A New Margin Bound

John Langford (and John Shawe-Taylor)

Problem: Classification

Given: m examples, $S = (x, y)^m \sim D^m$, $y \in \{-1, 1\}$, (D unknown)

Find: $h : X \rightarrow Y$

Goal: $e_D(h) = \text{Pr}_{x,y \sim D}(h(x) \neq y)$ small

Example: Predict cloudy or sunny given barometric pressure.

Problem: Classification

Given: m examples, $S = (x, y)^m \sim D^m$, $y \in \{-1, 1\}$, (D unknown)

Find: $h : X \rightarrow Y$

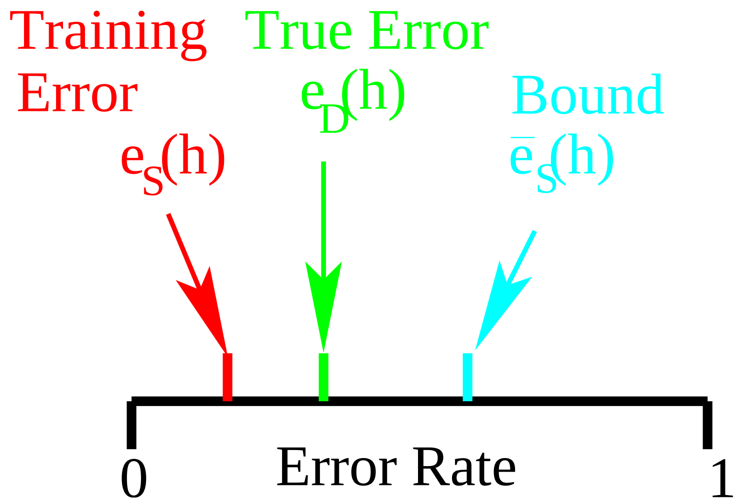
Goal: $e_D(h) = \text{Pr}_{x,y \sim D}(h(x) \neq y)$ small

Example: Predict cloudy or sunny given barometric pressure.

IBM example: Predict profitable customer or not.

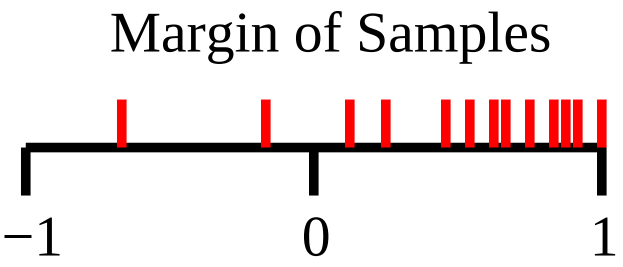
Bound

Bound: high confidence upper bound on $e_D(h)$.



Margin Bound

Specialize to $h(x) = \text{sign}\left(\sum_i w_i x_i\right)$



Margin $= \gamma(x, y) = y \sum_i w_i x_i$

- Margin Background
- Stochastic Margin Bound
- Proof Method
- Derandomized Margin Bound

Why Specialize?

Answer: Very common structure

1. Boosting

2. Bagging

3. Support Vector Machine

4. Maximum Entropy Classifier

5. Bayes-optimal Classifier

Why Specialize?

Answer: Empirically, averaging is a win.

Can we describe that mathematically?

Previous work

Two margin bound flavors:

1. "l_1/l_∞" $1 \geq \sum_i w_i$, $1 \geq \max_{i,x} |x_i|$ (SFBL98, LSM00)

2. "l_2/l_2" $1 \geq \sum_i w_i^2$, $1 \geq \max_x \sum_i x_i^2$ (S-TBWA98, B98)

Rob Schapire says: "you do l_2/l_2".

Limitations of previous work

1. Requires $\forall x, y: \ \gamma(x, y) > 0$ (S-TBWA98)

2. Requires $m = |S| > 10^6$ or more examples (SFBL98, S-TBWA98, B98)

3. Messy proof (SFBL98, S-TBWA98, B98, LSM00)

4. Not dependent on margin distribution (SFBL98, S-TBWA98, B98, LSM00)

This Bound

1. Simple proof

2. Uses full margin distribution.

3. Useful for $m = |S| < 10^2$

4. Margin can be negative.

- Margin Background
- Stochastic Margin Bound
- Proof Method
- Derandomized Margin Bound

The margin bound

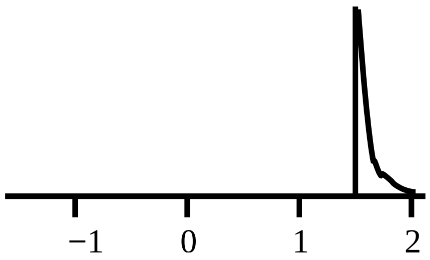
1. Exists large set of stochastic classifiers with error rate dependent upon margin

2. Derandomize large set of stochastic classifiers

Step 1: large set of stochastic classifiers

$\gamma > 0$ = margin of interest

$\hat{e}_\gamma(h) = \Pr_S(y \sum_i w_i x_i \leq \gamma)$

$\bar{F}(x) = \Pr_{z \sim N(0,1)}(z > x) =$ 

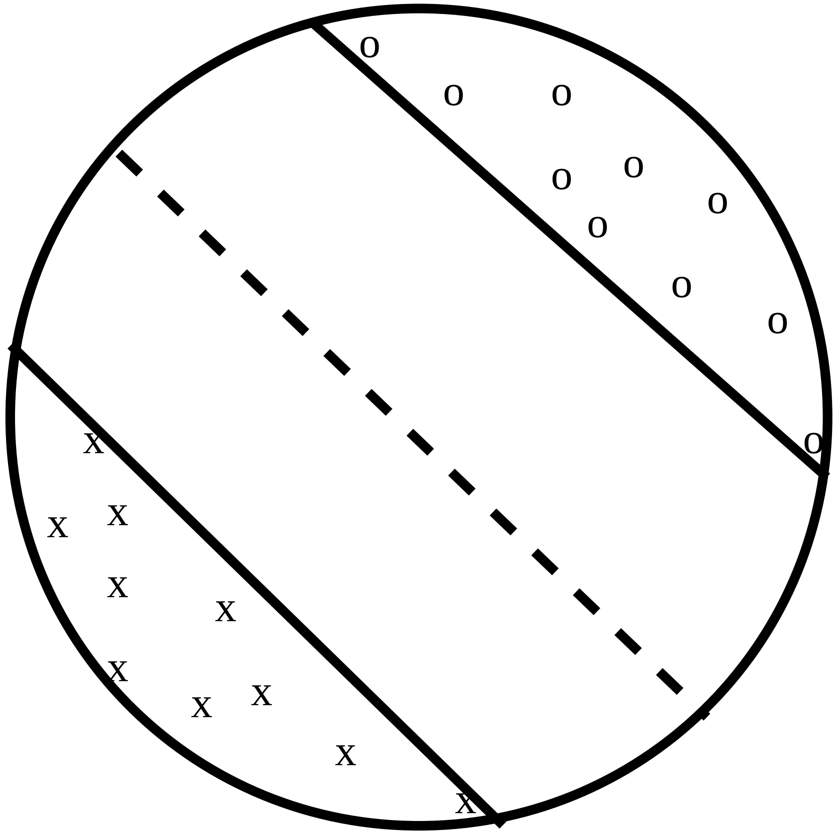
δ = probability of bound failure

$\epsilon > 0$ = extra error of interest

$Q(\vec{w}, \gamma, \epsilon)$ = distribution over classifiers

$$\Pr_{S \sim D^m}\left(\mathrm{KL}\left(\hat{e}_\gamma(h) + \epsilon || e_{Q(\vec{w},\gamma,\epsilon)}(h)\right) \leq \frac{\ln \frac{1}{\bar{F}\left(\frac{\bar{F}^{-1}(\epsilon)}{\gamma}\right)} + \ln \frac{m+1}{\delta}}{m} \right) \geq 1 - \delta$$

What does this mean?



What does this mean?

$\bar{F}(x) \simeq \frac{e^{-x^2/2}}{\sqrt{2\pi}}\frac{1}{x}$ for $x > 2$

assume $m = 100$, $\gamma = .5$, $\hat{e}_\gamma(h) = 0$, $\epsilon = 0.01$, $\delta = .1$

$$\Rightarrow e_{Q(\vec{w},\gamma,\epsilon)}(h) \leq \frac{\ln \frac{1}{\bar{F}\left(\frac{2}{\gamma}\right)} + \ln \frac{m}{\delta}}{m}$$

$$\Rightarrow e_{Q(\vec{w},\gamma,\epsilon)}(h) \leq \frac{\frac{2}{\gamma^2} + \ln \frac{2}{\gamma} + 0.5 \ln 2\pi + \ln \frac{m}{\delta}}{m}$$

$$\Rightarrow e_{Q(\vec{w},\gamma,\epsilon)}(h) \leq \frac{8+1.39+0.92+6.92}{100}$$

$$\Rightarrow e_{Q(\vec{w},\gamma,\epsilon)}(h) \leq \frac{17.2}{100}$$

Full margin distribution dependence

More general $\Rightarrow$ simpler!

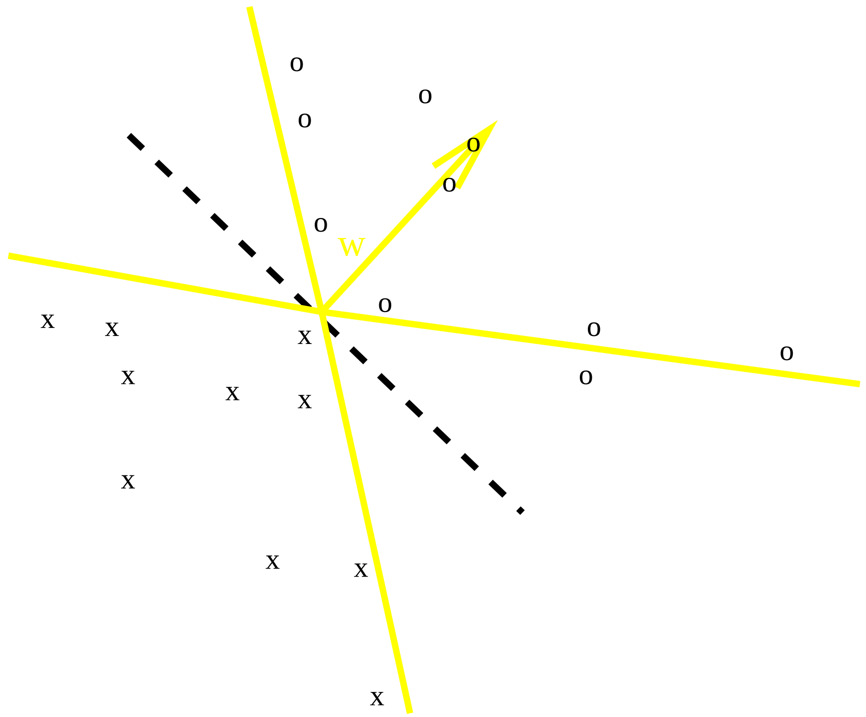
$$\vec{w} = w\vec{e}_{||}$$

$$\vec{x} = x_{||}\vec{e}_{||} + x_{\top}\vec{e}_{\top}$$

Choose $\mu > 0$ so $E_{x,y\sim S}\bar{F}\left(\frac{yx_{||}}{x_{\top}}\mu\right) = \epsilon$, then:

$$\Pr_{S\sim D^m}\left(\forall \vec{w}, \mu, \epsilon : \ \ \mathrm{KL}\left(\epsilon || e_{Q(\vec{w},\mu,\epsilon)}(h)\right) \leq \frac{\ln \frac{1}{\bar{F}(\mu)} + \ln \frac{m+1}{\delta}}{m}\right) \geq 1 - \delta$$

$\frac{x_{||}}{x_{\top}}$ important *not* $\sum_i x_i^2 < 1$



Errors are allowed: only expectation is important

- Margin Background
- Stochastic Margin Bound
- Proof Method
- Derandomized Margin Bound

Proof: Start with PAC-Bayes bound (McAllester 99)

$$e_{S,Q} = E_{h \sim Q} e_S(h)$$

$$e_{D,Q} = E_{h \sim Q} e_D(h)$$

$$\forall P \sim h$$

$$\Pr_{S \sim D^m} \left(\forall Q \sim h \ \mathsf{KL}(e_{S,Q} || e_{D,Q}) \leq \frac{\mathsf{KL}(Q||P) + \ln \frac{m+1}{\delta}}{m} \right) \geq 1 - \delta$$

$P = N(0,1) \times N(0,1) \times ...$ on every weight in average

$Q = Q_{||} \times Q_{\top}$

$Q_{\top} = N(0,1) \times N(0,1) \times ...$

$Q_{||}$ = rectified gaussian tail starting at μ

$\Rightarrow \mathrm{KL}(Q||P) = \mathrm{KL}(Q_{||}||N(0,1)) = \ln \frac{1}{\bar{F}(\mu)}$

Done.

One detail to check: $e_{S,Q} \leq \epsilon$

$$e_{S,Q} = E_S E_Q I(h(x) \neq y)$$

$$= E_{x,y\sim S} E_{\hat{w}_{||},\hat{w}_{\top}\sim Q} I(\mathrm{sign}(x_{||}\hat{w}_{||} + x_{\top}\hat{w}_{\top}) \neq y)$$

$$\leq E_{x,y\sim S} E_{\hat{w}_{||},\hat{w}_{\top}\sim Q} I(\mathrm{sign}(x_{||}\mu + x_{\top}\hat{w}_{\top}) \neq y)$$

$$= E_{x,y\sim S} \Pr\left(\hat{w}_{\top} > \frac{yx_{||}}{x_{\top}}\mu\right)$$

$$= E_{x,y\sim S} \bar{F}\left(\frac{yx_{||}}{x_{\top}}\mu\right)$$

$$= \epsilon$$

- Margin Background
- Stochastic Margin Bound
- Proof Method
- Derandomized Margin Bound

Derandomization methods

1. Simple: Pigeonhole principle

2. Complicated: run proof of stochastic margin bound backwards.

Simple better than complicated when stochastic error rate is small.

Pigeonhole Principle

1. Average of all random classifiers $=$ original

2. Error of average classifier $\leq 2\times$ error of stochastic classifier

	x^1	x^2	x^3	x^4
h_1	X			
h_2	X			
h_3		X		
h_4		X		

Complicated

$\gamma \rightarrow \frac{\gamma}{2}$

Stochastic Margin argument $\Rightarrow$ bound at $e_{\frac{\gamma}{2},Q}$ using $e_\gamma(h)$.

Same argument $\Rightarrow$ bound on $e_D(h)$.

(some trickiness pointed out and solved by David McAllester)

Things Learned

1. $\frac{x_{||}}{x_{\top}}$ is important rather than $||\vec{x}|| < 1$.

2. PAC-Bayes core instead of VC core $\Rightarrow$ tighter math.

3. Scales like $O\left(\frac{\frac{1}{\gamma^2}+\ln\frac{1}{\gamma}+\ln m+\ln\frac{1}{\delta}}{m}\right)$ to $O\left(\sqrt{\frac{\frac{1}{\gamma^2}+\ln\frac{1}{\gamma}+\ln m+\ln\frac{1}{\delta}}{m}}\right)$ with small constants (no $\frac{1}{\gamma^2}\ln^2 m$ dependence, no large constants)

4. no explicit dependence on feature space size (some implicit dependence)

Future work

Empirical: What happens if you optimize according to the bound? (Ashutosh Garg, others). Indirect evidence suggests better than SVM.

Theoretical: Can the simple and complicated derandomizations be unified?

Optimization

$\text{argmin}_{\vec{w},\mu}\mu$

s.t. $E_{x,y\sim S}\bar{F}(\frac{yx_{||}}{x_{\top}}\mu) \leq \epsilon$

where $x_{||} = \frac{\sum_i x_i w_i}{||w||}$ and $x_{\top} = \sqrt{||x||^2 - x_{||}^2}$.