Probing: Extracting Class Probability from a Classifier

John Langford

(with Bianca Zadrozny)

Reasons for The Probability Estimation Problem

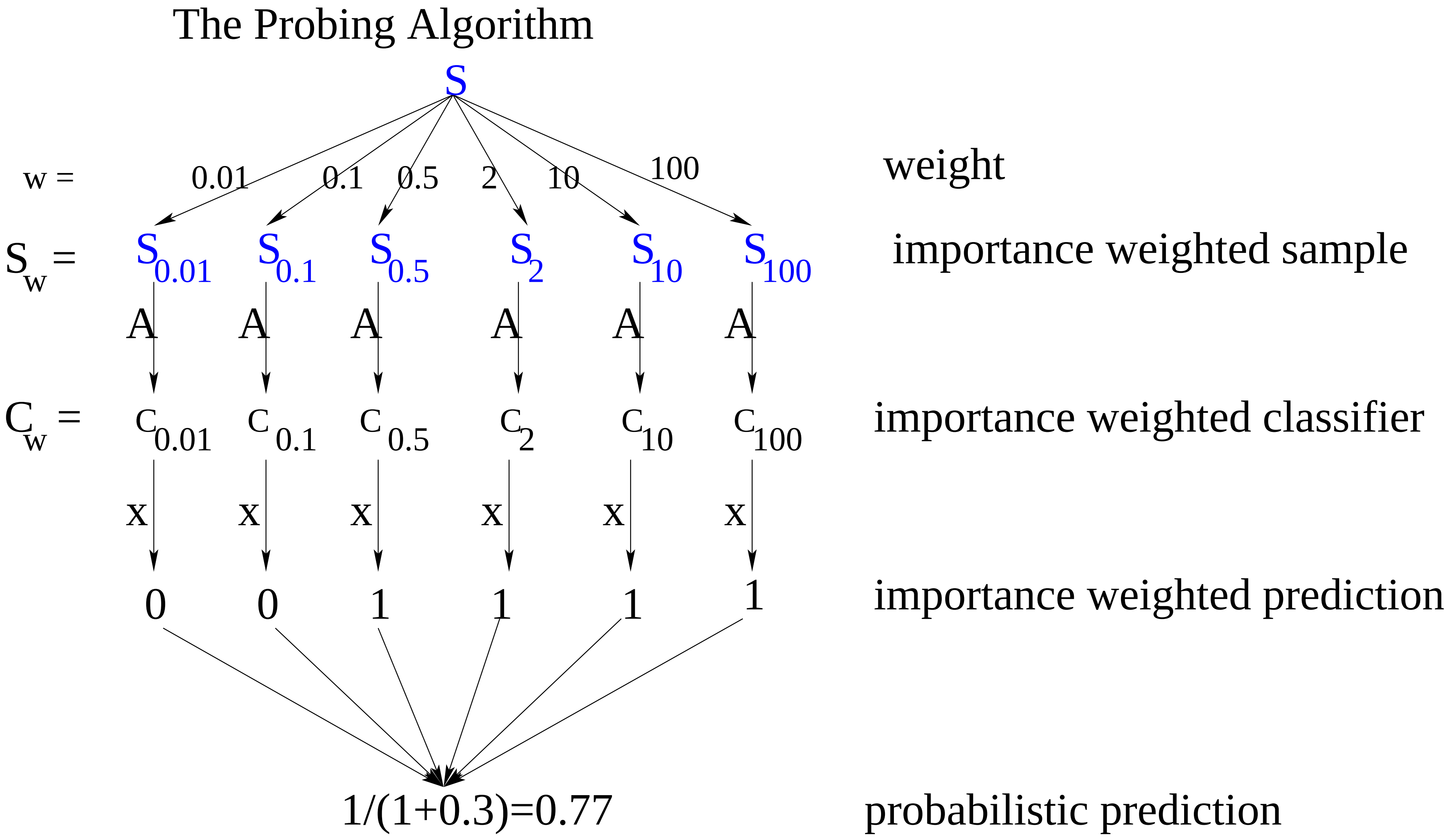
1. Doctor wants 'advice' from a machine, but not a decision.
2. Distributed system requires efficient communication of beliefs.
3. Compatibility between prediction and probabilistic prediction worlds.

The Probing Method: Observations

Observation: $c(x) = 1 \Rightarrow P(y = 1|x) > 0.5$

For importance-weighted classification:

If cost of false negative $= w$ and cost of false positive $= 1$ then
$c(x) = 1 \Rightarrow P(y = 1|x) > \frac{1}{1+w}$


The Probing Algorithm
S
w =
0.01
0.1
0.5
2
10
100
weight
S_w =
S_0.01
S_0.1
S_0.5
S_2
S_10
S_100
importance weighted sample
A
A
A
A
A
A
C_w =
c_0.01
c_0.1
c_0.5
c_2
c_10
c_100
importance weighted classifier
x
x
x
x
x
x
0
0
1
1
1
1
importance weighted prediction
1/(1+0.3)=0.77
probabilistic prediction

The Probing Method: Details

1. How do you make classifier take weights? Costing Reduction

2. How do you deal with nonmonotonic predictions? Sort

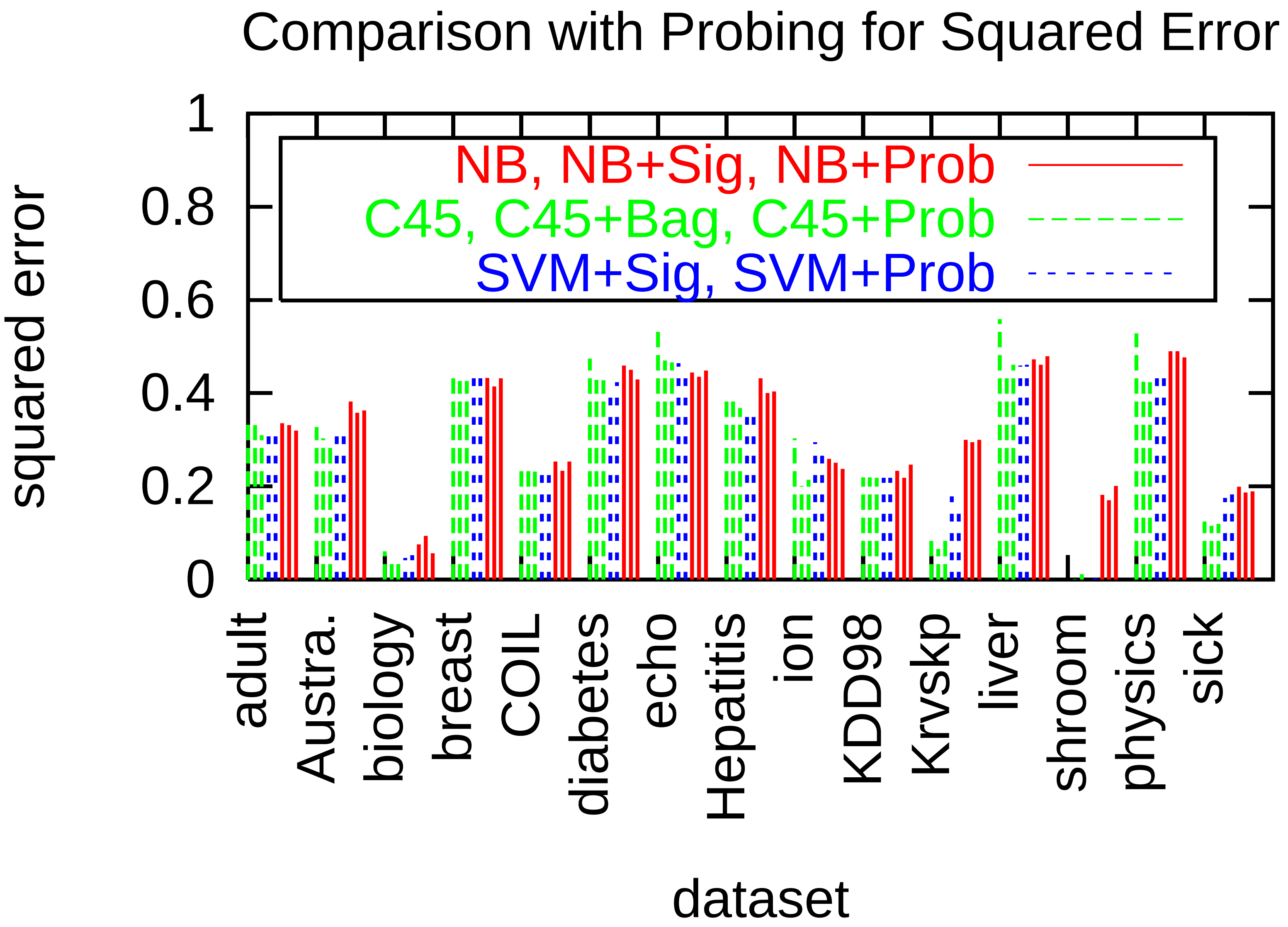
3. How do you discretize on w? Uniform grid in $p = \frac{1}{1+w}$, on demand

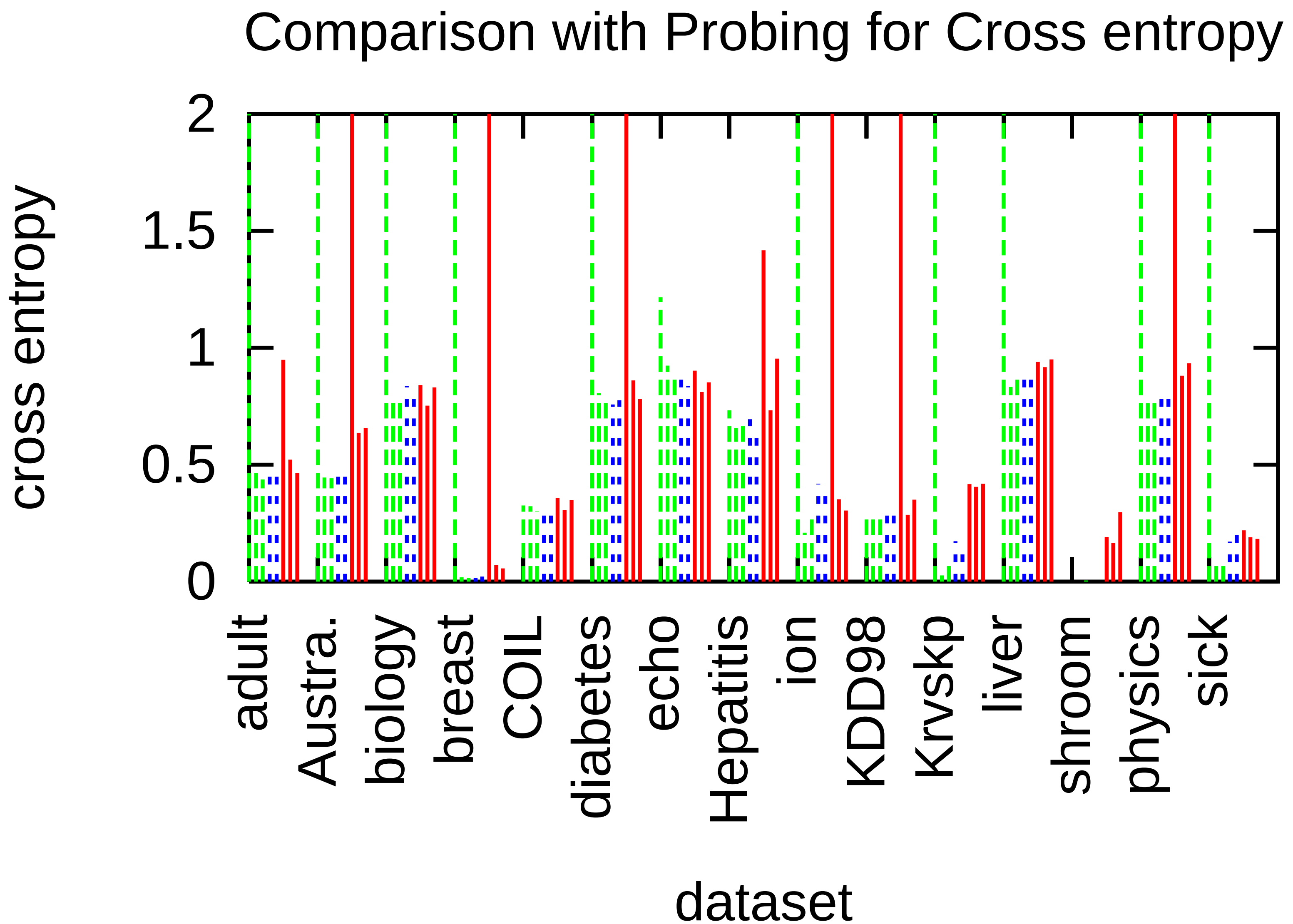
Other Methods

1. Fit a sigmoid to a margin. (+fast -margin specific -no performance guarantee)

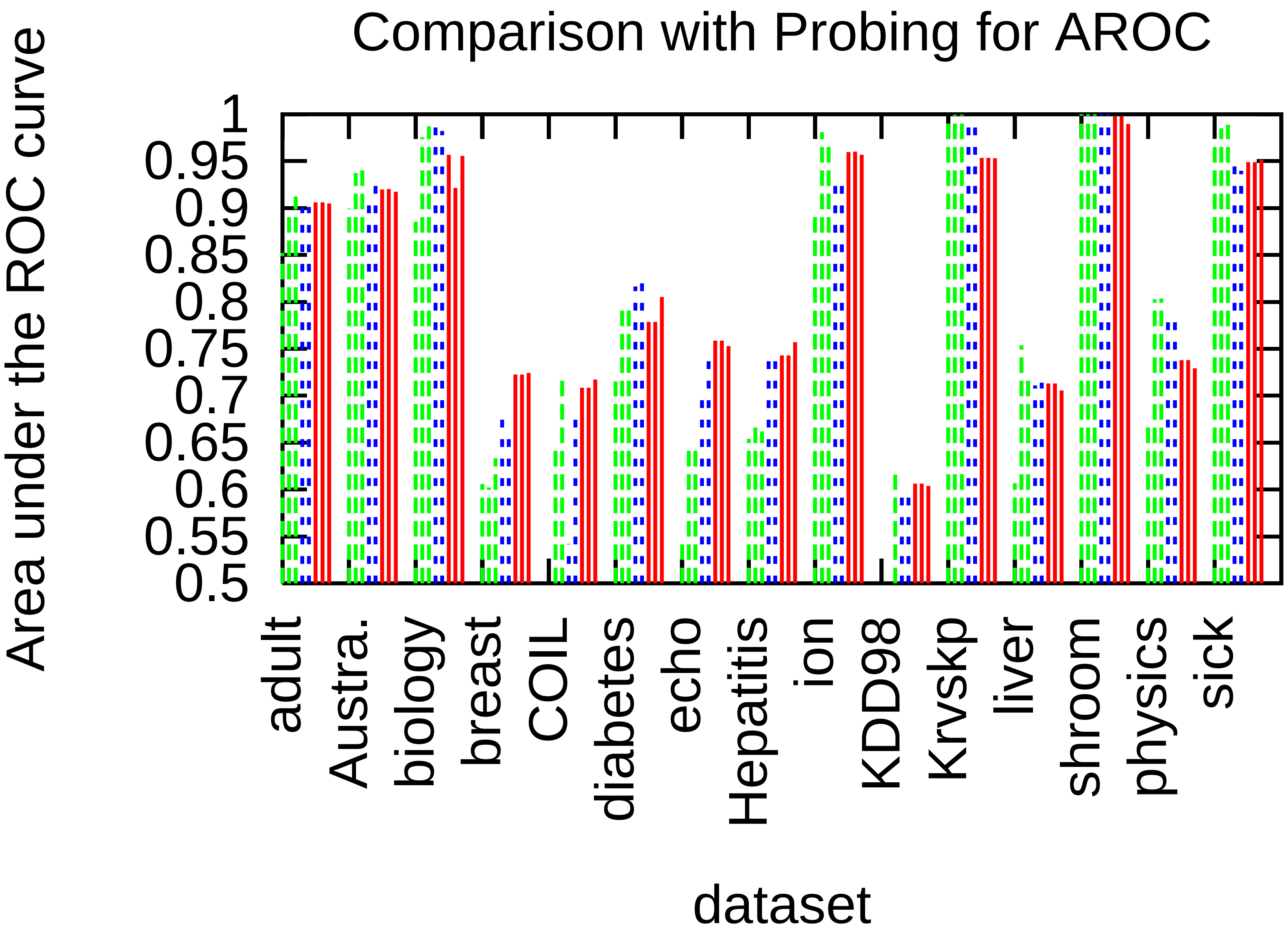
2. Count bagging proportion. (-result unrelated to $P(y|x)$ in general)

3. Classifier native. (+fast -classifier often terribly calibrated)


Comparison with Probing for Squared Error
squared error
dataset
0
0.2
0.4
0.6
0.8
1
NB, NB+Sig, NB+Prob
C45, C45+Bag, C45+Prob
SVM+Sig, SVM+Prob
adult
Austra.
biology
breast
COIL
diabetes
echo
Hepatitis
ion
KDD98
Krvskp
liver
shroom
physics
sick



Comparison with Probing for Cross entropy
cross entropy
2
1.5
1
0.5
0
adult
Austra.
biology
breast
COIL
diabetes
echo
Hepatitis
ion
KDD98
Krvskp
liver
shroom
physics
sick
dataset


Comparison with Probing for AROC
Area under the ROC curve
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
adult
Austra.
biology
breast
COIL
diabetes
echo
Hepatitis
ion
KDD98
Krvskp
liver
shroom
physics
sick
dataset

A Bit of Theory

Let $i_p(y) = \frac{1-p}{p} y + 1 - y$

Corollary: If c_p has minimal importance weighted loss

$$c_p = \arg\min_c E_{(x,y)\sim D} i_p(y) I(c(x) \neq y)$$

$\Rightarrow$ $\text{Probing}(x) = D(y|x)$

Note: This says nothing when error rates are not minimal.

Another Bit of Theory

Let $l_p(c) = \frac{1}{E_{(x,y)\sim D} i_p(y)} E_{(x,y)\sim D} i_p(y) I(c(x) \neq y)$

Theorem: Squared error is bounded by average relative classifier loss. If $c_p^* =$ optimal classifier at threshold p:

$\epsilon = E_{p \sim U(0,1)} l_p(c_p) - l_p(c_p^*)$

$\Rightarrow E_{x,y\sim D}(D(y|x) - \text{Probing}(x))^2 \leq 2 \max\{D(1), D(0)\}\epsilon$

The proof

$$\epsilon = E_{p\sim U(0,1)} l_p(c_p) - l_p(c_p^*)$$

$$= E_p \frac{1}{E_y i_p(y)} E_{(x,y)} \left(\frac{1-p}{p} y + 1 - y\right) \left(I(c_p(x) \neq y) - I(c_p^*(x) \neq y)\right)$$

$$= E_p \frac{1}{E_y i_p(y)} E_x \left[D(1|x)\frac{1-p}{p}(c_p^*(x) - c_p(x)) + D(0|x)(c_p(x) - c_p^*(x))\right]$$

$$= E_p \frac{1}{E_y i_p(y)} E_x \left[D(1|x)\frac{1-p}{p}(c_p^*(x) - c_p(x)) + (1 - D(1|x))(c_p(x) - c_p^*(x))\right]$$

$$= E_p \frac{1}{E_y i_p(y)} E_x (c_p^*(x) - c_p(x)) \left(\frac{1}{p} D(1|x) - 1\right)$$

The proof, part II

$$D(1|x) > p \Rightarrow c_p^*(x) = 1,\; D(1|x) < p \Rightarrow c_p^*(x) = 0$$

$$\Rightarrow \epsilon = E_p \frac{1}{p\left[\frac{p-1}{p}D(1)+D(0)\right]} E_{x\sim D} I(c_p(x) \neq c_p^*(x))|D(1|x) - p|$$

$$p\left[\frac{1-p}{p}D(1) + D(0)\right] = (1-p)D(1) + pD(0) \leq \max\{D(1), D(0)\}$$

$$\Rightarrow \epsilon \max\{D(1), D(0)\} \geq E_p E_{x\sim D} I(c_p(x) \neq c_p^*(x))|D(1|x) - p|$$

$$\Rightarrow \epsilon \max\{D(1), D(0)\} \geq E_{x\sim D} \int_{p=0}^{p=1} I(c_p(x) \neq c_p^*(x))|D(1|x) - p|dp$$

Proof III: Properties of most efficient error inducing method

We have a budget of ϵ errors. How can probability estimation errors be maximized?

1. Only classifiers c_p on one side of $D(1|x)$ err. (otherwise sort = cancellation of errors = wasted binary errors)

2. Errors for p closer to $D(1|x)$ are preferred over errors further from $D(1|x)$ (sort makes these equivalent, and the importance weighted loss is smaller for nearer errors)

$\Rightarrow \epsilon \max\{D(1), D(0)\} \geq E_{x \sim D} \int_{p=\text{Probing}(x)}^{D(1|x)} |D(1|x) - p| dp$

$\Rightarrow$ proof

Some Caveats

1. No bound holds for cross entropy. (Can't be done without extra assumptions/constraints.)

2. No direct bound on error rate for relative ranking.