PAC Bounds for Holdout Sets

John Langford

How can Probably Approximately Correct bounds be used for holdout set?

1. Review: Bounds for specific procedures

2. Combined Training set/Test set bounds

3. Improving the Cross-Validation bound

4. Important final details

What kind of Probably Approximately Correct bounds apply to holdout procedures?

1. Holdout Set

2. K-fold Cross validation

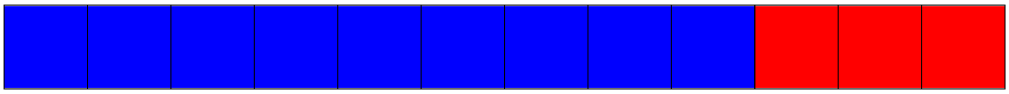
3. Progressive validation

"Beating the Holdout: Bounds for KFold and Progressive Cross-Validation" with Avrim Blum and Adam Kalai at COLT99

Vocabulary

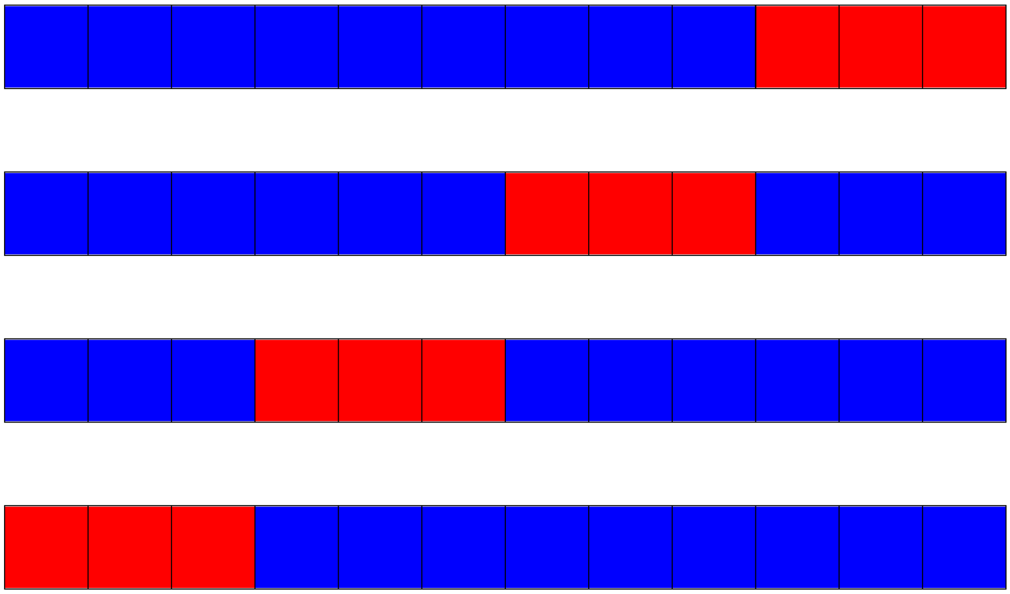
Object	Definition
X	input space
$Y = \{0, 1\}$	output space
$h : X \to Y$	a hypothesis
D	Distribution over $X \times Y$
$\overline{e}(h) = E_D(h(x) \neq y)$	true error
S	Sample set iid from D
$m = \|S\|$	sample count
$\widehat{e}(h) = E_S(h(x) \neq y)$	empirical error
H	Hypothesis set

Holdout Set



$$\mathrm{Pr}(\bar{e}(h) - \hat{e}_{test}(h) > \epsilon) \leq e^{-2m_{test}\epsilon^2}$$

K-fold Cross Validation

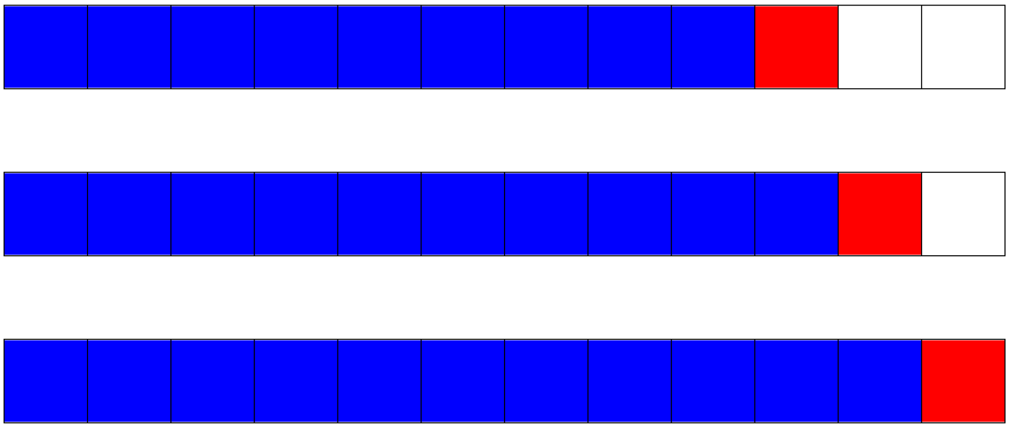


$$\forall n \geq 1 : E|\overline{e}(h) - \widehat{e}_{\frac{m}{k}}(h)|^n \geq E|\overline{e}_{cv} - \widehat{e}_{cv}|^n$$

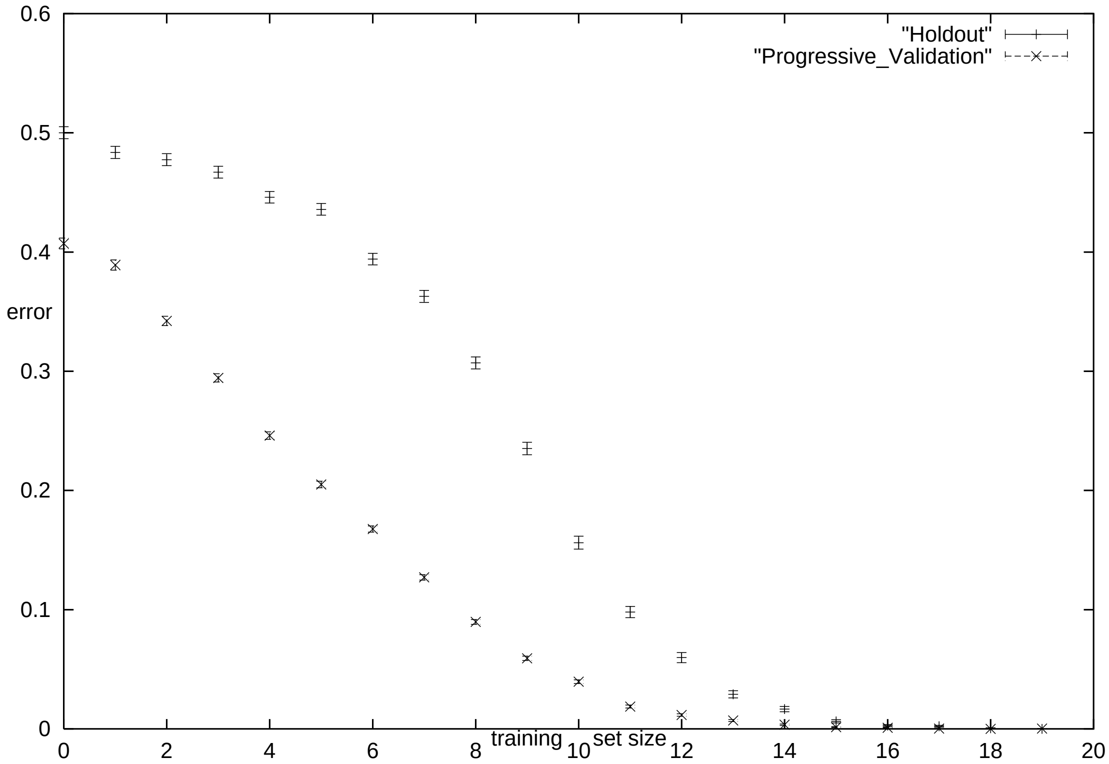
Pathological Cross Validation

- $X = \emptyset$
- $h_1(x) = 1,\ h_0(x) = 0$
- $\Pr_D(Y = 0) = 0.5$
- $\Pr_D(\bar{e}(h) - \hat{e}(h) > \epsilon) =?$

Progressive Validation



$$\Pr(\bar{e}_{pv} - \hat{e}_{pv} > \epsilon) \leq e^{-2m_{pv}\epsilon^2}$$


"Holdout"
"Progressive_Validation"
error
training set size
0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
10
12
14
16
18
20

Combining Training and Test set

Training set bound:

$$\Pr(\exists h \in H : \bar{e}(h) - \hat{e}_{train}(h) > \epsilon) \leq |H| e^{-2m_{train}\epsilon^2}$$

Training set bound occasionally tighter then a holdout bound.

How can we get the best of both bounds?

1. Bound $\Rightarrow$ worst case cumulative distribution over errors

2. Threshold on $F_{train}(\hat{e}_{train})F_{test}(\hat{e}_{test}) \leq \epsilon$

3. Calculate $\Pr(F_{train}(\hat{e}_{train})F_{test}(\hat{e}_{test}) \leq \epsilon)$

4. Choose k small enough to get low probability of error

Improving Cross Validation bounds

$\bar{e}(h_1, h_2) = \mathrm{Pr}_D(h_1(x) \neq h_2(x))$

$\hat{e}(h_1, h_2) = \mathrm{Pr}_S(h_1(x) \neq h_2(x))$

Then:

$$\mathrm{Pr}(\exists h_1, h_2 \in H : \bar{e}(h_1, h_2) - \hat{e}(h_1, h_2) + \epsilon)$$

$$\leq \frac{|H|(|H| + 1)}{2} e^{-2m\epsilon^2}$$

If $\bar{e}(h_1, h_2)$ near 0 then we have "hypothesis stability"

$\Rightarrow$ tighter cross validation bounds

M. Anthony and B. Holden at Colt98

Final Details

- Use instead: Pr$(p - \hat{p} > \epsilon) \leq e^{-mD(p-\epsilon||p)}$

- Progressive validation and holdout bounds: pretty tight and quantitatively applicable

- Cross validation bounds: theory much looser then reality

Code at http://www.cs.cmu.edu/~jcl/research