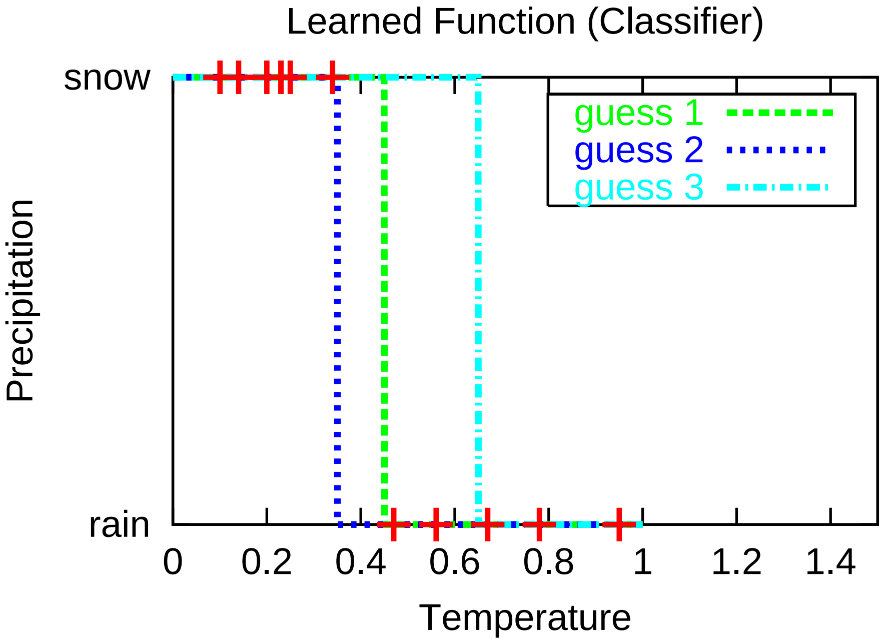
The Classification Problem:

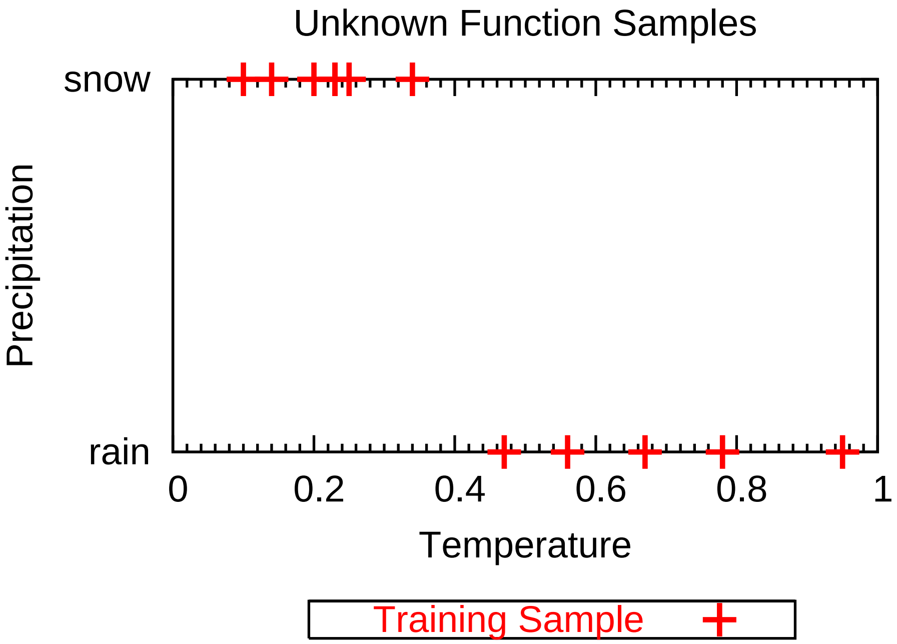
1. Get (input, class label) pairs from unknown function
2. Find the unknown function
3. ... to minimize errors on future inputs

Practical Prediction Theory

John Langford

IBM Research

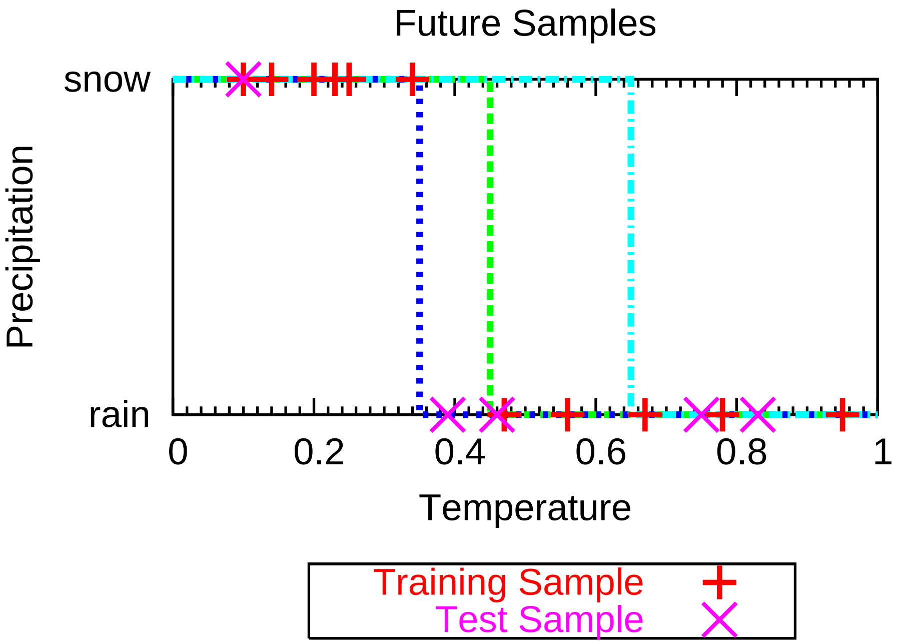

Learned Function (Classifier)
snow
rain
Precipitation
Temperature
0
0.2
0.4
0.6
0.8
1
1.2
1.4
guess 1
guess 2
guess 3


Unknown Function Samples
snow
rain
Precipitation
Temperature
0
0.2
0.4
0.6
0.8
1
Training Sample

Learning = Prediction ability

- We can't expect any prediction ability, in general.
- We can expect prediction ability, if examples come independently, sometimes.

Here we study prediction ability, assuming indepedence.



Better Methods for Learning & Verification

Standard technique:

1. Divide samples into train and test set
2. Train on train set
3. Test on test set

We can do better.

Why study prediction theory?

1. Better methods for learning and verifying predictive ability
2. To gain insight into learning.

Outline

1. The Basic Model
2. The Test Set Bound
3. Occam's Razor Bound
4. PAC-Bayes Bound
5. Sample Compression Bound

To gain insight into learning

1. Overfitting: sample complexity quantifies overfitting.
2. Learning algorithm design: What is a good pruning criteria? Why are large margins good? What other algorithms are likely to yield good results?

Model: Basic Assumption

All samples are drawn independently from some unknown distribution $D(x, y)$.

$S = (x, y)^m \sim D^m$ is a sample set.

Model: Definitions

X = input space

$Y = \{0, 1\}$ = output space

$c : X \rightarrow Y$ = classifier

Model: Derived quantities

The thing we want to know:

$$c_D \equiv \Pr_{x,y\sim D}(c(x) \neq y) = \text{true error}$$

The thing we have:

$$\hat{c}_S \equiv \Pr_{x,y\sim S}(c(x) \neq y) = \frac{1}{m}\sum_{i=1}^{m} I\left[c(x) \neq y\right]$$

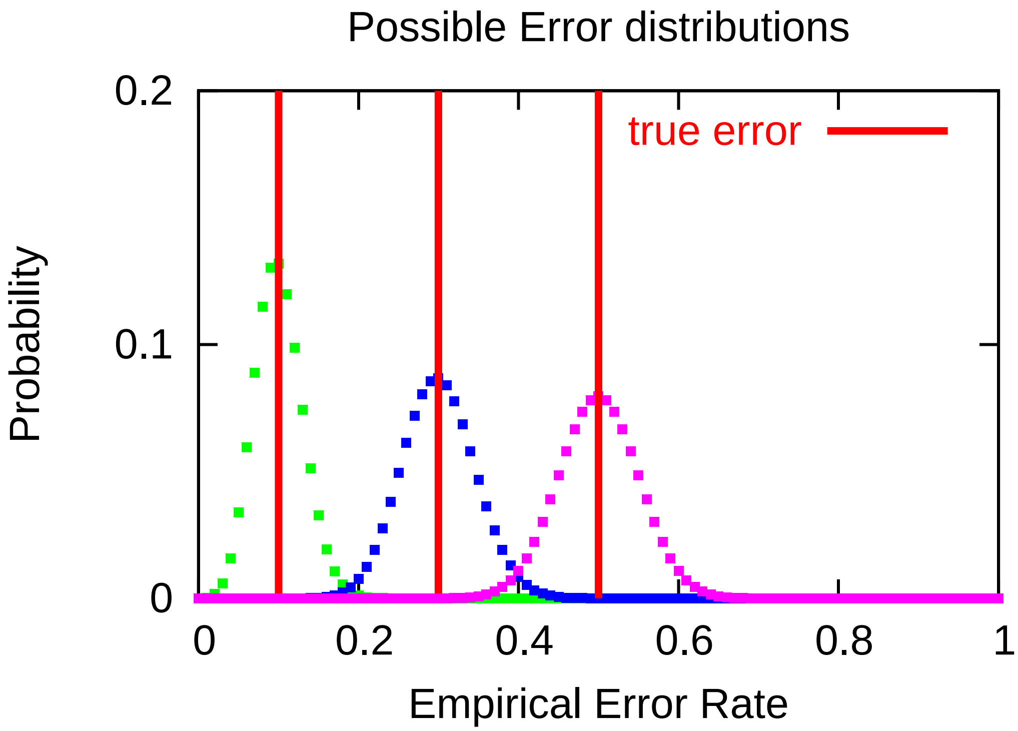
$=$ “train error”, “test error”, or “observed error”, depending on context.

(note: we identify the set S with the uniform distribution on S)

Model: Derived quantities

The thing we want to know:

$$c_D \equiv \Pr_{x,y\sim D}(c(x) \neq y) = \text{true error}$$



Model: Basic Observations

Q: What is the distribution of $\hat{c}_S$?

A: A Binomial.

$$\Pr_{S\sim D^m}\left(\hat{c}_S = \frac{k}{m}\middle| c_D\right) = \binom{m}{k} c_D^k (1-c_D)^{m-k}$$

= probability of k heads (errors) in m flips of a coin with bias c_D.

Model: basic quantities

Need confidence intervals ⇒ use the pivot of the cumulative instead

$$\overline{\mathrm{Bin}}\left(\frac{k}{m}, \delta\right) = \max\left\{p : \mathrm{Bin}\left(\frac{k}{m}, p\right) \geq \delta\right\}$$

= the largest true error such that the probability of observing k or fewer “heads” (errors) is at least δ.

Model: basic quantities

We use the cumulative:

$$\begin{aligned} \mathrm{Bin}\left(\tfrac{k}{m}, c_D\right) &= \quad \mathrm{Pr}_{S\sim D^m}\left(\hat{c}_S \leq \tfrac{k}{m} \middle| c_D\right) \\ &= \textstyle\sum_{i=0}^{k} \left(\begin{array}{c} m \\ i \end{array}\right) c_D^i (1-c_D)^{m-i} \end{aligned}$$

= probability of observing k or fewer “heads” (errors) with m coins.

Test Set Bound: Setting

Standard technique:

1. Cut the data into train set and test set

2. Train on the train set

3. Test on the test set

What does Sample Complexity say about this method?

Outline

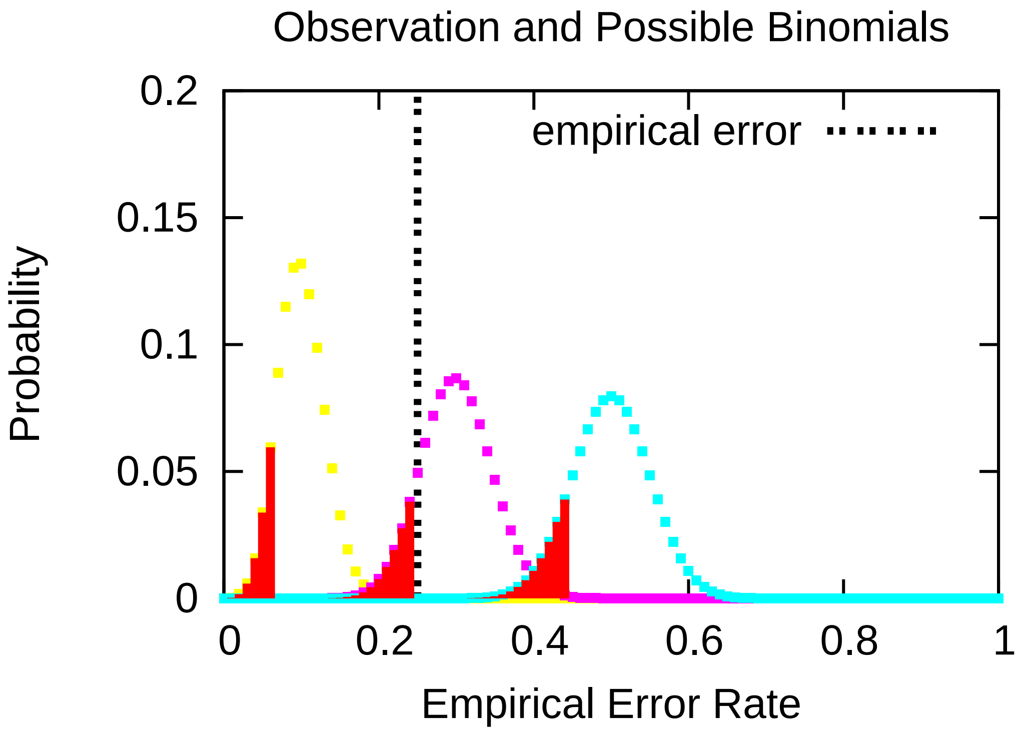
1. The Basic Model

2. The Test Set Bound

3. Occam's Razor Bound

4. PAC-Bayes Bound

5. Sample Compression Bound



Test Set Bound: Theorem

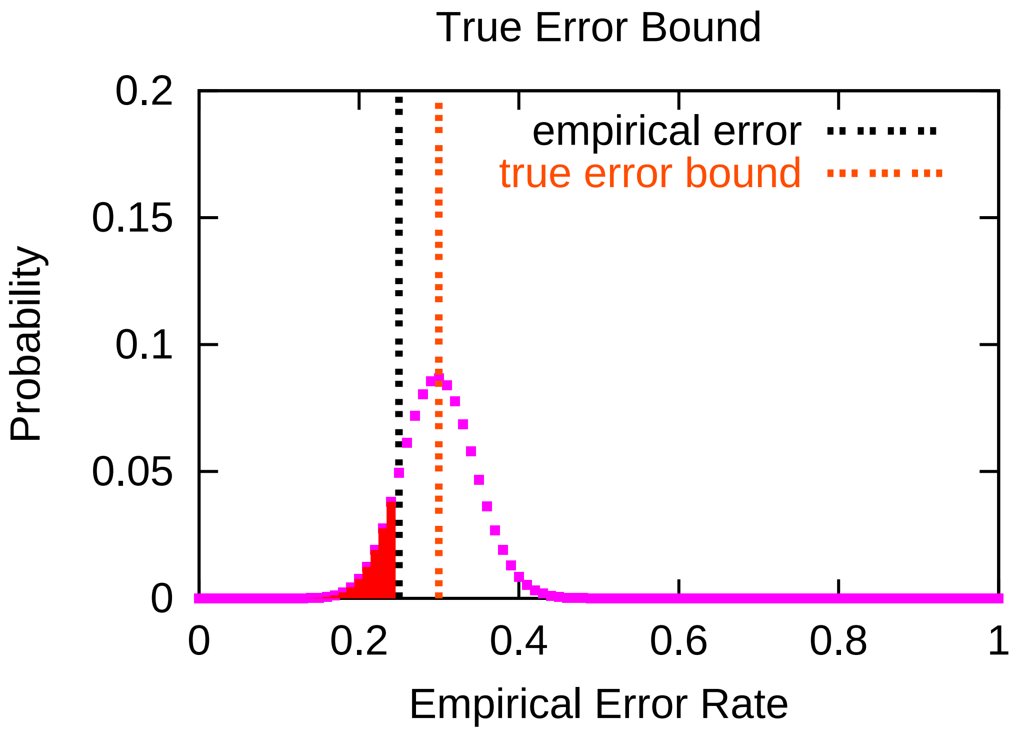
Theorem: (Test Set Bound) For all classifiers c, for all D, for all $\delta \in (0, 1]$:

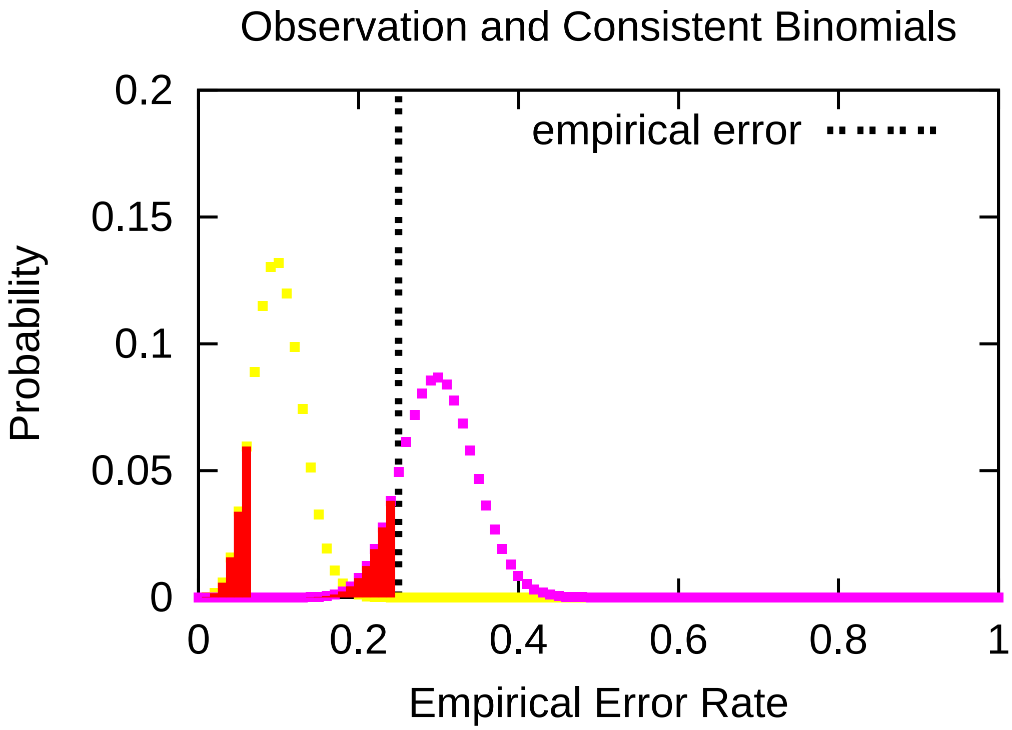
$$\Pr_{S \sim D^m}\left(c_D \leq \overline{\text{Bin}}\left(\hat{c}_S, \delta\right)\right) \geq 1 - \delta$$

World's easiest proof: (by contradiction).

Assume $\text{Bin}\left(\frac{k}{m}, c_D\right) \geq \delta$ (which is true with probability $1 - \delta$).

Then by definition, $\overline{\text{Bin}}\left(\hat{c}_S, \delta\right) \geq c_D$


True Error Bound
Probability
Empirical Error Rate
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1
empirical error
true error bound


Observation and Consistent Binomials
Probability
Empirical Error Rate
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1
empirical error

What does Test Set Bound mean?

Corollary: For all classifiers c, for all D, for all $\delta \in (0,1]$:

$$\Pr_{S \sim D^m}\left(c_D \leq \hat{c}_S + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}\right) \geq 1 - \delta$$

Proof: Use the Chernoff approximation,

$$\overline{\text{Bin}}\left(\hat{c}_S, \delta\right) \leq \hat{c}_S + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$$

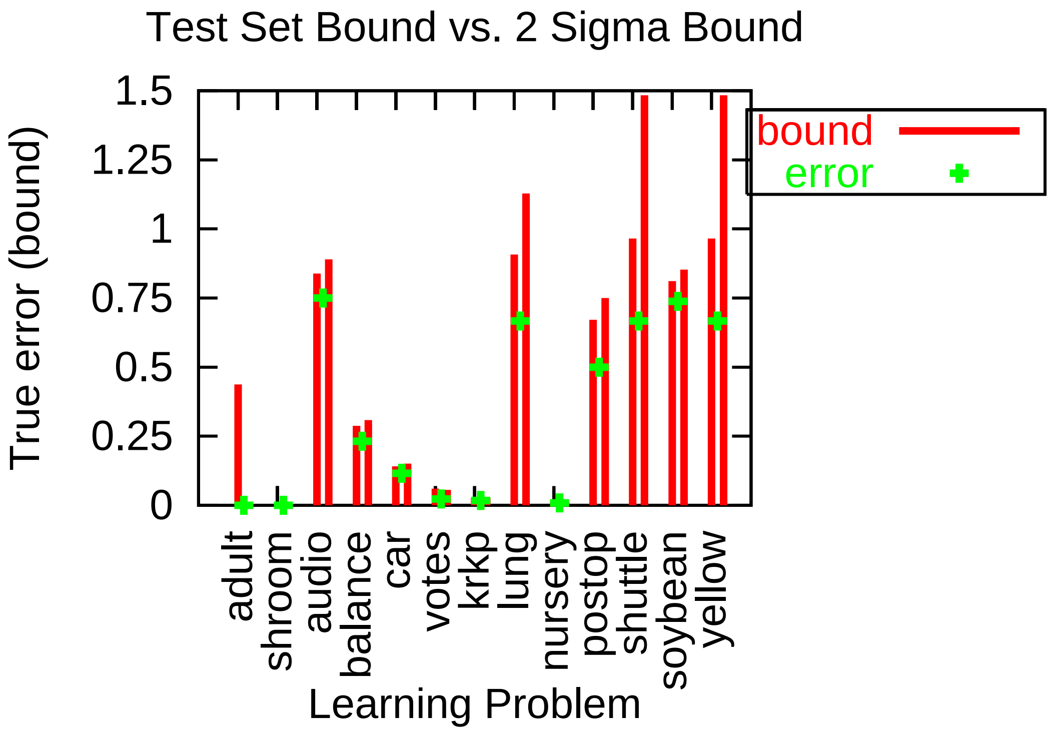
Note: NOT tight when $\hat{c}_S$ near 0 (our goal!)

Test Set Bound Notes

Perfectly tight: There exist true error rates achieving the bound

Lower bound of the same form.

Primary use: verification of succesful learning



Test Set Bound Comparison: Empirical "confidence" intervals

k = number of test errors, m = number of examples

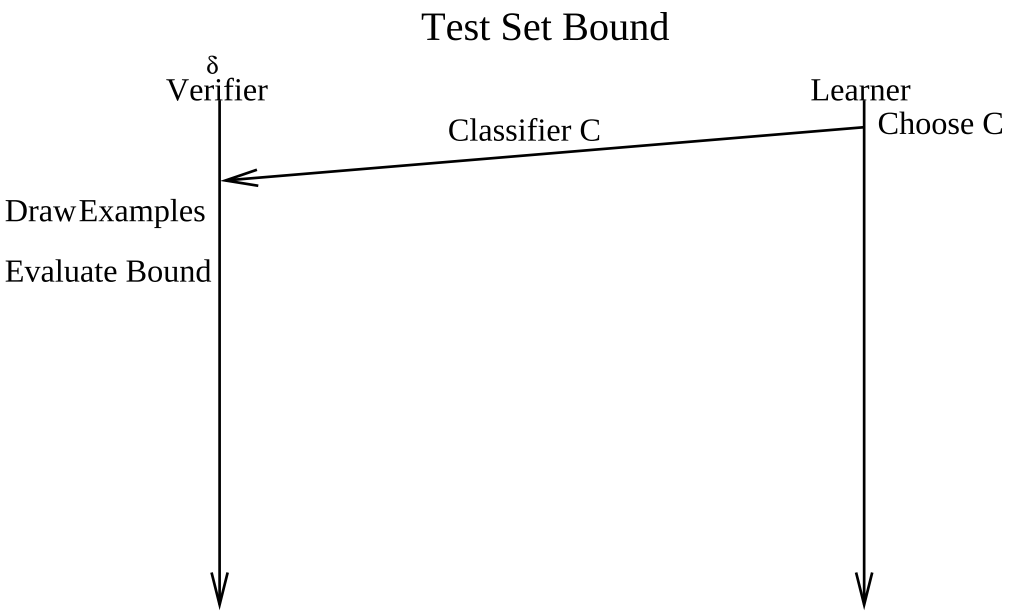
$\mu = \frac{k}{m}$

$\sigma^2 = \frac{1}{m-1} \sum_{i=1}^{m} (\mu - I\left[c(x_i) \neq y_i\right])^2$

pick bound $= \frac{k}{m} + 2\sigma$

How do they compare?

Interpretation: Interactive Proof of Learning



Test Set Bound

δ

Verifier

Learner

Classifier C

Choose C

Draw Examples

Evaluate Bound

Test Set Bound vs Empirical Confidence Interval

1. empirical confidence intervals are sometimes pessimistic

2. empirical confidence intervals are sometimes optimistic

3. the test set bound always works

Occam's Razor bounds

- Sometimes the holdout set is *critical* for learning.
- Sometimes we want bounds to guide learning

⇒Train set bounds

Occam's Razor bound is the simplest train set bound.

Outline

1. The Basic Model
2. The Test Set Bound
3. Occam's Razor Bound
4. PAC-Bayes Bound
5. Sample Compression Bound

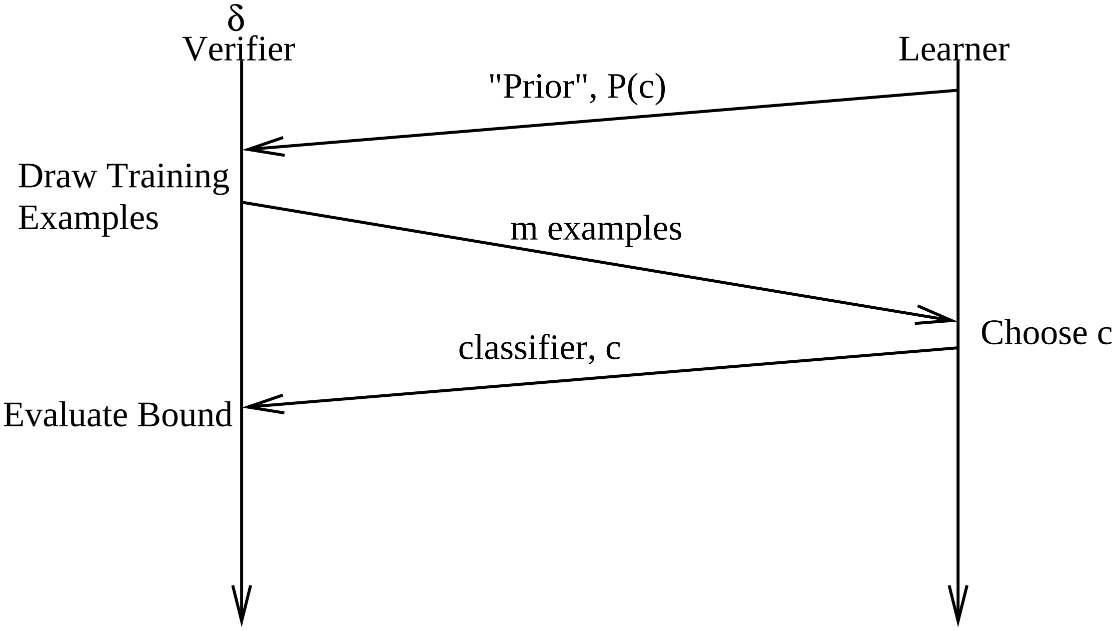
Occam's Razor Bound

Theorem: (Occam's Razor Bound) For all "priors" $P(c)$ over the classifiers c, for all D, for all $\delta \in (0,1]$:

$$\Pr_{S \sim D^m}\left(\exists c:\ c_D \leq \overline{\mathrm{Bin}}\left(\hat{c}_S, \delta P\left(c\right)\right)\right) \geq 1 - \delta$$

Compare with test set bound: $\delta \to \delta P(c)$.

Occam's Razor Bound Protocol



Occam's Razor Bound: Proof

Test set bound ⇒

$$\forall c \quad \Pr_{S\sim D^m}\Big(c_D \leq \overline{\mathsf{Bin}}\left(\hat{c}_S, \delta P(c)\right)\Big) \geq 1 - \delta p(c)$$

Negate to get:

$$\forall c \quad \Pr_{S\sim D^m}\Big(c_D > \overline{\mathsf{Bin}}\left(\hat{c}_S, \delta P(c)\right)\Big) < \delta p(c)$$

Occam's Razor Bound: Proof

Test set bound ⇒

$$\forall c \quad \Pr_{S\sim D^m}\Big(c_D \leq \overline{\mathsf{Bin}}\left(\hat{c}_S, \delta P(c)\right)\Big) \geq 1 - \delta p(c)$$

Occam's Razor Bound: Proof

Test set bound ⇒

$$\forall c \quad \Pr_{S\sim D^m}\left(c_D \leq \overline{\text{Bin}}\left(\hat{c}_S, \delta P(c)\right)\right) \geq 1 - \delta p(c)$$

Negate to get:

$$\forall c \quad \Pr_{S\sim D^m}\left(c_D > \overline{\text{Bin}}\left(\hat{c}_S, \delta P(c)\right)\right) < \delta p(c)$$

Apply union bound: $\Pr(A \text{ or } B) \leq \Pr(A) + \Pr(B)$ repeatedly.

$$\Pr_{S\sim D^m}\left(\exists c: \quad c_D > \overline{\text{Bin}}\left(\hat{c}_S, \delta P(c)\right)\right) < \sum_c \delta P(c) = \delta$$

Negate again to get proof.

Next: Graphical proof

Occam's Razor Bound: Proof

Test set bound ⇒

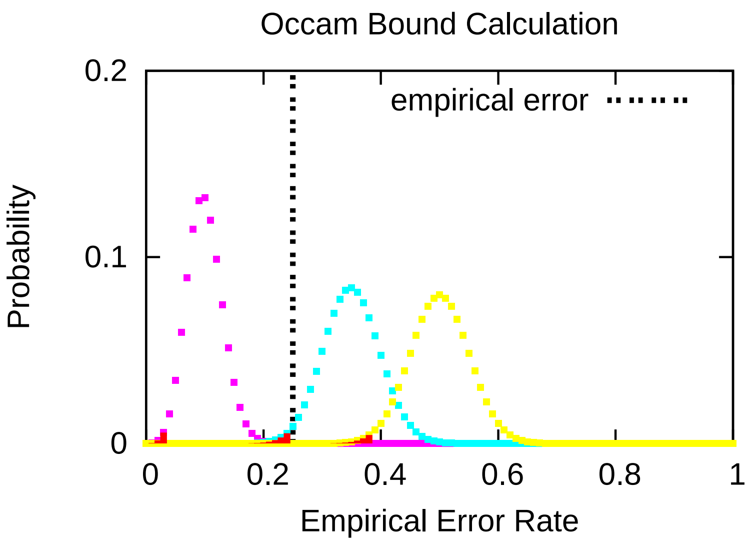
$$\forall c \quad \Pr_{S\sim D^m}\left(c_D \leq \overline{\text{Bin}}\left(\hat{c}_S, \delta P(c)\right)\right) \geq 1 - \delta p(c)$$

Negate to get:

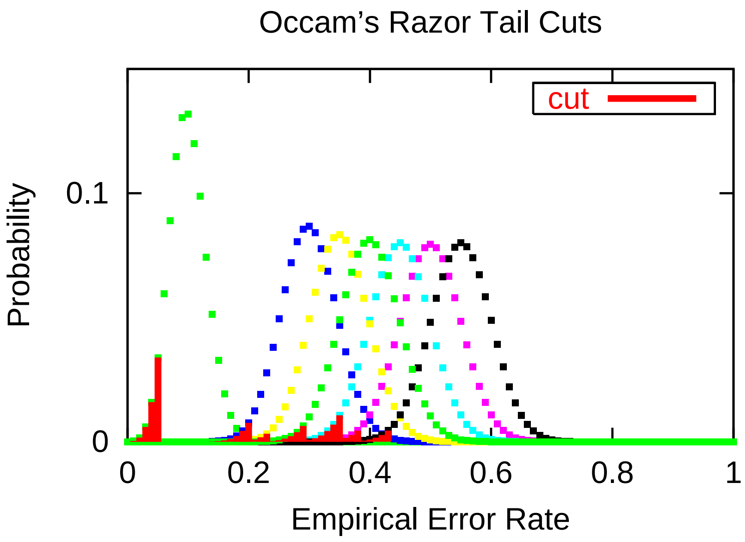
$$\forall c \quad \Pr_{S\sim D^m}\left(c_D > \overline{\text{Bin}}\left(\hat{c}_S, \delta P(c)\right)\right) < \delta p(c)$$

Apply union bound: $\Pr(A \text{ or } B) \leq \Pr(A) + \Pr(B)$ repeatedly.

$$\Pr_{S\sim D^m}\left(\exists c: \quad c_D > \overline{\text{Bin}}\left(\hat{c}_S, \delta P(c)\right)\right) < \sum_c \delta P(c) = \delta$$



The chosen classifier has an unknown true error rate.

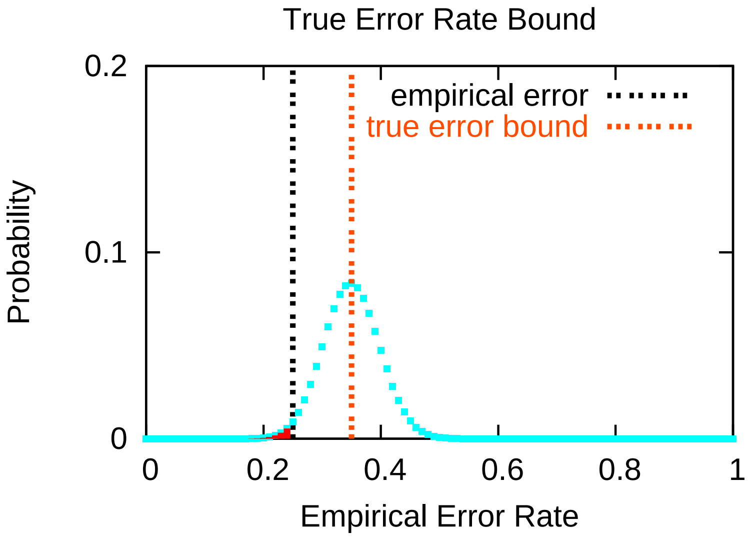


Each classifier is a Binomial with a different size tail cut.

With high probability no error falls in any tail.

Occam's Razor Bound Results Decision Trees

- ID3 decision tree + pruning
- probability of failure = $\delta = 0.1$
- Discrete problems from UCI database of Machine Learning problems.
- 100% of data used for training set bounds
- 80%/20% Train/Test split for test set bounds
- Minimal selection bias



Bound = the largest true error rate for which the observation is not in the tail.

Outline

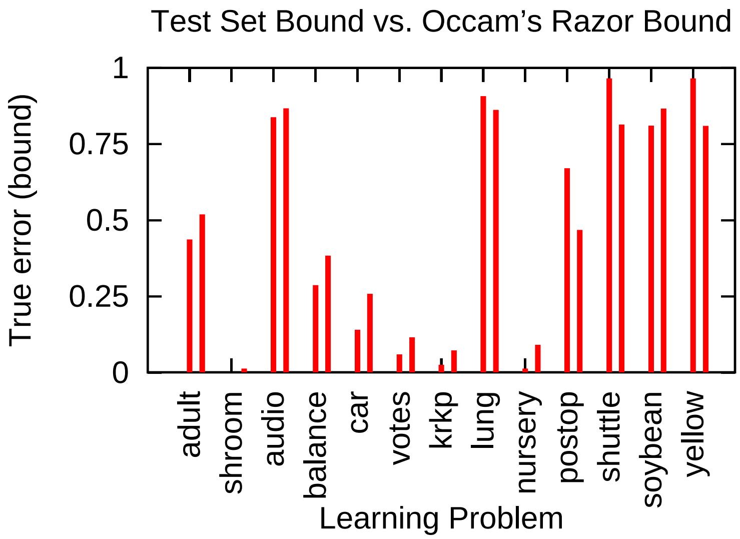
1. The Basic Model

2. The Test Set Bound

3. Occam's Razor Bound

4. PAC-Bayes Bound

5. Sample Compression Bound



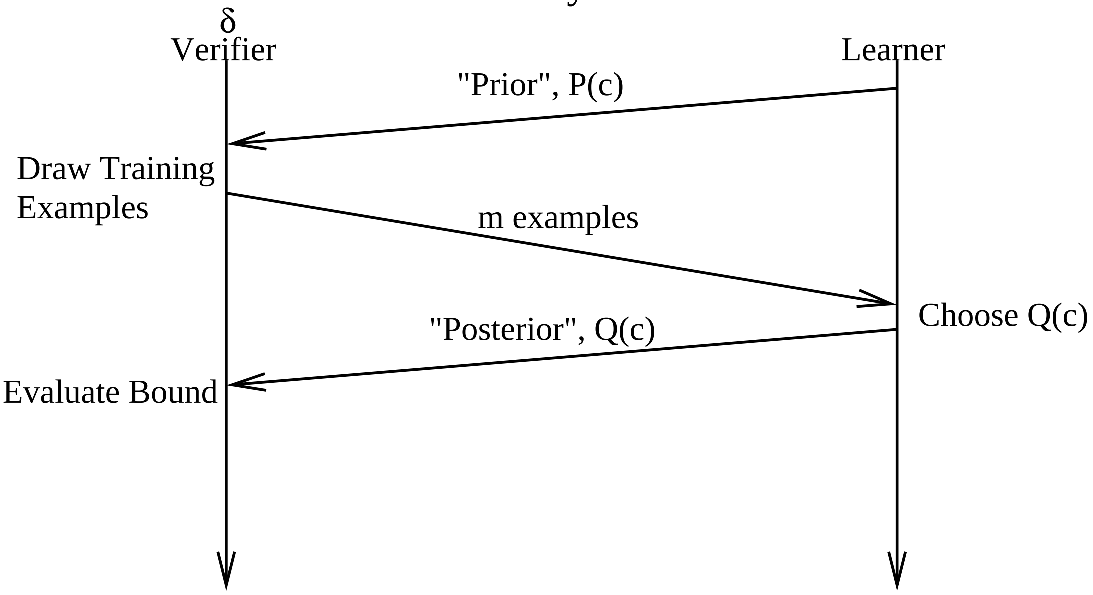
Left bar = holdout bound, right bar = Occam's Razor Bound

PAC-Bayes Bound: Basic quantities

$Q_D \equiv E_{c \sim Q} c_D$ = average true error

$\hat{Q}_S \equiv E_{c \sim Q} \hat{c}_S$ = average train error

PAC−Bayes Bound



Lemma 1: For all $P(c)$, for all D, for all $\delta \in (0,1]$:

$$\Pr_{S\sim D^m}\left(E_{c\sim P}\frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)} \leq \frac{m+1}{\delta}\right) \geq 1-\delta$$

PAC-Bayes Bound: Theorem

Theorem: (PAC-Bayes Bound) For all "priors" $P(c)$ over the classifiers c, for all D, for all $\delta \in (0,1]$:

$$\Pr_{S\sim D^m}\left(\forall Q(c):\ \mathrm{KL}\left(\hat{Q}_S||Q_D\right) \leq \frac{\mathrm{KL}(Q||P) + \ln\frac{m+1}{\delta}}{m}\right) \geq 1-\delta$$

where: $\mathrm{KL}(Q||P) = E_{c\sim Q}\ln\frac{Q(c)}{P(c)}$

$\mathrm{KL}(q||p) = q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p}$ for $q < p$

The proof uses two quick lemmas

Lemma 1: For all $P(c)$, for all D, for all $\delta \in (0,1]$:

$$\Pr_{S\sim D^m}\left(E_{c\sim P}\frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)} \le \frac{m+1}{\delta}\right) \ge 1-\delta$$

Proof:

$$\forall c\ \ E_{S\sim D^m}\frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)} = \sum_{\frac{k}{m}} \Pr_{S\sim D^m}\left(\hat{c}_S = \frac{k}{m}\right)\frac{1}{\Pr_{S\sim D^m}\left(\hat{c}_S = \frac{k}{m}\right)} = m+1$$

$$\Rightarrow E_{S\sim D^m}E_{c\sim P}\frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)} = m+1$$

Lemma 1: For all $P(c)$, for all D, for all $\delta \in (0,1]$:

$$\Pr_{S\sim D^m}\left(E_{c\sim P}\frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)} \le \frac{m+1}{\delta}\right) \ge 1-\delta$$

Proof:

$$\forall c\ \ E_{S\sim D^m}\frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)} = \sum_{\frac{k}{m}} \Pr_{S\sim D^m}\left(\hat{c}_S = \frac{k}{m}\right)\frac{1}{\Pr_{S\sim D^m}\left(\hat{c}_S = \frac{k}{m}\right)} = m+1$$

Lemma 2: For all $Q(c)$: $\frac{E_{c\sim Q} \ln \frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)}}{m} \geq \mathrm{KL}(\hat{Q}_S || Q_D)$

Lemma 1: For all $P(c)$, for all D, for all $\delta \in (0,1]$:

$$\Pr_{S\sim D^m}\left(E_{c\sim P} \frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)} \leq \frac{m+1}{\delta} \right) \geq 1-\delta$$

Proof:

$$\forall c \;\; E_{S\sim D^m} \frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)} = \sum_{\frac{k}{m}} \Pr_{S\sim D^m}\left(\hat{c}_S = \frac{k}{m}\right) \frac{1}{\Pr_{S\sim D^m}\left(\hat{c}_S = \frac{k}{m}\right)} = m+1$$

$$\Rightarrow E_{S\sim D^m} E_{c\sim P} \frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)} = m+1$$

Use the Markov inequality ($X \geq 0$, $EX = \mu$, $\Rightarrow \Pr(X > \frac{\mu}{\delta}) \leq \delta$):

$$\forall P \;\; \Pr_{S\sim D^m}\left(E_{c\sim P} \frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)} \leq \frac{m+1}{\delta} \right) \leq \delta$$

$\Rightarrow$lemma

Lemma 2: For all $Q(c)$: $\frac{E_{c\sim Q}\ln\frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)}}{m} \geq \mathrm{KL}(\hat{Q}_S||Q_D)$

Proof:$\frac{1}{m}E_{c\sim Q}\ln\frac{1}{\binom{m}{m\hat{c}_S}c_D^{m\hat{c}_S}(1-c_D)^{m(1-\hat{c}_S)}}$

$$= E_{c\sim Q}\left[\hat{c}_S\ln\frac{1}{c_D}+(1-\hat{c}_S)\ln\frac{1}{1-c_D}\right]-\frac{E_{c\sim Q}\ln\binom{m}{m\hat{c}_S}}{m}$$

Jensen's inequality (f concave $\Rightarrow$ $Ef(X)\geq f(EX)$):

$$E_{c\sim Q}\left[\hat{c}_S\ln\frac{1}{c_D}+(1-\hat{c}_S)\ln\frac{1}{1-c_D}\right]\geq\hat{Q}_S\ln\frac{1}{Q_D}+\left(1-\hat{Q}_S\right)\ln\frac{1}{1-Q_D}$$

Lemma 2: For all $Q(c)$: $\frac{E_{c\sim Q}\ln\frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)}}{m} \geq \mathrm{KL}(\hat{Q}_S||Q_D)$

Proof:$\frac{1}{m}E_{c\sim Q}\ln\frac{1}{\binom{m}{m\hat{c}_S}c_D^{m\hat{c}_S}(1-c_D)^{m(1-\hat{c}_S)}}$

$$= E_{c\sim Q}\left[\hat{c}_S\ln\frac{1}{c_D}+(1-\hat{c}_S)\ln\frac{1}{1-c_D}\right]-\frac{E_{c\sim Q}\ln\binom{m}{m\hat{c}_S}}{m}$$

PAC-Bayes bound: Proof

Let

$$P_G(c) = \frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)\, E_{c\sim P}\frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)}} P(c)$$

Lemma 2: For all $Q(c)$: $\frac{E_{c\sim Q} \ln \frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)}}{m} \geq \mathrm{KL}(\hat{Q}_S || Q_D)$

Proof: $\frac{1}{m} E_{c\sim Q} \ln \frac{1}{\binom{m}{m\hat{c}_S} c_D^{m\hat{c}_S}(1-c_D)^{m(1-\hat{c}_S)}}$

$$= E_{c\sim Q}\left[\hat{c}_S \ln \frac{1}{c_D} + (1-\hat{c}_S)\ln\frac{1}{1-c_D}\right] - \frac{E_{c\sim Q}\ln \binom{m}{m\hat{c}_S}}{m}$$

Jensen's inequality (f concave $\Rightarrow$ $Ef(X) \geq f(EX)$):

$$E_{c\sim Q}\left[\hat{c}_S \ln \frac{1}{c_D} + (1-\hat{c}_S)\ln\frac{1}{1-c_D}\right] \geq \hat{Q}_S \ln \frac{1}{Q_D} + \left(1-\hat{Q}_S\right)\ln\frac{1}{1-Q_D}$$

and $\frac{E_{c\sim Q}\ln \binom{m}{m\hat{c}_S}}{m} \leq \frac{E_{c\sim Q}\ln e^{mH(\hat{c}_S)}}{m} = \frac{E_{c\sim Q} mH(\hat{c}_S)}{m} \leq H\left(\hat{Q}_S\right)$

$\Rightarrow$lemma

PAC-Bayes bound: Proof

Let

$$P_G(c) = \frac{1}{\mathsf{Pr}_{S\sim D^m}\left(\hat{c}_S\right) E_{c\sim P}\frac{1}{\mathsf{Pr}_{S\sim D^m}(\hat{c}_S)}} P(c)$$

$$\Rightarrow 0 \leq \mathsf{KL}(Q||P_G) = E_{c\sim Q} \ln \frac{Q(c)}{P(c)} \mathop{\mathsf{Pr}}_{S\sim D^m}\left(\hat{c}_S\right) E_{c\sim P}\frac{1}{\mathsf{Pr}_{S\sim D^m}\left(\hat{c}_S\right)}$$

$$= \mathsf{KL}(Q||P) - E_{c\sim Q} \ln \frac{1}{\mathsf{Pr}_{S\sim D^m}\left(\hat{c}_S\right)} + \ln E_{c\sim P}\frac{1}{\mathsf{Pr}_{S\sim D^m}\left(\hat{c}_S\right)}$$

$$\Rightarrow E_{c\sim Q} \ln \frac{1}{\mathsf{Pr}_{S\sim D^m}\left(\hat{c}_S\right)} \leq \mathsf{KL}(Q||P) + \ln E_{c\sim P}\frac{1}{\mathsf{Pr}_{S\sim D^m}\left(\hat{c}_S\right)}$$

Lemma 1&2 ⇒proof

PAC-Bayes bound: Proof

Let

$$P_G(c) = \frac{1}{\mathsf{Pr}_{S\sim D^m}\left(\hat{c}_S\right) E_{c\sim P}\frac{1}{\mathsf{Pr}_{S\sim D^m}(\hat{c}_S)}} P(c)$$

$$\Rightarrow 0 \leq \mathsf{KL}(Q||P_G) = E_{c\sim Q} \ln \frac{Q(c)}{P(c)} \mathop{\mathsf{Pr}}_{S\sim D^m}\left(\hat{c}_S\right) E_{c\sim P}\frac{1}{\mathsf{Pr}_{S\sim D^m}\left(\hat{c}_S\right)}$$

$$= \mathsf{KL}(Q||P) - E_{c\sim Q} \ln \frac{1}{\mathsf{Pr}_{S\sim D^m}\left(\hat{c}_S\right)} + \ln E_{c\sim P}\frac{1}{\mathsf{Pr}_{S\sim D^m}\left(\hat{c}_S\right)}$$

PAC-Bayes Margin bound

$\bar{F}(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ = cumulative distribution of a Gaussian

$Q(\vec{w},\mu) = N(\mu,1) \times N(0,1)^{n-1}$ where first direction parallel to $\vec{w}$

$\gamma(\vec{x},y) = \frac{y\vec{w}\cdot\vec{x}}{||\vec{w}||||\vec{x}||}$ = normalized margin

$\hat{Q}(\vec{w},\mu)_S = E_{\vec{x},y\sim S}\bar{F}\left(\mu\gamma(\vec{x},y)\right)$ = stochastic error rate

Corollary: (PAC-Bayes Margin Bound) For all distributions D, for all $\delta \in (0,1]$:

$$\Pr_{S\sim D^m}\left(\forall \vec{w},\mu>0:\ \mathrm{KL}\left(\hat{Q}(\vec{w},\mu)_S||Q(\vec{w},\mu)_D\right) \le \frac{\frac{\mu^2}{2}+\ln\frac{m+1}{\delta}}{m}\right) \ge 1-\delta$$

PAC-Bayes Bound: Application

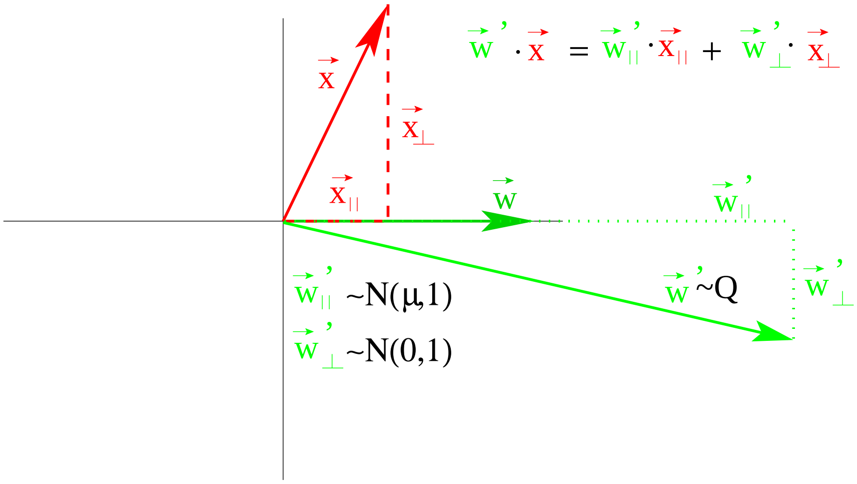
Is the PAC-Bayes bound tight enough to be useful?

Application: true error bounds for Support Vector Machines.

Classifier form:

$$c(x) = \mathrm{sign}\left(\vec{w}\cdot\vec{x}\right)$$

Also note: Work by Mattias Seeger applying to Gaussian Processes.


$\vec{w}' \cdot \vec{x} = \vec{w}'_{||} \cdot \vec{x}_{||} + \vec{w}'_{\perp} \cdot \vec{x}_{\perp}$
$\vec{x}$
$\vec{x}_{\perp}$
$\vec{x}_{||}$
$\vec{w}$
$\vec{w}'_{||}$
$\vec{w}' \sim Q$
$\vec{w}'_{\perp}$
$\vec{w}'_{||} \sim N(\mu,1)$
$\vec{w}'_{\perp} \sim N(0,1)$

$$\hat{Q}(\vec{w},\mu)_S = E_{\vec{x},y\sim S,\vec{w}'\sim Q(\vec{w},\mu)} I\left(y \neq \text{sign}\left(\vec{w}' \cdot \vec{x}\right)\right)$$

$$= E_{\vec{x},y\sim S} E_{w'_{||}\sim N(\mu,1)} E_{w'_{\perp}\sim N(0,1)} I\left(y(w'_{||}x_{||} + w'_{\perp}x_{\perp}) \leq 0\right)$$

Use properties of Gaussians to finish proof

PAC-Bayes Margin Bound: Proof

Start with PAC-Bayes bound:

$$\forall P(c) \quad \Pr_{S\sim D^m}\left(\forall Q(c):\ \text{KL}\left(\hat{Q}_S||Q_D\right) \leq \frac{\text{KL}(Q||P) + \ln\frac{m+1}{\delta}}{m}\right) \geq 1-\delta$$

Set $P = N(0,1)^n$

$Q(\vec{w},\mu) = N(\mu,1) \times N(0,1)^{n-1}$ with first direction parallal to $\vec{w}$

Gaussian $\Rightarrow$ coordinate system reorientable

$$\Rightarrow \text{KL}(Q||P) = \text{KL}(N(0,1)^{n-1}||N(0,1)^{n-1}) + \text{KL}(N(\mu,1)||N(0,1))$$

$$= \frac{\mu^2}{2}$$

PAC-Bayes: Application to SVM

SVM classifier:

$$c(x) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i k(x_i, x)\right)$$

k is a kernel $\Rightarrow \exists \vec{\Phi}: \quad k(x_i, x) = \vec{\Phi}(x_i) \cdot \vec{\Phi}(x)$ so:

$\vec{w} \cdot \vec{x} = \sum_{i=1}^{m} \alpha_i k(x_i, x)$ $\qquad$ $\vec{w} \cdot \vec{w} = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j)$

$$\Rightarrow \gamma(x, y) = \frac{\sum_{i=1}^{m} \alpha_i k(x_i, x)}{\sqrt{k(x,x) \sum_{i,j=1,1}^{m,m} \alpha_i \alpha_j k(x_i, x_j)}}$$

$\Rightarrow$ Margin bound applies to support vector machines.s

PAC-Bayes Margin proof: the end

$$= E_{\vec{x},y\sim S} E_{z'\sim N(0,1)} E_{w'_\perp \sim N(0,1)} I\left(y\mu \le -yz' - yw'_\perp \frac{x_\perp}{x_{||}}\right)$$

The sum of two Gaussians is a Gaussian $\Rightarrow$

$$= E_{\vec{x},y\sim S} E_{v\sim N\left(0,1+\frac{x_\perp^2}{x_{||}^2}\right)} I\left(y\mu \le -yv\right)$$

$$= E_{\vec{x},y\sim S} E_{v\sim N\left(0,\frac{1}{\gamma(\vec{x},y)^2}\right)} I\left(y\mu \le -yv\right)$$

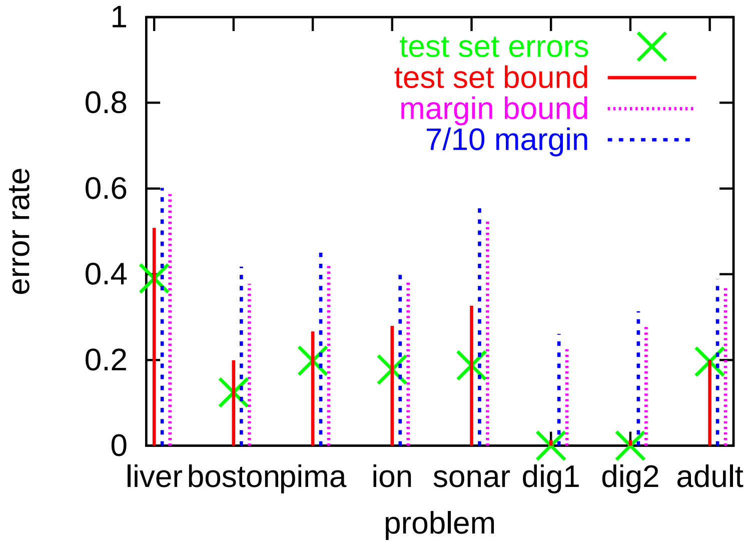
$$= E_{\vec{x},y\sim S} \bar{F}\left(\mu\gamma(\vec{x},y)\right)$$

$\Rightarrow$ Corollary

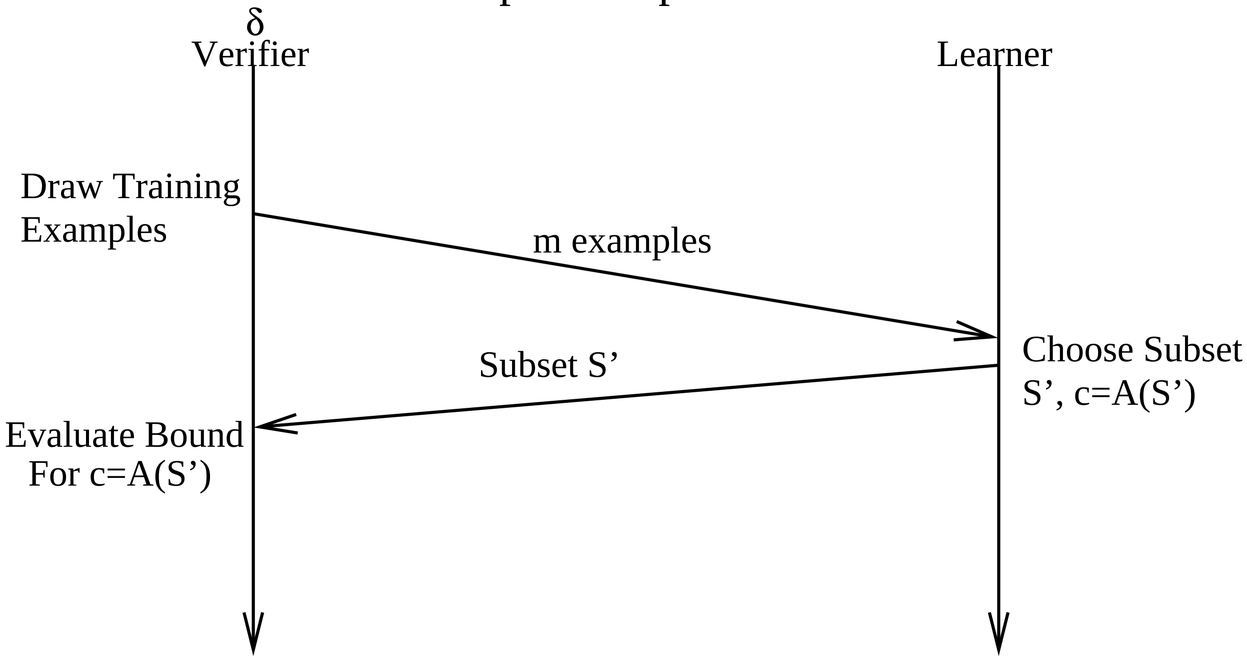
Outline

1. The Basic Model
2. The Test Set Bound
3. Occam's Razor Bound
4. PAC-Bayes Bound
5. Sample Compression Bound

PAC-Bayes Margin Bound Results



Sample Compression Bound



δ
Verifier

Learner

Draw Training Examples

m examples

Choose Subset S', c=A(S')

Subset S'

Evaluate Bound For c=A(S')

Sample Compression Bound: Basic Quantities

A = learning algorithm

$S' \subseteq S$ = subset of training set.

The idea: if we knew S' in advance that $S - S'$ would act as a test set. We don't know S' in advance so the bound is looser.

Sample Compression Bound: Proof

(Note that S' is unambiguous before S is drawn)

$$\forall S' \subseteq S \text{ with } c = A(S') : \Pr_{S\sim D^m}\left(c_D \leq \overline{\mathsf{Bin}}\left(\hat{c}_{S-S'}, \frac{\delta}{m\binom{m}{|S-S'|}}\right)\right) \geq 1 - \frac{\delta}{m\binom{m}{|S-S'|}}$$

Sample Compression Bound: Theorem

Theorem: (Sample Compression Bound) For all algorithms A, for all distributions D, for all $\delta \in (0,1]$:

$$\Pr_{S\sim D^m}\left(\forall S' \subseteq S \text{ with } c = A(S') : \;\; c_D \leq \overline{\mathsf{Bin}}\left(\hat{c}_{S-S'}, \frac{\delta}{m\binom{m}{|S'|}}\right)\right) \geq 1 - \delta$$

Sample Compression Bound: Proof

(Note that S' is unambiguous before S is drawn)

$$\forall S' \subseteq S \text{ with } c = A(S') : \Pr_{S\sim D^m}\left(c_D \leq \overline{\text{Bin}}\left(\hat{c}_{S-S'}, \frac{\delta}{m\binom{m}{|S-S'|}}\right)\right) \geq 1 - \frac{\delta}{m\binom{m}{|S-S'|}}$$

Negate to get:

$$\forall S' \subseteq S \text{ with } c = A(S') : \Pr_{S\sim D^m}\left(c_D > \overline{\text{Bin}}\left(\hat{c}_{S-S'}, \frac{\delta}{m\binom{m}{|S-S'|}}\right)\right) < \frac{\delta}{m\binom{m}{|S-S'|}}$$

Use union bound ($\Pr(A \text{ or } B) \leq \Pr(A) + \Pr(B)$) over each S'

$$\Rightarrow \Pr_{S\sim D^m}\left(\exists S' \subseteq S \text{ with } c = A(S') : \; c_D > \overline{\text{Bin}}\left(\hat{c}_{S-S'}, \frac{\delta}{m\binom{m}{|S-S'|}}\right)\right) < \delta$$

Sample Compression Bound: Proof

(Note that S' is unambiguous before S is drawn)

$$\forall S' \subseteq S \text{ with } c = A(S') : \Pr_{S\sim D^m}\left(c_D \leq \overline{\text{Bin}}\left(\hat{c}_{S-S'}, \frac{\delta}{m\binom{m}{|S-S'|}}\right)\right) \geq 1 - \frac{\delta}{m\binom{m}{|S-S'|}}$$

Negate to get:

$$\forall S' \subseteq S \text{ with } c = A(S') : \Pr_{S\sim D^m}\left(c_D > \overline{\text{Bin}}\left(\hat{c}_{S-S'}, \frac{\delta}{m\binom{m}{|S-S'|}}\right)\right) < \frac{\delta}{m\binom{m}{|S-S'|}}$$

Sample Compression Bound Application: Support Vector Machines

If S'= set of support vectors than $A(S') = A(S)$.

How well does the sample compression bound work with a support vector machine?

Note work by Mario Marchand and John Shawe-Taylor using Sample Compression bound variant for "Set Covering Machine".

Sample Compression Bound: Proof

(Note that S' is unambiguous before S is drawn)

$$\forall S' \subseteq S \text{ with } c = A(S') : \Pr_{S\sim D^m}\left(c_D \leq \overline{\text{Bin}}\left(\hat{c}_{S-S'}, \frac{\delta}{m\binom{m}{|S-S'|}}\right)\right) \geq 1 - \frac{\delta}{m\binom{m}{|S-S'|}}$$

Negate to get:

$$\forall S' \subseteq S \text{ with } c = A(S') : \Pr_{S\sim D^m}\left(c_D > \overline{\text{Bin}}\left(\hat{c}_{S-S'}, \frac{\delta}{m\binom{m}{|S-S'|}}\right)\right) < \frac{\delta}{m\binom{m}{|S-S'|}}$$

Use union bound ($\Pr(A \text{ or } B) \leq \Pr(A) + \Pr(B)$) over each S'

$$\Rightarrow \Pr_{S\sim D^m}\left(\exists S' \subseteq S \text{ with } c = A(S') : \ c_D > \overline{\text{Bin}}\left(\hat{c}_{S-S'}, \frac{\delta}{m\binom{m}{|S-S'|}}\right)\right) < \delta$$

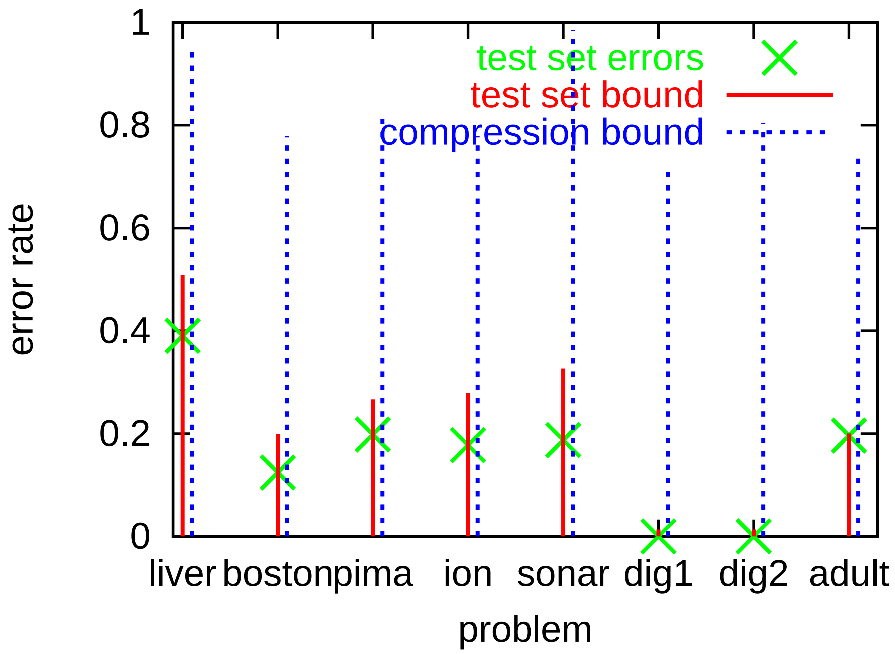
Negate again $\Rightarrow$ theorem

Sample Compression Bound Results

Answer: Sample Compression bound not very tight on SVM.

Why not?

The SVM learning algorithm achieves 'incidental' sparsity rather than optimizing for it explicitly (in contrast to the margin).



Conclusion

1. Use real confidence intervals to compare classifiers.
2. Test set bound very simple.
3. Train set bounds on the threshold of quantitatively useful.

Code for bound calculation at:

http://www-2.cs.cmu.edu/~jcl/programs/bound/bound.html