equilibrium margin (Emargin) bound

Theorem 3. *(Wang et al., 2011) If $8 < |\mathcal{H}| < \infty$, then for any $\delta > 0$, with probability at least $1-\delta$ over the random choice of the training set S of size $m > 1$, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ such that*

$$q_0 = \Pr_S\left[yf(x) \le \sqrt{8/|\mathcal{H}|}\right] < 1 \tag{3}$$

satisfies the following bound:

$$\Pr_D[yf(x) < 0] \le \frac{\ln|\mathcal{H}|}{m} + \underline{\inf_{q\in\{q_0,q_0+\frac{1}{m},\cdots,1\}} KL^{-1}(q;u[\hat{\theta}(q)])},$$

$O(\log m\ /m)$

Proved to be tighter than Breiman's bound

where

$$u[\hat{\theta}(q)] = \frac{1}{m}\left(\frac{8\ln|\mathcal{H}|}{\hat{\theta}^2(q)}\ln\frac{2m^2}{\ln|\mathcal{H}|} + \ln|\mathcal{H}| + \ln\frac{m}{\delta}\right)$$

and $\hat{\theta}(q) = \sup\left\{\theta \in \left(\sqrt{8/|\mathcal{H}|}, 1\right] : \Pr_S[yf(x) \le \theta] \le q\right\}$. *Also, the Emargin is given by* $\theta^* \in \arg\inf_{q\in\{q_0,q_0+\frac{1}{m},\cdots,1\}} KL^{-1}(q;u[\hat{\theta}(q)])$.

- Considered factors different from Schapire et al. and Breiman's bounds
- No intuition to optimize

First presented in [Wang, Sugiyama, Yang, Zhou, Feng, COLT'08]

The *k*th margin bound

Given a sample S of size m, we define the *kth margin* as the kth smallest margin over sample S, i.e., the kth smallest value in $\{y_i f(x_i), i \in [m]\}$

Theorem 4. *For any $\delta > 0$ and $k \in [m]$, if $\theta = \hat{y}_k f(\hat{x}_k) > \sqrt{8/|\mathcal{H}|}$, then with probability at least $1-\delta$ over the random choice of sample with size m, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound:*

$$\Pr_D[yf(x) < 0] \le \frac{\ln|\mathcal{H}|}{m} + KL^{-1}\left(\frac{k-1}{m}; \frac{q}{m}\right), \qquad (4)$$

where

$$q = \frac{8\ln(2|\mathcal{H}|)}{\theta^2} \ln\frac{2m^2}{\ln|\mathcal{H}|} + \ln|\mathcal{H}| + \ln\frac{m}{\delta}.$$

The minimum margin bound and Emargin bound are special cases of the kth margin bound, both are single-margin bound (not margin distribution bound)

Finally, our margin distribution bound

Theorem 8. *For any $\delta > 0$, with probability at least $1 - \delta$ over the random choice of sample S with size $m \geq 5$, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound:*

$$\Pr_D[yf(x) < 0] \leq \frac{2}{m} + \inf_{\theta \in (0,1]} \left[\Pr_S[yf(x) < \theta] + \underbrace{\frac{7\mu + 3\sqrt{3\mu}}{3m} + \sqrt{\frac{3\mu}{m} \Pr_S[yf(x) < \theta]}}_{O(\log m / m)} \right]$$

where

$$\mu = \frac{8}{\theta^2} \ln m \ln(2|\mathcal{H}|) + \ln \frac{2|\mathcal{H}|}{\delta}.$$

- ✓ Uniformly tighter than Breiman's as well as Schapire et al.' bounds
- ✓ Considers the same factors as Schapire et al. and Breiman

thus, defends the margin theory against Breiman's doubt

[Gao & Zhou, AIJ 2013]

New insight?

Theorem 9. *For any $\delta > 0$, with probability at least $1-\delta$ over the random choice of sample S with size $m \geq 5$, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound:*

$$\Pr_D[yf(x) < 0] \leq \frac{1}{m^{50}} + \inf_{\theta\in(0,1]} \boxed{\Pr_S[yf(x) < \theta] + m^{-2/(1-\boxed{E_S^2[yf(x)]}+\theta/9)} + \frac{3\sqrt{\mu}}{m^{3/2}} + \frac{7\mu}{3m} + \sqrt{\frac{3\mu}{m}\boxed{\hat{\mathcal{I}}(\theta)}}}$$

related to average margin

$O(\log m/m)$

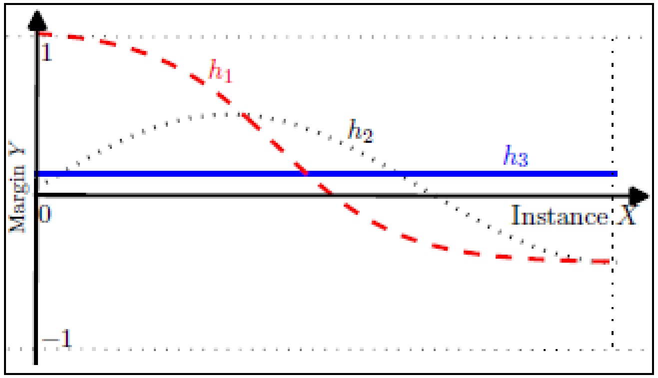
where

$$\mu = 144\ln m \ln(2|\mathcal{H}|)/\theta^2 + \ln(2|\mathcal{H}|/\delta),$$
$$\hat{\mathcal{I}}(\theta) = \Pr_S[yf(x) < \theta]\Pr_S[yf(x) \geq 2\theta/3].$$

related to margin variance

We should pay attention to not only the average margin, but also the margin variance !

In practice



Margin variance really important

Figure from [Gao & Zhou, AIJ 2013]

Figure 1: Each curve represents a voting classifier. The X-axis and Y-axis denote example and margin, respectively, and uniform distribution is assumed on the example space. The voting classifiers h_1, h_2 and h_3 have the same average margin but with different generalization error rates: 1/2, 1/3 and 0.

[Shivaswamy & Jebara, NIPS 2011] tried to design new boosting algorithms by maximizing average margin and minimizing margin variance simultaneously, and the results are encouraging

Long march of margin theory for AdaBoost

- 1989, [Kearns & Valiant], open problem
- 1990, [Schapire], proof by construction, the first Boosting algorithm
- 1993, [Freund], another impractical boosting algorithm by voting
- 1995/97, [Freund & Schapire], AdaBoost

- 1998, [Schapire, Freund, Bartlett & Lee], Margin theory
- 1999, [Breiman], serious doubt by minimum margin bound
- 2006, [Reyzin & Schapire], finding the model complexity issue in exps, emphasizing the importance of margin distribution
- 2008, [Wang, Sugiyama, Yang, Zhou & Feng], Emargin bound, believed to be a margin distribution bound
- 2013, [Gao & Zhou], a real margin distribution bound, sheding new insight ; margin theory defensed

Joint work with my student

W. Gao and Z.-H. Zhou. On the doubt about margin explanation of boosting. Artificial Intelligence, 2013, 203: 1-18.
(arXiv:1009.3613, Sept.2010)

Wei Gao
(高尉)

An easy-to-read article:

Z.-H. Zhou. Large margin distribution learning. ANNPR 2014, Montreal, Canada, LNAI 8774, pp.1-11 (keynote article)

Thanks!