A Tale of Two Bounds

John Langford

IBM Research

What is a bound?

You have a classifier, $c : X \to Y$.

You have some distribution D over $X \times Y$.

How often will the classifier err in the future?

$$e_D(c) = \Pr_{x,y \sim D}(c(x) \neq y)$$

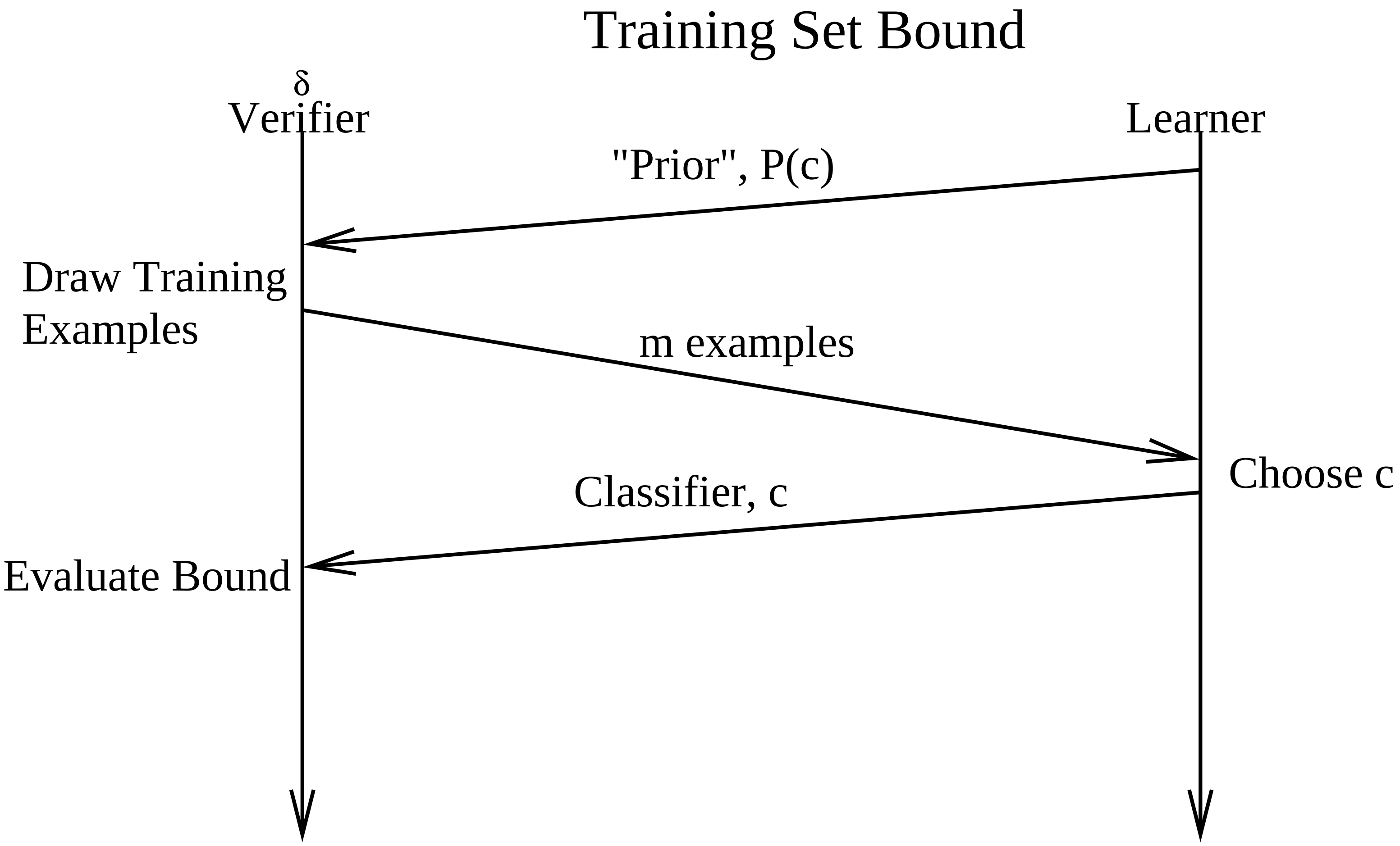
A "bound" is an upper bound on the future error rate.

Types of Bounds

1. Test set bounds (use 'uncontaminated' examples for bounding)

2. Train set bounds (use examples for training and bounding)

This talk: train set bounds.


Training Set Bound
δ
Verifier
Learner
"Prior", P(c)
Draw Training
Examples
m examples
Choose c
Classifier, c
Evaluate Bound

A Brief History of Bounds

1. Test Set bound (aka Binomial confidence interval) ??
2. VC bound (Vapnik and Chervonenkis, 1971)
3. Compression Bound (Littlestone and Warmuth, 1986)
4. Occam's Razor Bound (BEHW, 1987)
5. Margin Bounds (SFBL, 1998, Bartlett 1998)
6. PAC-Bayes Bound (McAllester, 1999)

Outline

1. PAC-Bayes Margin bound for SVMs

2. Compression bound for SVMs

PAC-Bayes Bound

$P(c)$: distribution over classifiers

$Q(c)$: Training example dependent distribution over classifiers

$\hat{e}_S(c)$ = error rate on sample set S.

For all $P(c)$:

$$\Pr_{S\sim D^m}\left(\forall Q(c):\ \mathrm{KL}(E_{c\sim Q}\hat{e}_S(c)||E_{c\sim Q}e_D(c)) \le \frac{\mathrm{KL}(Q||P) + \ln\frac{m+1}{\delta}}{m}\right) \ge 1-\delta$$

$m =$ number of examples

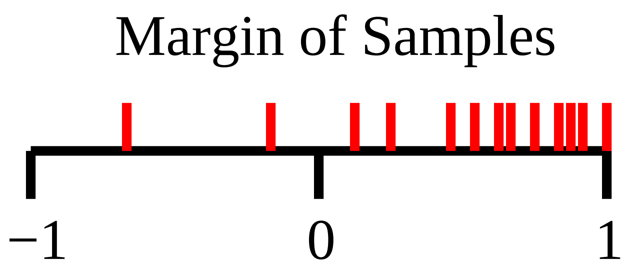
$\mathrm{KL}(q||p) = q \ln \frac{q}{p} + (1 - q) \ln \frac{1-q}{1-p}$ and $\mathrm{KL}(Q||P) = E_{c\sim Q} \ln \frac{Q(c)}{P(c)}$

Specialize to weighted sum classifiers

Consider:

$$c(x) = \text{sign}\left(\sum_{i=1}^{n} w_i x_i\right)$$

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



PAC-Bayes Margin Bound (NIPS2002)

$$\Pr_{S\sim D^m}\left(\forall Q_{\mu,w}: \ \mathrm{KL}(E_{c\sim Q}\hat{e}(c)||E_{c\sim Q}e_D(c)) \leq \frac{\ln\frac{1}{\bar{F}(\mu)} + \ln\frac{m+1}{\delta}}{m}\right) \geq 1-\delta$$

$P = N(0,1)^n$ = isotropic normal

$Q = N(0,1)^{n-1} \times R_\mu$

$\bar{F}(\mu) = 1 - \int_{-\infty}^{\mu} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ = cumulative of the normal

$R_\mu(x) = \frac{I(x>\mu)}{\sqrt{2\pi}\bar{F}(\mu)} e^{-\frac{x^2}{2}}$ = renormalized normal tail

A few things to note

$Q_{\mu,w} = N(0,1)$ in all directions perpendicular to w.

$\epsilon \equiv E_{x,y\sim S}\bar{F}\left(\frac{yx_{||}}{x_{\perp}}\mu\right) \geq E_{c\sim Q}\hat{e}(c)$

which implies:

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

The Margin Bound for Support Vector Machines

Kernelize bound (Herbrich and Graepel 2001) $\gamma(x,y) = \frac{y(w \cdot x)}{\sqrt{w \cdot w}\sqrt{x \cdot x}}$

$$\epsilon = E_{x,y \sim U(S)} \bar{F}\left(\frac{\gamma}{1-\gamma^2}\mu\right)$$

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

k is a kernel $\Rightarrow k(x_i, x) = \Phi(x_i) \cdot \Phi(x)$ where Φ is some projection into another space.

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

Dataset/Learning details

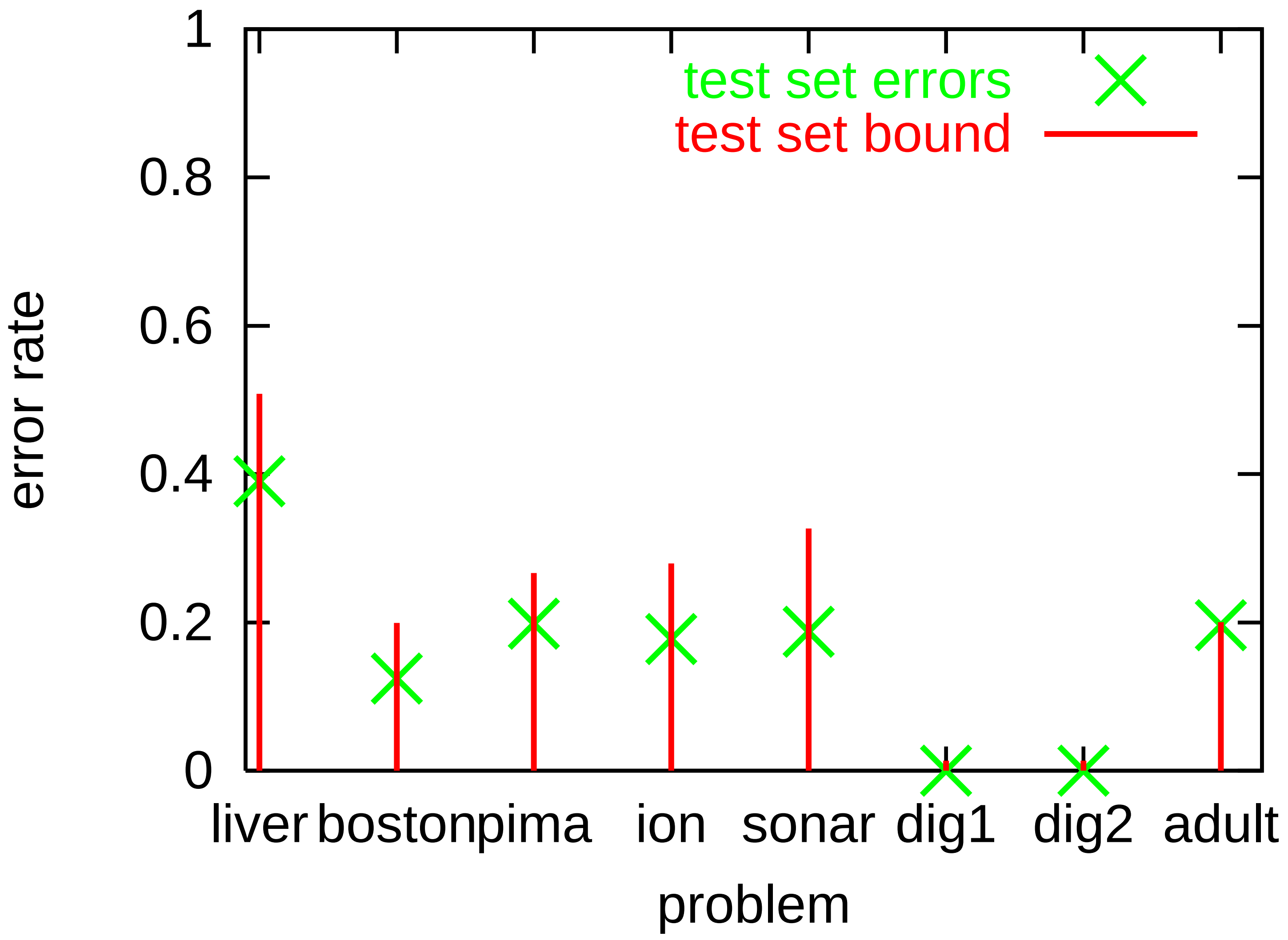
Use “radial basis function” kernel, $K(x_1, x_2) = e^{-\gamma ||x_1 - x_2||^2}$

8 datasets from UCI database, picked for pure numerical features

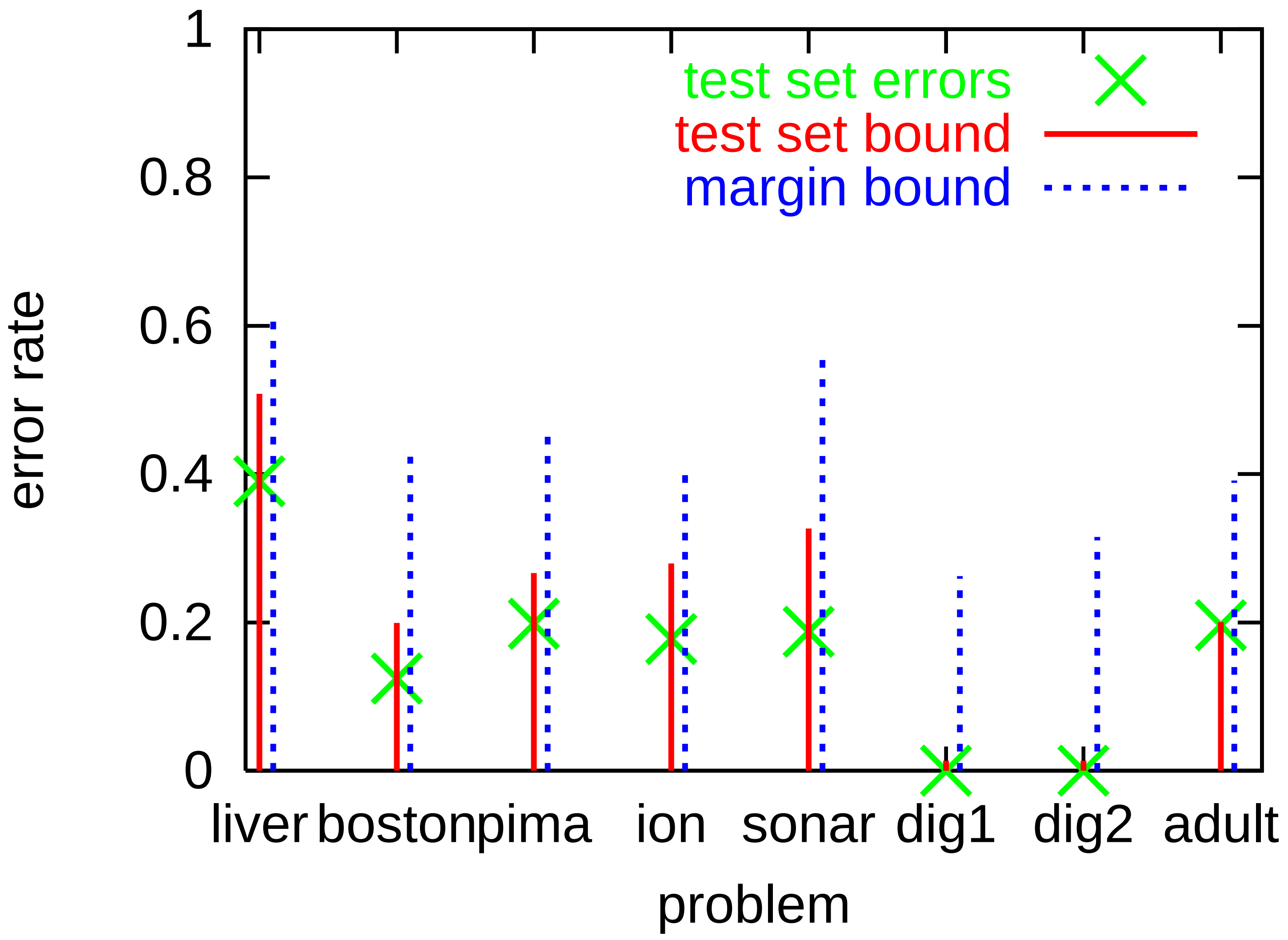
7:3 Train:Test split (except “adult” which is to large)

145 to 1428 examples

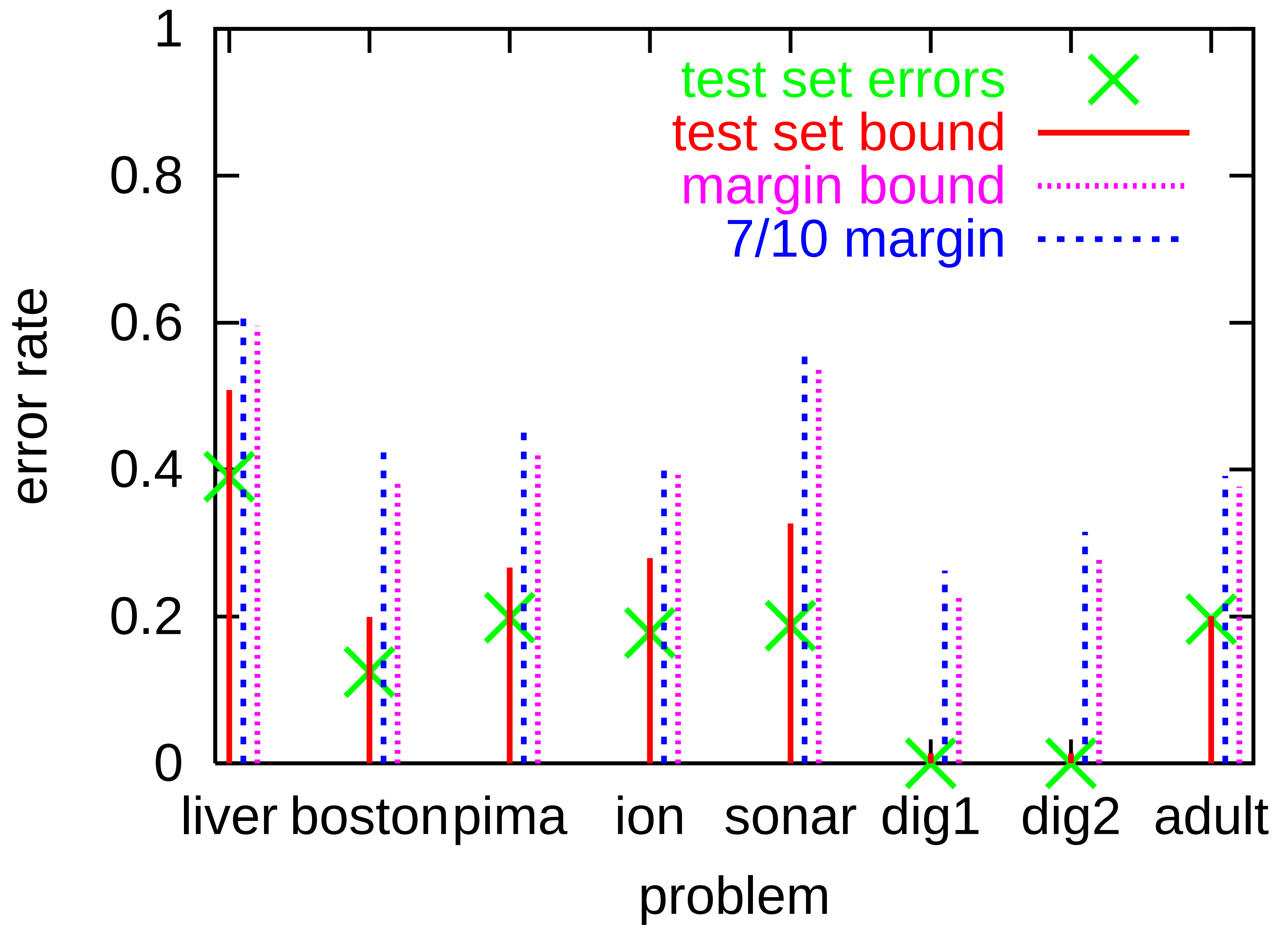
Everything implemented as a postprocessor for SVMlight.



test set errors
test set bound
error rate
1
0.8
0.6
0.4
0.2
0
liver
boston
pima
ion
sonar
dig1
dig2
adult
problem



test set errors
test set bound
margin bound
error rate
1
0.8
0.6
0.4
0.2
0
liver
boston
pima
ion
sonar
dig1
dig2
adult
problem


test set errors
test set bound
margin bound
7/10 margin
error rate
1
0.8
0.6
0.4
0.2
0
liver
boston
pima
ion
sonar
dig1
dig2
adult
problem

What about optimizing?

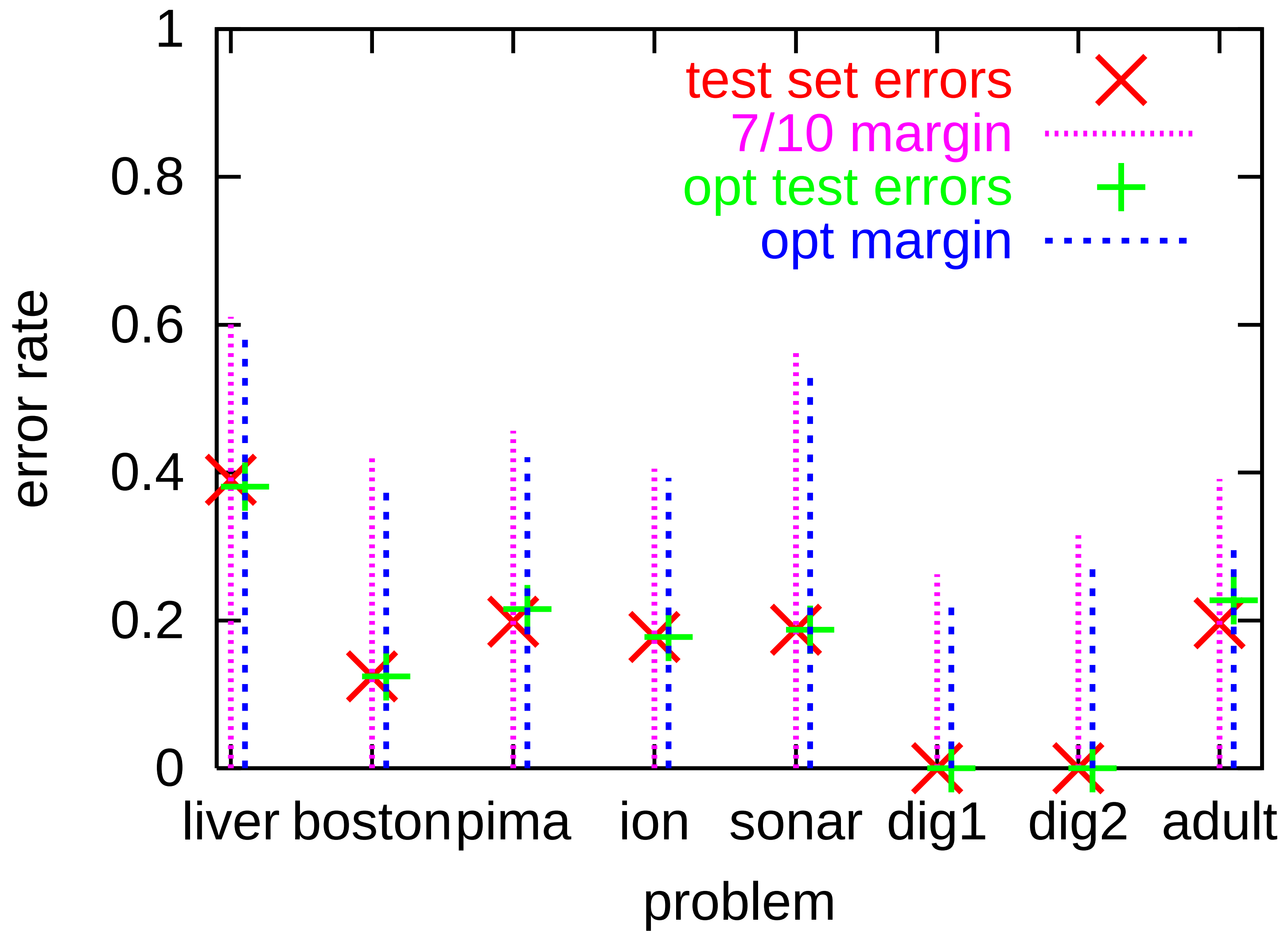
Use the gradient:

$$\frac{\partial \bar{e}}{\partial \alpha_i} = \frac{\ln \frac{\bar{e}(1-\hat{e})}{\hat{e}(1-\bar{e})}}{\frac{\bar{e}-\hat{e}}{\bar{e}(1-\bar{e})}} \frac{\partial \epsilon}{\partial \alpha_i}$$

$$\frac{\partial \epsilon}{\partial \alpha_i} = -\mu E_{x,y\sim S} f\left(\frac{\gamma}{\sqrt{1-\gamma^2}}\mu\right) \frac{1}{(1-\gamma^2)^{\frac{3}{2}}} \frac{\partial \gamma}{\partial \alpha_i}$$

$$\frac{\partial \gamma}{\partial \alpha_i} = \frac{yk(x,x_i)}{\sqrt{w\cdot w}\sqrt{k(x,x)}} - y_i \gamma \gamma_i \sqrt{k(x_i,x_i)}$$

In practice, about x10 slower than SVMlight.


test set errors
7/10 margin
opt test errors
opt margin
error rate
1
0.8
0.6
0.4
0.2
0
liver
boston
pima
ion
sonar
dig1
dig2
adult
problem

Outline

1. PAC-Bayes Margin bound for SVMs

2. Compression bound for SVMs

The Sample Compression Bound

Suppose we knew in advance that the learning algorithm will not depend upon some subset of the examples.

Then, the "undependent" subset acts like a test set and gives us a test set bound.

Let $\mathrm{Bin}(m,k,p) = \sum_{i=0}^{k} \binom{m}{i} p^i (1-p)^{m-i}$ = cumulative of Binomial

Let S' = "undependent subset"

$$\Pr_{S \sim D^m}\left(e_D(c) \leq \mathrm{argmax}\left\{ p : \mathrm{Bin}(|S'|, |S'|\hat{e}_{S'}(c), p) \geq \delta \right\} \right) \geq 1 - \delta$$

The Sample Compression Bound II

Problem: We don't know S' in advance.

Solution: Assign $\frac{\delta}{m\binom{m}{k}}$ confidence to any subset of size k. Note:

$$\sum_{k=1}^{m} \sum_{|S'|=k} \frac{\delta}{m\binom{m}{k}} = \delta$$

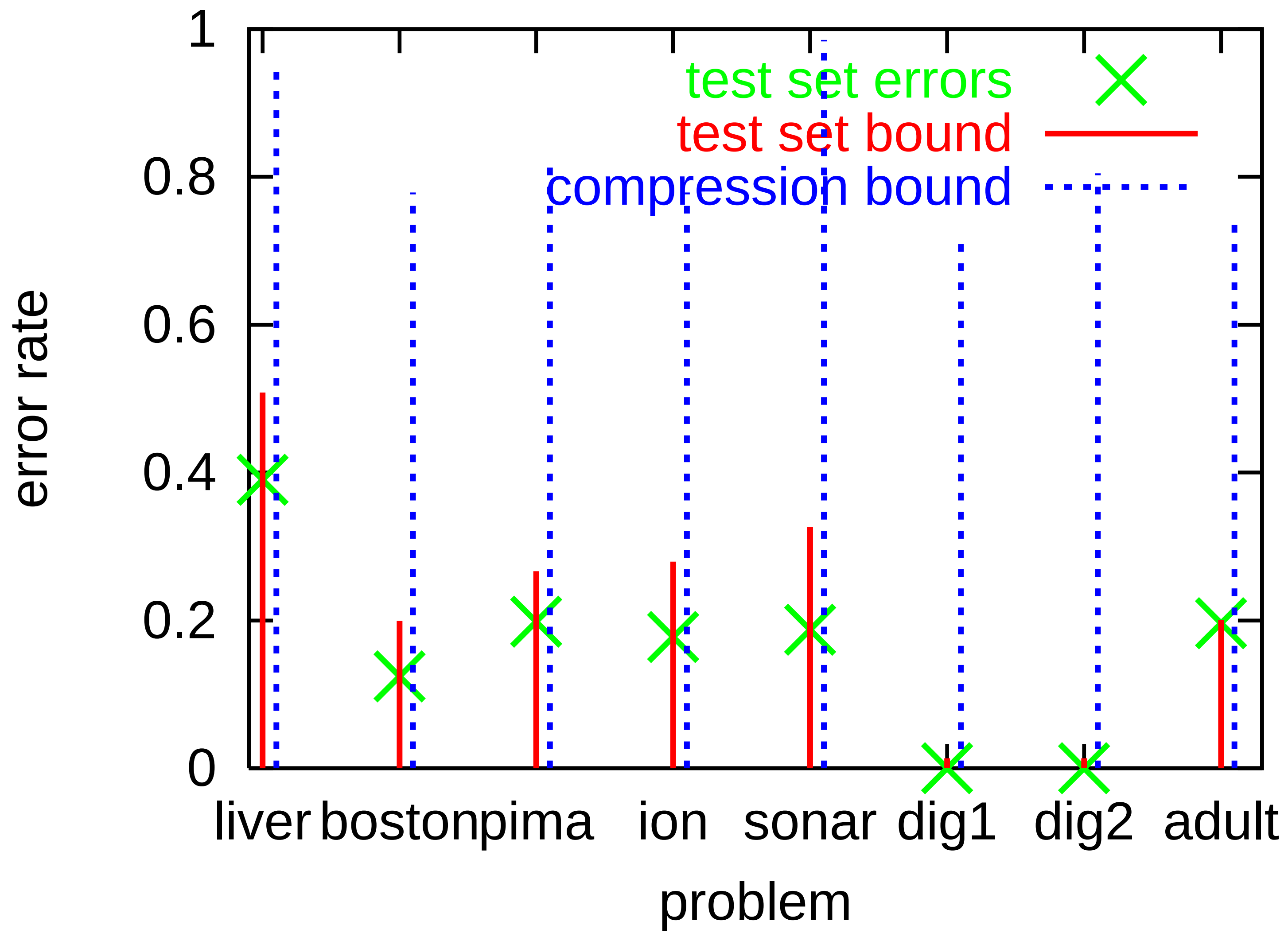
Also note: $\Pr(A \text{ or } B) \leq \Pr(A) + \Pr(B)$. Apply repeatedly to get that with probability $1-\delta$:

$$\forall S' : \ e_D(c) \leq \operatorname{argmax}\left\{ p : \operatorname{Bin}(|S'|, |S'|\hat{e}_{S'}(c), p) \geq \frac{\delta}{m\binom{m}{|S'|\hat{e}_{S'}(c)}} \right\}$$

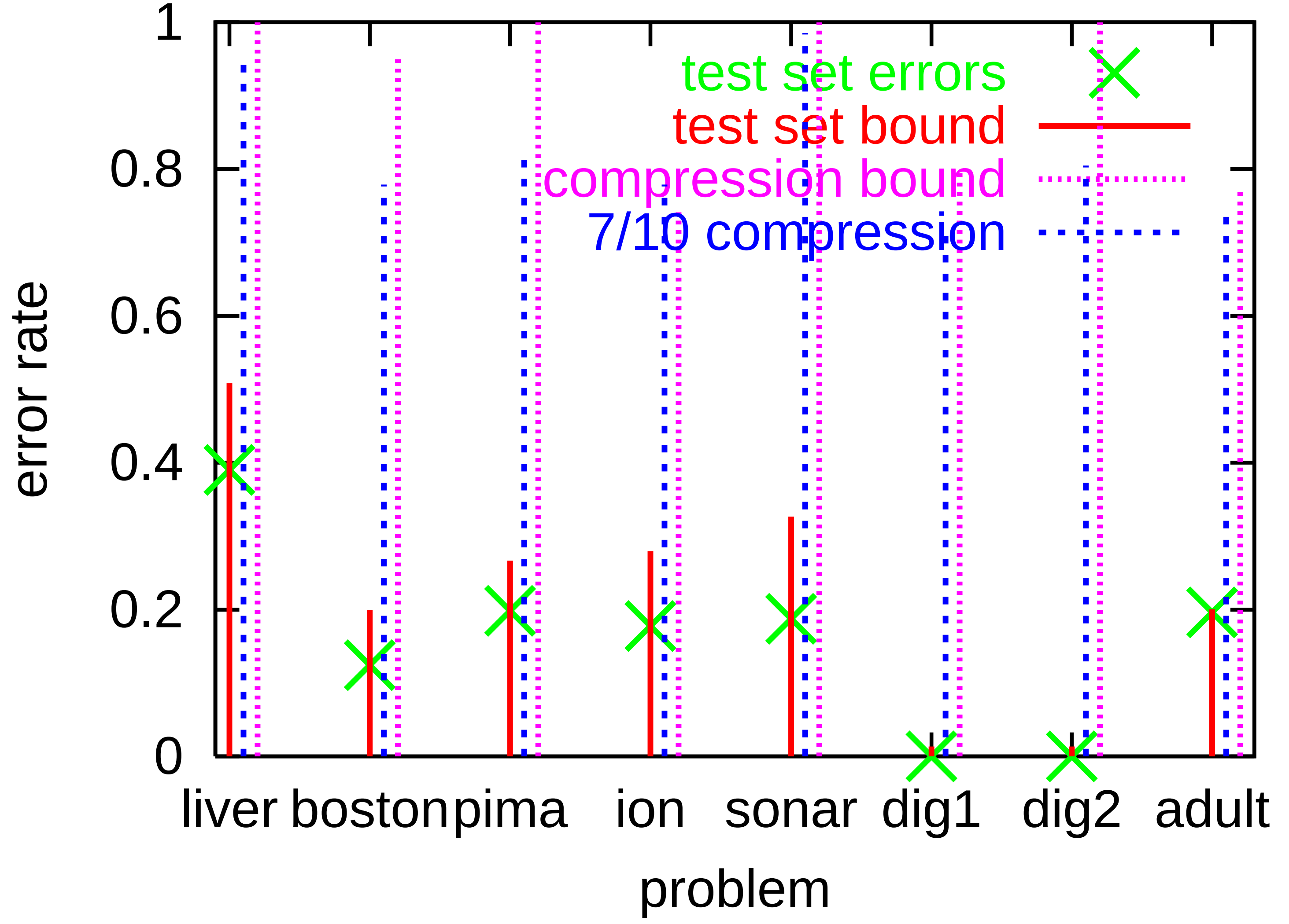
Sample Compression Bound application

A support Vector machine only depends upon support vectors. Nonsupport vectors are "undependent".

$\Rightarrow$ Use performance on the nonsupport vector training set.



1
0.8
0.6
0.4
0.2
0
error rate
test set errors
test set bound
compression bound
liver
boston
pima
ion
sonar
dig1
dig2
adult
problem

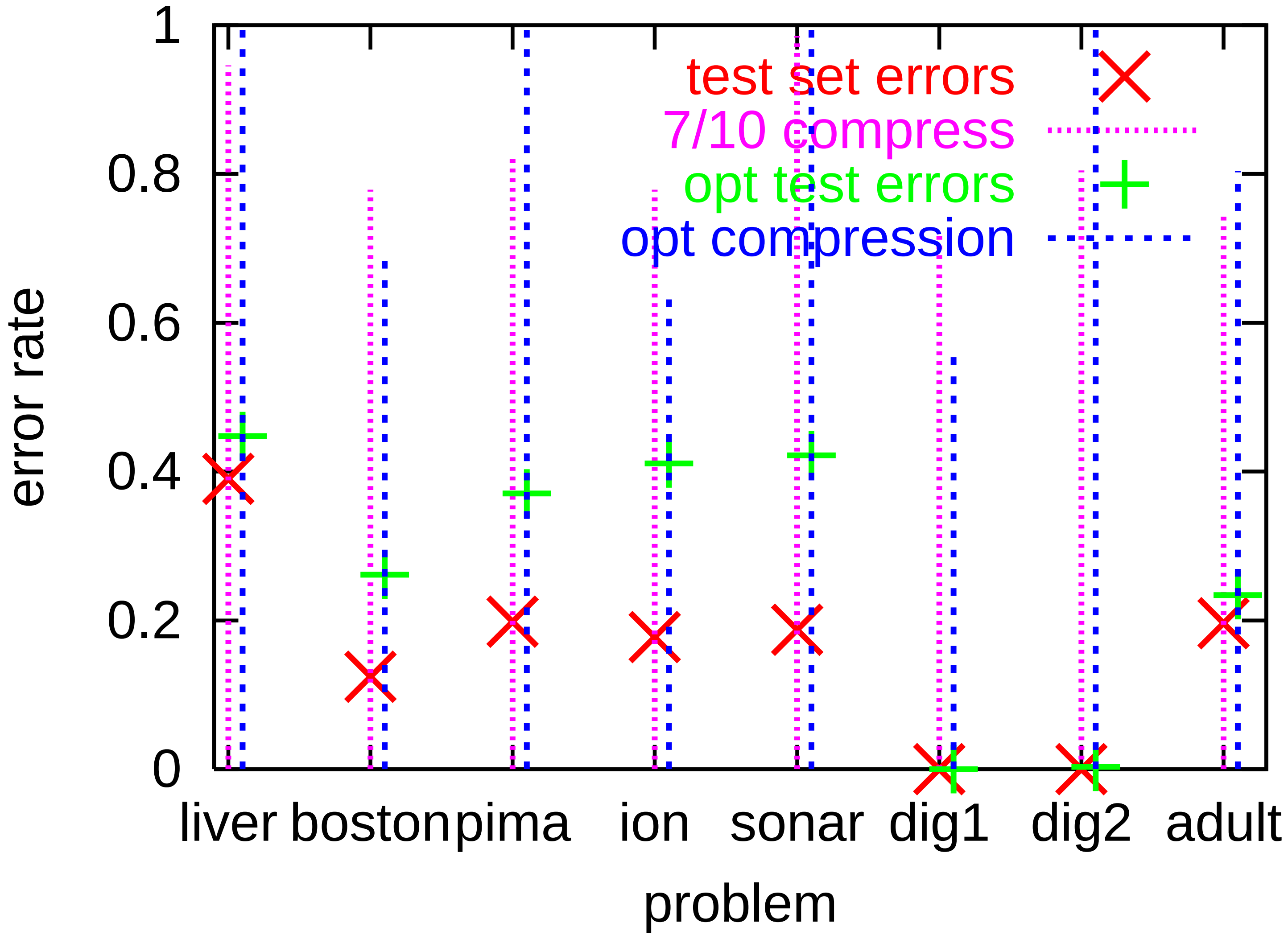

test set errors
test set bound
compression bound
7/10 compression
error rate
problem
1
0.8
0.6
0.4
0.2
0
liver
boston
pima
ion
sonar
dig1
dig2
adult

What about Optimizing?

SVMlight has an option '-i 1' = throw out the inconsistent examples and retrain.

$\Rightarrow$ much fewer support vectors

$\Rightarrow$ more errors in the “undependent” set



1
0.8
0.6
0.4
0.2
0
error rate
test set errors
7/10 compress
opt test errors
opt compression
liver
boston
pima
ion
sonar
dig1
dig2
adult
problem

Conclusion

1. PAC-Bayes Margin bound is good enough to provide useful information for SVMs.

   (a) May be worth combining with test set bound for combined train-test bound

   (b) Approach probably works for Maximum Entropy classifiers

2. Compression bound does not work as well for SVMs.

   (a) Maybe works well for Relevance Vector Machine?

   (b) Maybe more direct sparsity optimization is fruitful?