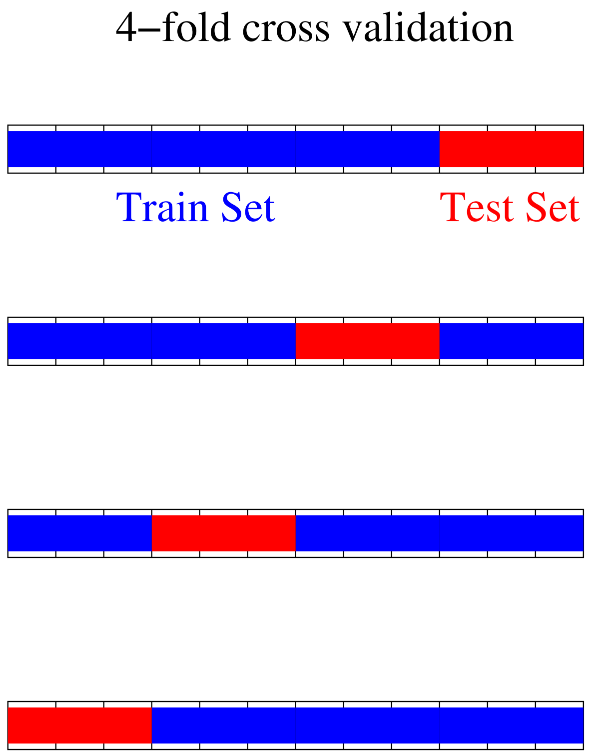
# The Cross Validation Problem

John Langford

COLT 2005




Everybody does it. Nobody knows why.

## The Problem

Derive a classifier from the $K$ classifiers with a lower true error rate bound than other approaches. The bound should have the form:

For all learning algorithms $A$, fold numbers $K$,

$$\Pr_{S\sim D^m}(e(D,h) \leq f(S,h,h_1,...,h_K)) \geq 1-\delta$$

where $h$ is the classifier derived from the cross validation process on $m$ examples producing classifiers $h_1,...,h_K$, $f$ is the bound, and $e(D,h) = \Pr_{(x,y)\sim D}(h(x) \neq y)$ is the true error rate of $h$.

## What is Known

1. Devroye, Rogers, and Wagner: bounds for specific algorithms

2. Kearns and Ron: additional stability assumption $\Rightarrow$ the deviation for the classifier which learns on all the data is not much worse than for a test set of size $m/K$.

3. Blum, Kalai, and Langford: you can do at least as well as a test set of size $m/K$ using the randomized classifier which draws uniformly from the $K$ learned classifiers.

4. Bengio and Grandvalet: There is no unbiased estimator of variance.

5. Matti Kaariainen: safe derandomization of a stochastic classifier using unlabeled data without additional assumptions.

Prizes:

1. \$100 for any improvement (no extra assumptions allowed).

2. \$500 for a significant improvement (judge = John).

3. Something in between for something in between (judge = John).

Some other discussion: Machine Learning (Theory) blog

http://hunch.net under "problems"