Confidence Intervals

Observation: It does not matter how careful we are with our model evaluation techniques, there remains a fundamental uncertainty about the ability of our data set D to effectively represent our (possibly infinite) data universe.

This uncertainty reflects into our model evaluation. If D is a poor representation then the models we construct using D will generalize poorly to the rest of the data universe. If D is a good representation of the data universe then we can expect that our model will generalize well.

Here we will deal with this uncertainty using *confidence intervals*.

Perhaps most surprising is that we will use D itself in order to estimate this uncertainty using the *bootstrap*.

Confidence Intervals

First, let us define *error confidence intervals* formally.

Given a model error err_D over some data set D, then the error confidence interval is defined as the probability p that our model error err_D lies between some lower bound lb and some upper bound ub,

$$Pr\left(\text{lb} \leq \text{err}_D \leq \text{ub}\right) = p.$$

Paraphrasing this equation with $p = 95\%$:

> *We are 95% percent sure that our error err_D is not better than lb and not worse than ub.*

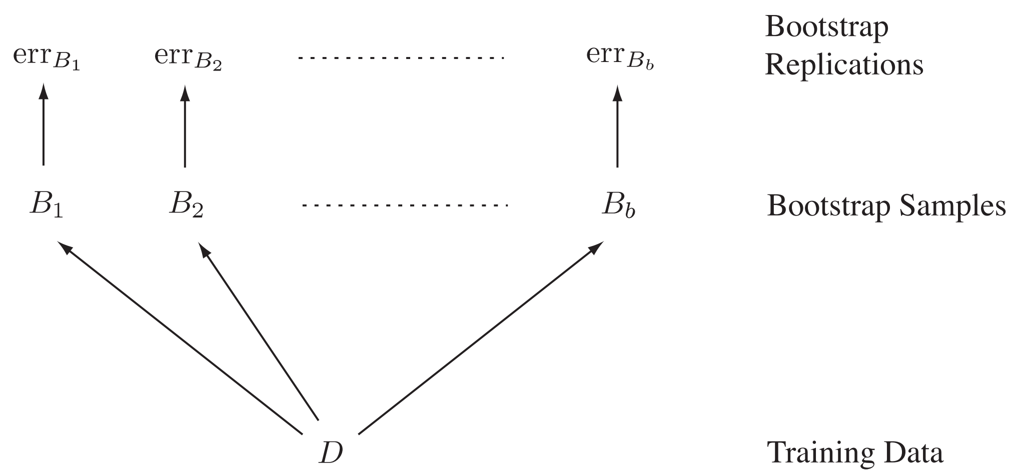
The Bootstrap

A particular effective and computationally straightforward way to estimate the lower and upper bounds of confidence intervals is the *bootstrap*.

What is remarkable about the bootstrap is that we use the data set D itself to capture the uncertainty with which it represents the data universe at large.

In the bootstrap we create b bootstrap samples of our data set D using sampling with replacement.

We use the variation among the bootstrap samples to compute the variation in the respective model errors.



The Bootstrap

given data set D
for $i = 1$ **to** 200 **do**
  $B[i] \leftarrow$ sample D with replacement, note $|B[i]| = |D|$.
  $\mathsf{err}[i] \leftarrow$ compute model error using parameter set (k^*, λ^*, C^*) and $B[i]$.
end for
sort err in ascending fashion
$\mathsf{ub} \leftarrow \mathsf{err}[195]$
$\mathsf{lb} \leftarrow \mathsf{err}[5]$
return $(\mathsf{lb}, \mathsf{ub})$

The algorithm to compute a 95% error confidence interval.

Model Comparisons

By now it should be clear that a single performance number computed on D is perhaps a poor indicator for models.

As an example, consider the model $\hat{f}_D[k^*, \lambda^*, C^*]$ with a cross-validated error,

$$\text{CVE}_D[k^*, \lambda^*, C^*] = 0.1,$$

and a 95% confidence interval $[0.08, 0.12]$. Consider another model $\hat{f}_D[k^\bullet, \lambda^\bullet, C^\bullet]$ with a cross-validated error,

$$\text{CVE}_D[k^\bullet, \lambda^\bullet, C^\bullet] = 0.05,$$

and a 95% confidence interval $[0.01, 0.09]$.

By just looking at the cross-validated error we are tempted to say that the second model is superior to the first model.

However, the confidence intervals *overlap*, meaning that the performance difference between the two models is *statistically not significant*.