It can not be emphasized enough that no claim whatsoever is being made in this paper that all algorithms are equivalent in practice, *in the real world. In particular, no claim is being made that one should not use cross-validation in the real world.*
—Wolpert, 1994

Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection

Ron Kohavi
Stanford University

IJCAI-95

(ronnyk@CS.Stanford.EDU)

Motivation & Summary of Results

You have ten induction algorithms. Which one is the best for a given dataset?

Answer: run them all and pick the one with the highest estimated accuracy.

1. Which accuracy estimation?
   Resubstitution? Holdout? Cross-validation? Bootstrap?

2. How sure would you be of your choice?

3. If you had spare CPU cycles, would you alway choose leave-one-out?

> For accuracy estimation: stratified 10 to 20-fold CV.
> For model selection: multiple runs of 3-5 CV.

Talk Outline

☞ ① Accuracy estimation: the problem.

② Accuracy estimation methods:

- ⋆ Holdout & random subsampling.
- ⋆ Cross-validation.
- ⋆ Bootstrap.

③ Experiments, recent experiments.

④ Summary.

Basic Definitions

1. Let $\mathcal{D}$ be a dataset of n labelled instances.

2. Assume $\mathcal{D}$ is an i.i.d. sample from some underlying distribution on the set of labelled instances.

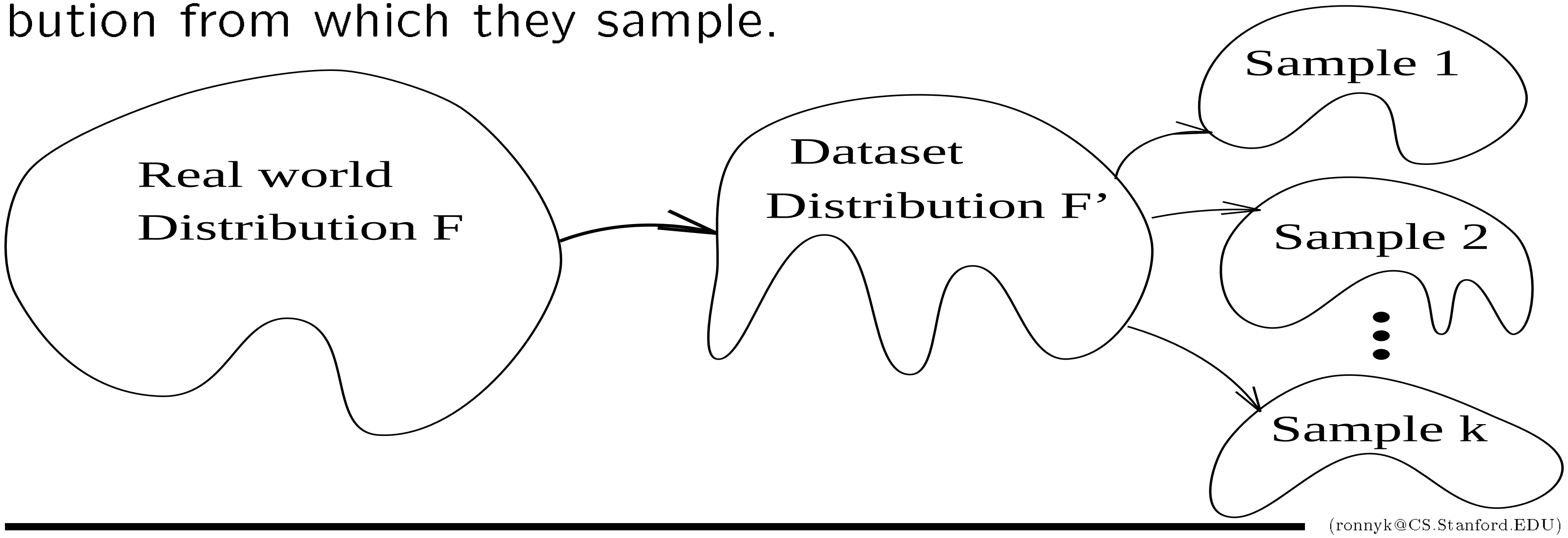
3. Let $\mathcal{I}$ be an induction algorithm that produces a classifier from data $\mathcal{D}$.

The Problem

Estimate the accuracy of the classifier induced from the dataset.

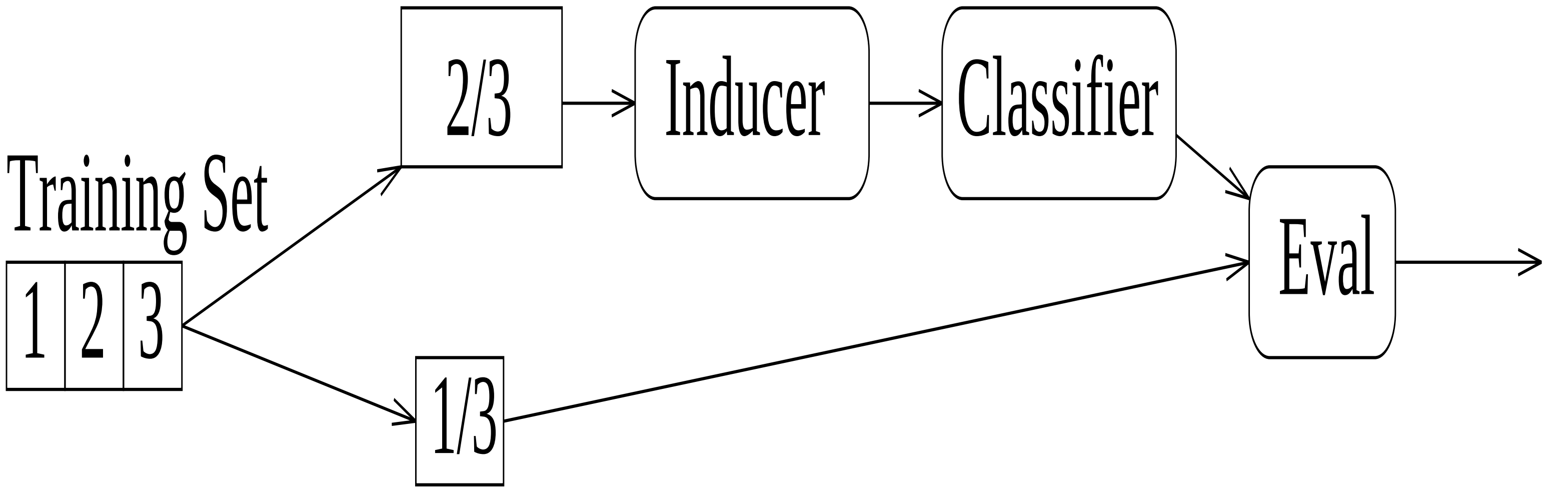
The **accuracy** of a classifier is the probability of correctly classifying a randomly selected instance from the distribution.

All resampling methods (holdout, cross-validation, bootstrap) will use a uniform distribution on the given dataset as a distribution from which they sample.



Holdout

Holdout partition the data into two mutually exclusive subsets called a training set and a test set. The training set is given to the inducer, and the induced classifier is tested on the test set.



Pessimistic estimator because only a portion of the data is given to the inducer for training.

Confidence Bounds

The classification of each test instance can be viewed as a Bernoulli trial: correct or incorrect prediction.

Let S be the number of correct classifications on the test set of size h, then S is distributed binomially and S/h is approximately normal with mean $\mathtt{acc}$ and variance of $\mathtt{acc} * (1 - \mathtt{acc})/h$.

$$\Pr\left\{-z < \frac{\mathtt{acc}_h - \mathtt{acc}}{\sqrt{\mathtt{acc}(1 - \mathtt{acc})/h}} < z\right\} \approx \gamma \qquad (1)$$

(ronnyk@CS.Stanford.EDU)

Random Subsampling

Repeating the holdout many times is called random subsampling.

Common mistake:

> One cannot compute the standard deviation of the samples' mean accuracy in random subsampling because the test instances are not independent.
>
> The t-tests you commonly see with significance levels are usually wrong.

Example: random concept (50% for each class). One induction algorithm predicts 0, the other 1. Given a sample of 100 instances, chances are it won't be 50/50, but something like 53/47. Leaving 1/3 out gives s.d.=8.7%.

Audience comments?

Pedro Domingos *et al.*, please keep the comments to the end.

Accuracy as defined here is the prediction accuracy on instances sampled from the parent population, not from the small dataset we have at hand.

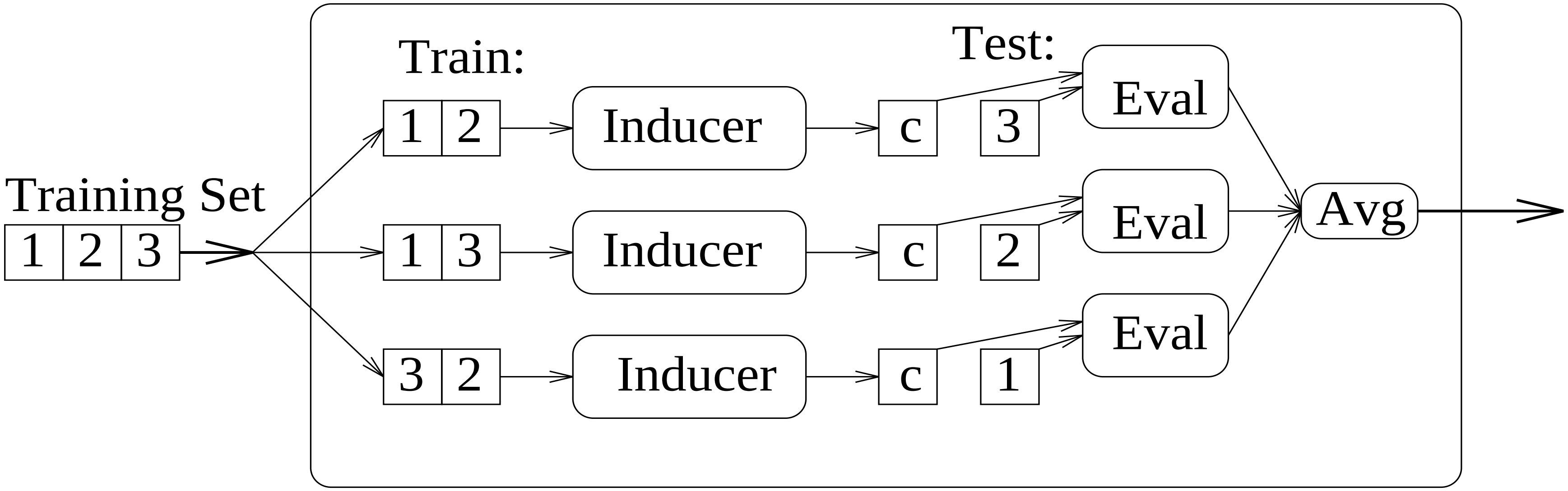
t-tests with random subsampling tests the hypothesis that one algorithm is better than another, assuming the distribution is uniform and each instance in the dataset has $1/n$ probability.

This is not an interesting hypothesis, except that we know how to compute it.

Cross Validation

In k-fold cross-validation, the dataset $\mathcal{D}$ is randomly split into k mutually exclusive subsets (the folds) $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ of approximately equal size.

The inducer is trained and tested k times; each time $t \in \{1, 2, \ldots, k\}$ it is trained on $\mathcal{D} \setminus \mathcal{D}_t$ and tested on $\mathcal{D}_t$.



Standard Deviation

Two ways to look at cross-validation:

1. A method that works for stable inducers, *i.e.*, inducers that do not change their predictions much when given similar datasets (see paper).

2. A method of computing the *expected accuracy* for a dataset of size $n - k/n$, where k is the number of folds.

Since each instance is used only once as a test instance, the standard deviation can be estimated as $\texttt{acc}_{\text{CV}} \cdot (1 - \texttt{acc}_{\text{CV}})/n$, where n is the number of instances in the dataset.

Failure of cross-validation/leave-one-out

Leave-one-out is n-fold cross-validation.

Fisher's iris dataset contains 50 instances of each class (three classes), leading one to expect that a majority inducer should have accuracy about 33%.

When an instance is deleted from the dataset, its label is a minority in the training set; thus the majority inducer predicts one of the other two classes and always errs in classifying the test instance.

The leave-one-out estimated accuracy for a majority inducer on the iris dataset is therefore 0%$\pm$0%

.632 Bootstrap

A **bootstrap sample** is created by sampling n instances uniformly from the data (with replacement).

Given a number b, the number of bootstrap samples, let $\epsilon 0_i$ be the accuracy estimate for bootstrap sample i on the instances not included in the sample. The .632 bootstrap estimate is defined as

$$\mathtt{acc}_{\text{boot}} = \frac{1}{b} \sum_{i=1}^{b} (0.632 \cdot \epsilon 0_i + .368 \cdot \mathtt{acc}_s)$$

where $\mathtt{acc}_s$ is the resubstitution accuracy estimate on the full dataset.

Bootstrap failure

The .632 bootstrap fails to give the expected result when the classifier can perfectly fit the data (*e.g.*, an unpruned decision tree or a one nearest neighbor classifier) and the dataset is completely random, say with two classes.

The apparent accuracy is 100%, and the $\epsilon 0$ accuracy is about 50%. Plugging these into the bootstrap formula, one gets an estimated accuracy of about 68.4%, far from the real accuracy of 50%.

Experimental design: Q & A

1. Which induction algorithm to use? C4.5 and Naive Bayes. Reason: FAST inducers.

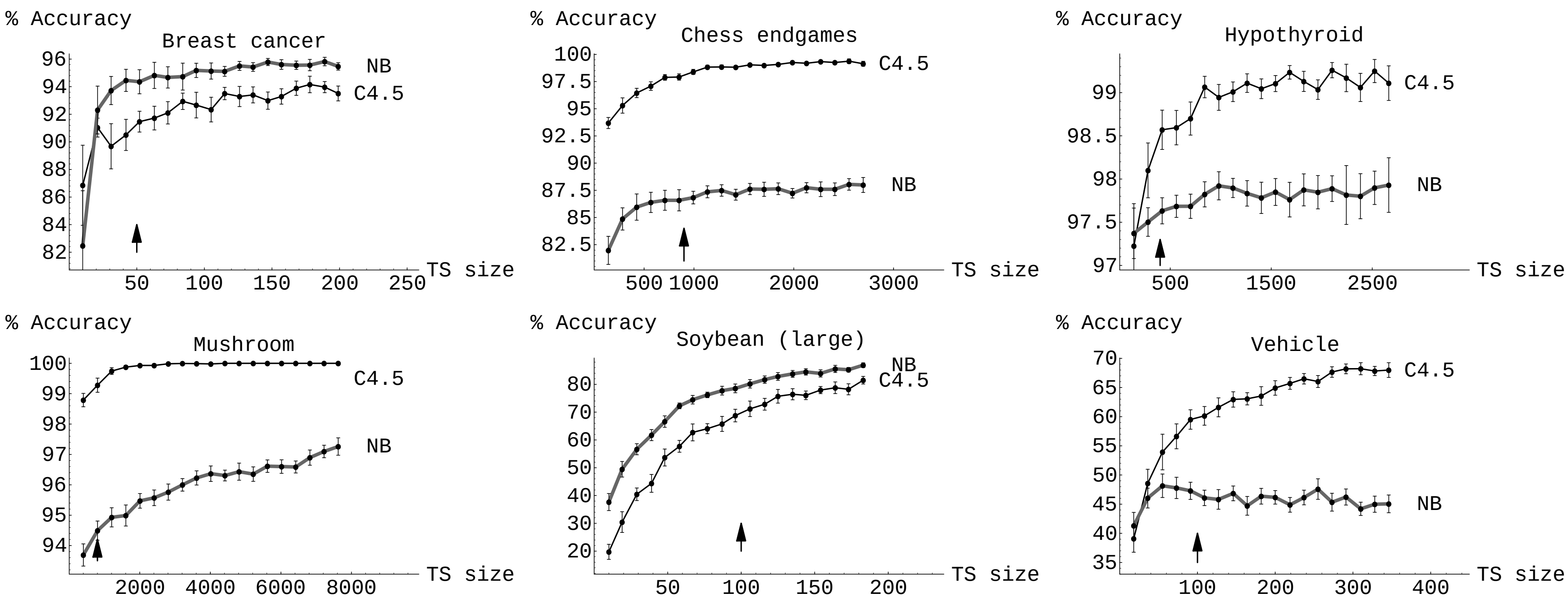
2. Which datasets to use? Seven datasets were chosen. Reasons:

   (a) Wide variety of domains.

   (b) At least 500 instances.

   (c) Learning curve that did not flatten too early.

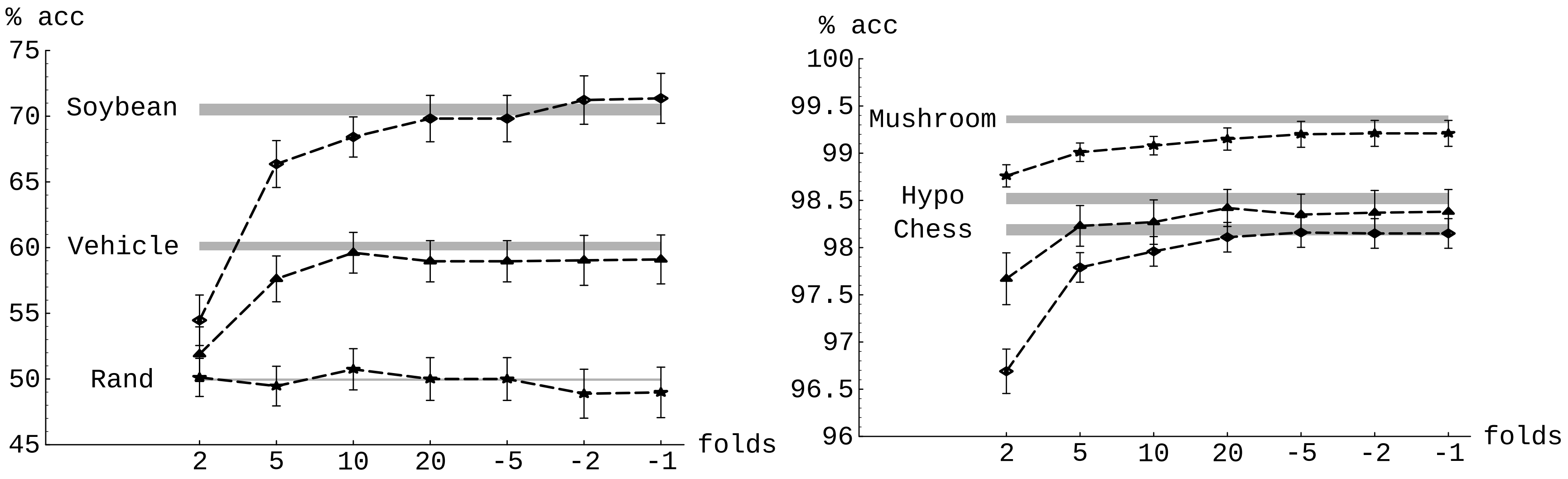
Learning Curves



Arrow indicate sampling points.

The bias (CV)

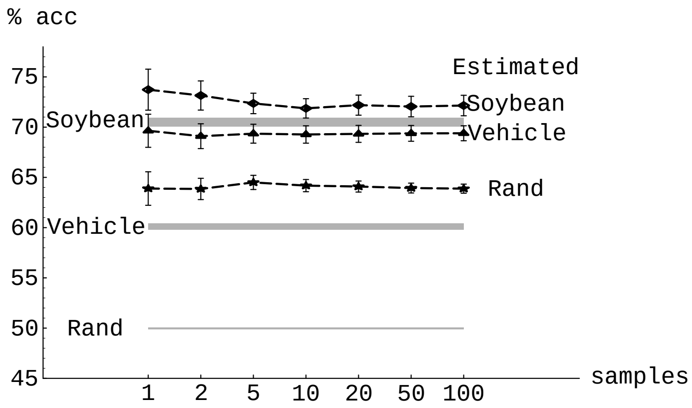
The **bias** of a method that estimates a parameter θ is defined as the expected estimated value minus the value of θ. An unbiased estimation method is a method that has zero bias.

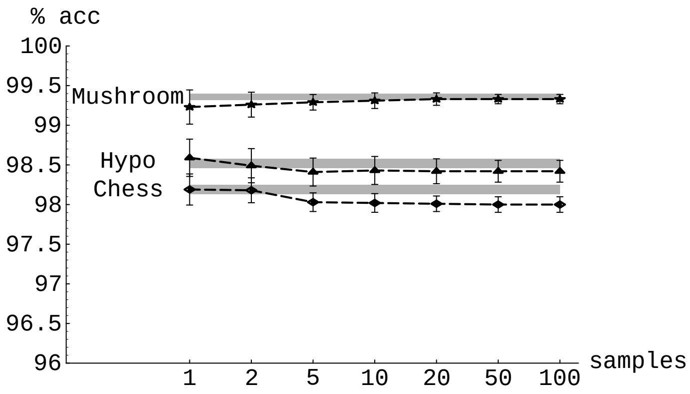


C4.5: The bias of cross-validation with varying folds. A negative k folds stands for leave-k-out. Error bars are 95% confidence intervals for the mean. The gray regions indicate 95% confidence intervals for the true accuracies.

Bootstrap Bias

Bootstrap is right on the mark for three out of six, but biased for soybean and extremely biased for vehicle and rand.

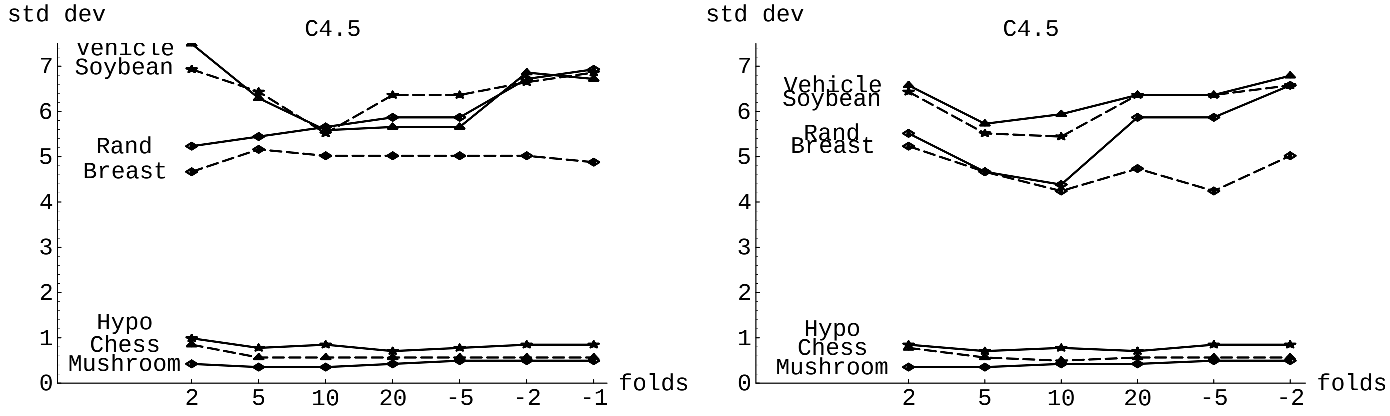




C4.5: The bias of bootstrap.

The Variance of CV

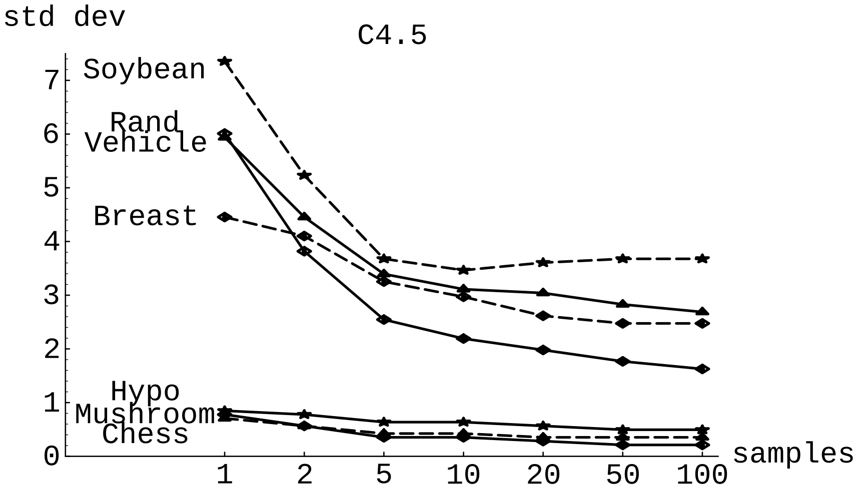
The variance is high at the ends: two fold and leave-{1,2}-out

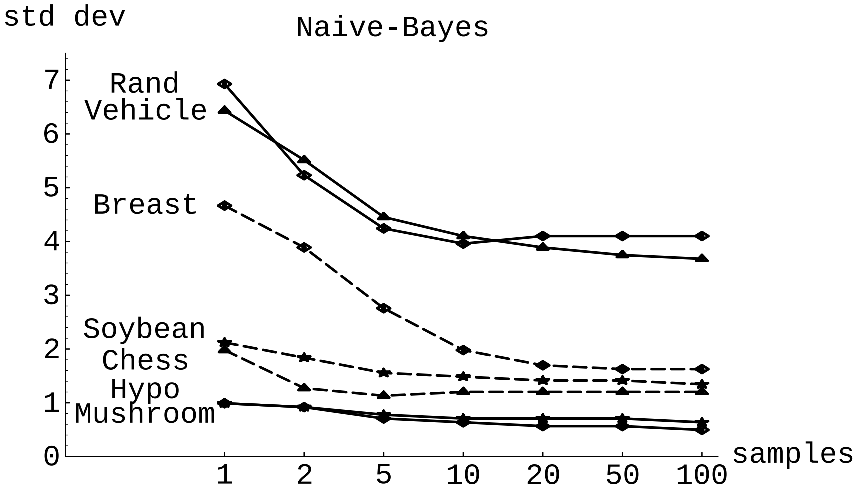


Regular (left), Stratified (right).

Stratification slightly reduces both bias and variance.

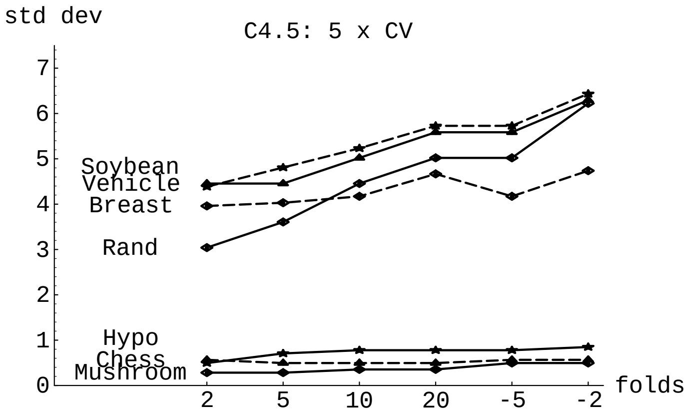
The Variance of Bootstrap

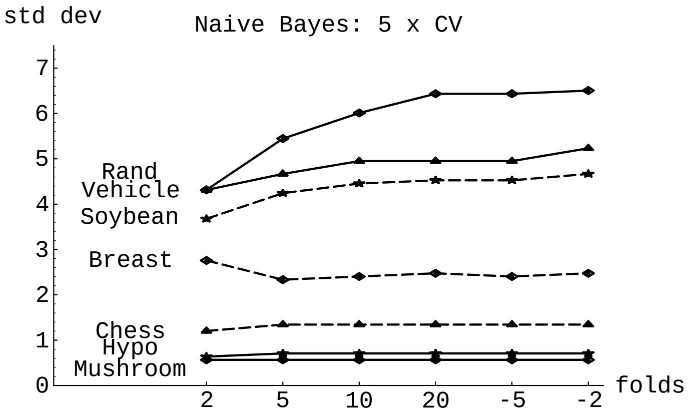




Multiple Times

To stabilize the cross-validation, we can execute CV multiple times, shuffling the instances each time.





The graphs show the effect on C4.5 and NB when CV was run five times. The variance for lower folds was due to the variance because of smaller dataset size.

Summary

1. Reviewed accuracy estimation: holdout, cross-validation, stratified CV, bootstrap. All methods fail in some cases.
2. Bootstrap has low variance, but is extremely biased.
3. With cross-validation, you can determine the bias/variance tradeoff that is appropriate, and run multiple times to reduce variance.
4. Standard deviations can be computed for cross-validation, but are incorrect for random subsampling (holdout).

Recommendation

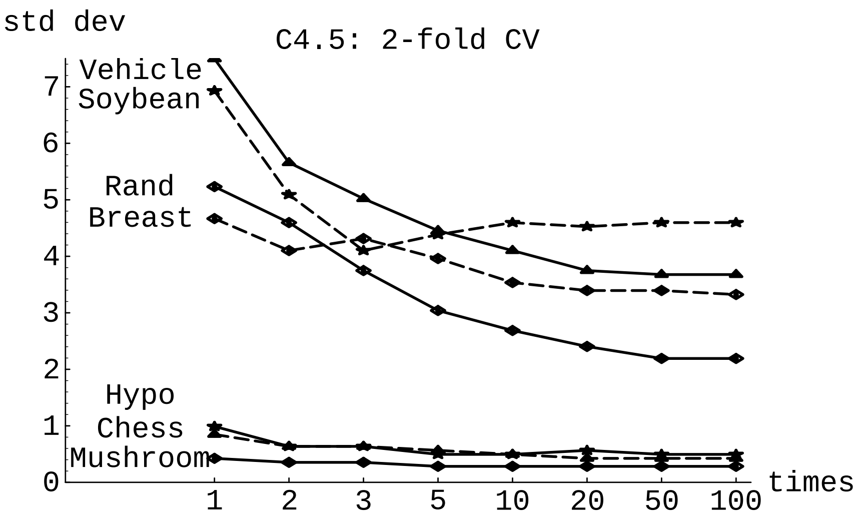
For accuracy estimation: stratified 10 to 20-fold CV.
For model selection: multiple runs of 3-5 folds (dynamic).

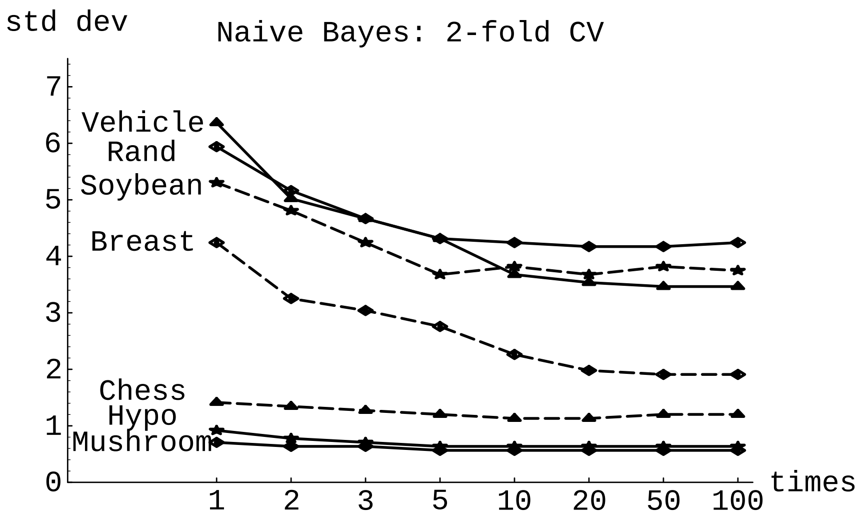
What is the "true" accuracy

To estimate the true accuracy at the sampling points, we sampled 500 times and estimated the accuracy using the unseen instances (holdout).

Dataset	no. of attr.	sample-size / total size	no. of categories	C4.5	Naive-Bayes
Breast cancer	10	50/699	2	91.37 ± 0.10	94.22 ± 0.10
Chess	36	900/3196	2	98.19 ± 0.03	86.80 ± 0.07
Hypothyroid	25	400/3163	2	98.52 ± 0.03	97.63 ± 0.02
Mushroom	22	800/8124	2	99.36 ± 0.02	94.54 ± 0.03
Soybean large	35	100/683	19	70.49 ± 0.22	79.76 ± 0.14
Vehicle	18	100/846	4	60.11 ± 0.16	46.80 ± 0.16
Rand	20	100/3000	2	49.96 ± 0.04	49.90 ± 0.04

How Many Times Should We Repeat?





Multiple runs with two-fold CV. The more, the better.