1 Validation vs. Testing

Suppose you have 10 classifier learning algorithms and a classifier problem. You don't know which learning algorithm is the best, so you want to apply all of the algorithms and choose the best one. In order to do this, you split the data into a training set (of size 300) and a test set (of size 100). For each learning algorithm you learn a classifier using the training set and test on the test set.

Let $c_1, ..., c_{10}$ be the 10 classifiers and $\hat{e}_1, ..., \hat{e}_{10}$ be the error rates of the test set on each of the classifiers.

Suppose you pick the classifier with the smallest test error (or, equivalently, the smallest true error bound based upon the test set).

1. Is it reasonable to report the test error rate as a predictor of the true error rate? (yes or no, and please explain why. If you can, quantify the answer.)

2. Does this answer change if our "10 classifier learning algorithms" is actually the same learning algorithm with 10 different choices of parameters?

2 A Problem of Science

Suppose you are the chair of the program committee of a machine learning conference. 10 papers have been submitted, and you must decide to publish at most one. Each of these is a paper where the authors report the results of a new learning algorithm they invented on a common (shared) learning problem with a common (shared) division of data between training and test sets. Let $\hat{e}_1, ..., \hat{e}_{10}$ be the error rates on the test set achieved in each of the 10 papers.

1. Which paper do you publish, if any? (state an answer and justify it)

2. Suppose you attend the conference. No list of unpublished papers exists, and you have only a vague idea how many papers were rejected. How much do you trust the results published at the conference? (answer and justify the answer)

Side note: For *real* machine learning conferences, there is no single person deciding what is accepted and what is rejected. Most results for new learning algorithms are *not* reported on a common dataset or division of the dataset. Also, most results are *not* reported in a common format or evaluation criteria. Each of these makes evaluation much more difficult.