Bagging Generalities

Stephen Vardeman
Analytics Iowa LLC
ISU Statistics and IMSE

Bootstrap samples from the training set

A way to try to prevent a prediction methodology from producing $\hat{f}$ "too sensitive" to exact characteristics of a training sample is to employ "**boostrapping**." This involves some large number, B, of "bootstrap" samples of size N from the training set $\mathbf{T}$. Each of these, $\mathbf{T}_1^*, \mathbf{T}_2^*, \ldots, \mathbf{T}_B^*$, is a random sample *with replacement* of size N from $\mathbf{T}$. Applying a fixed method of prediction B times produces for each $b = 1, \ldots, B$

$$\text{predictor} \quad \hat{f}^{*b} \text{ based on } \quad \mathbf{T}_b^*$$

Note that (in cases where all training cases are different) for large N on average $\mathbf{T}_b^*$ fails to contain about 37% of training cases. The probability that a particular training case is missed in a bootstrap sample is $\left(1 - N^{-1}\right)^N \approx e^{-1} \approx .37$ for N of any reasonable size.

SEL bagging

"Bootstrap aggregation" or "**Bagging**" for SEL is then the use of

$$\hat{f}^{B}_{\text{bag}}(\mathbf{x}) \equiv \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(\mathbf{x})$$

The hope is to average (not-perfectly-correlated as they are built on not-completely-overlapping bootstrap samples) low-bias/high-variance predictors to reduce variance while maintaining low bias.

Limiting SEL bagged predictor

Even for fixed training set $\mathbf{T}$ and input $\mathbf{x}$ a bagging predictor $\hat{f}^B_{\text{bag}}(\mathbf{x})$ is *random* (varying with the selection of the bootstrap samples). Let E^* denote averaging over the creation of a single bootstrap sample and $\hat{f}^*$ be the predictor derived from such a bootstrap sample. Then

$$\mathsf{E}^*\hat{f}^*(\mathbf{x}) = \hat{f}_{\text{bag}}(\mathbf{x})$$

is the "true"/large-B bagging predictor with simulation-based approximation $\hat{f}^B_{\text{bag}}(\mathbf{x})$. (Unless the operations applied to a training set to produce $\hat{f}$ are linear, $\mathsf{E}^*\hat{f}^*(\mathbf{x})$ will differ from the predictor computed from the full training data, $\hat{f}(\mathbf{x})$.)

$\hat{f}^B_{\text{bag}}(\mathbf{x}) \rightarrow \mathsf{E}^*\hat{f}^*(\mathbf{x}) = \hat{f}_{\text{bag}}(\mathbf{x})$ as $B \rightarrow \infty$ by the law of large numbers.

Bagging classifiers

A bagged predictor in the 0-1 loss classification case is

$$\hat{f}^{*}_{\text{bag}}(\mathbf{x}) = \arg\max_{k} \sum_{b=1}^{B} I\left[\hat{f}^{*b}(\mathbf{x}) = k\right]$$

(a majority vote combination of the individual classification trees). One here expects that for each k a law of large numbers will imply that

$$\frac{1}{B}\sum_{b=1}^{B} I\left[\hat{f}^{*b}(\mathbf{x}) = k\right] \to P^{*}\left[\hat{f}^{*}(\mathbf{x}) = k\right] \text{ as } B \to \infty$$

so that there is a limiting classifier

$$\arg\max_{k} P^{*}\left[\hat{f}^{*}(\mathbf{x}) = k\right]$$

for which $\hat{f}^{*}_{\text{bag}}(\mathbf{x})$ is a simulation-based approximation.

Out-of-bag predictions for training cases

It is common practice to make a kind of running cross-validation estimate of error based on "out-of-bag" (OOB) samples as one builds a bagged predictor. Then, for each b suppose one keeps track of the set of (OOB) indices $I(b) \subset \{1, 2, \ldots, N\}$ for which the corresponding training vector does not get included in the bootstrap training set $\mathbf{T}_b^*$. In SEL contexts let

$$\hat{y}_{iB}^* = \frac{1}{\#\text{ of indices } b \leq B \text{ such that } i \in I(b)} \sum_{b \leq B \text{ such that } i \in I(b)} \hat{f}^{*b}(\mathbf{x}_i)$$

and in 0-1 loss classification contexts let

$$\hat{y}_{iB}^* = \arg\max_k \sum_{b \leq B \text{ such that } i \in I(b)} I\left[\hat{f}^{*b}(\mathbf{x}_i) = k\right]$$

OOB estimates of Err

Then in SEL regression contexts, a running cross-validation type of estimate of Err is

$$\text{OOB}(B) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_{iB}^{*}\right)^2$$

and a correpsonding estimate for 0-1 loss classification contexts is

$$\text{OOB}(B) = \frac{1}{N}\sum_{i=1}^{N} I\left[y_i \neq \hat{y}_{iB}^{*}\right]$$

As B increases, one can expect $\hat{f}_{\text{bag}}^{B}(\mathbf{x})$ to better approximate its limit $\hat{f}_{\text{bag}}(\mathbf{x})$ and OOB(B) to better approximate Err for $\hat{f}_{\text{bag}}(\mathbf{x})$.

Plotting and convergence of OOB(B)

Plotting OOB(B) versus B and determining when B is large enough that OOB(B) seems to have leveled off at some limiting value is a common way of determining when both 1) the extra/non-intrinsic noise introduced into the creation of a predictor by the bootstrap sampling has been averaged away and 2) a reliable measure of efficacy for the bagged predictor has been arrived at. Note that in spite of the fact that for small B the (random) predictor $\hat{f}_B^*$ is built on a small number of samples trees and is fairly simple, B is **not** really a *complexity* parameter, but **is** rather a *convergence* parameter.)

Where losses other than SEL or 0-1 loss are involved, exactly how to "bag" bootstrapped versions of a predictor is not altogether obvious, and even what might look like sensible possibilities can do poorly.