“Bumping” and Active Set Selection

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

"Bumping"

Another/different thing one might do with bootstrap versions of a predictor is to "pick a winner" based on performance on the training data. This is the "bumping"/stochastic perturbation idea of HTF's Section 8.9. That is, let $\hat{f}^{*0} = \hat{f}$ be the predictor computed from the training data, and define

$$\widehat{b} = \underset{b=0,1,\ldots,B}{\arg\min} \sum_{i=1}^{N} \left(y_i - \hat{f}^{*b}(\mathbf{x}_i) \right)^2$$

and take

$$\hat{f}_{\text{bump}}(\mathbf{x}) = \hat{f}^{*\widehat{b}}(\mathbf{x})$$

The idea here is that if a few cases in the training data are responsible for making a basically good method of predictor construction perform poorly, eventually a bootstrap sample will miss those cases and produce an effective predictor.

Active set selection

Rick (Wen) Zhou in his ISU PhD dissertation made another use of bootstrapping, motivated by a real 2-class classification problem with "covariate shift." $\mathbf{x}$ values in an important test set were mostly unlike input vectors $\mathbf{x}_i$ available in a fairly small training set. With relatively little information available in the training set, highly flexible methods like nearest neighbor classification seemed unlikely to be effective. But a single simple application of a less flexible methodology (like one based on logistic regression) also seemed unlikely to be effective, because most test case input vectors were "near" at most "a few" training case input vectors and extrapolation of some kind was unavoidable.

What Zhou settled on and ultimately found to be relatively effective was to use (locally defined) bootstrap classifiers based on weighted bootstrap samples, with weights were chosen to depend upon an $\mathbf{x}$ at which one is classifying.

Zou's method

For a test input vector $\mathbf{x} \in \Re^p$ define weights for training case inputs $\mathbf{x}_i$ by

$$w_i(\mathbf{x}) = \exp\left(-\eta \|\mathbf{x} - \mathbf{x}_i\|^2\right)$$

for some appropriate $\eta > 0$. For $w(\mathbf{x}) = \sum_{i=1}^{N} w_i(\mathbf{x})$ a single "weighted bootstrap" sample tailored to the input $\mathbf{x}$ can be made by sampling N training cases iid according to the distribution over $i = 1, 2, \ldots, N$ with probabilities $p_i(\mathbf{x}) = w_i(\mathbf{x}) / w(\mathbf{x})$. Upon fitting a simple form of classifier to B such tailored samples and using majority voting of those classifiers, one has a classification decision for input $\mathbf{x}$. It is one that respects both the likelihood that training cases close to the input are most relevant to decisions about its likely response and the need to enforce simplicity on the prediction.