Classification Trees

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Classification trees/decision trees

The "classification trees" version of binary tree prediction is little different from the continuous y, SEL/"regression tree" version of this material. One needs only to define an empirical loss to associate with a given tree parallel to SSE used for SEL.

First note that in a K-class problem corresponding to a particular rectangle R_m is the fraction of training vectors with classification k,

$$\widehat{p_{mk}} = \frac{1}{\#\text{ training input vectors in } R_m} \sum_{\substack{i \text{ with } \mathbf{x}_i \\ \text{in } R_m}} I\left[y_i = k\right]$$

and a plausible classifier based on I rectangles is

$$\hat{f}_I\left(\mathbf{x}\right) = \arg\max_k \widehat{p_{m(\mathbf{x})k}}$$

the class that is most heavily represented in the rectangle to which $\mathbf{x}$ belongs.

Rectangle splitting criteria

The empirical mis-classification rate for this predictor (that can be used as a rectangle-splitting criterion) is

$$\overline{\text{err}} = \frac{1}{N}\sum_{i=1}^{N} I\left[y_i \neq \hat{f}_l\left(\mathbf{x}_i\right)\right] = \frac{1}{N}\sum_{m=1}^{l} N_m\left(1-\widehat{p_{mk(m)}}\right)$$

where $N_m = \#$ training input vectors in R_m, and $k\left(m\right) = \arg\max_k \widehat{p_{mk}}$.

Two other popular splitting criteria are "the Gini index"

$$\frac{1}{N}\sum_{m=1}^{l} N_m\left(\sum_{k=1}^{K}\widehat{p_{mk}}\left(1-\widehat{p_{mk}}\right)\right)$$

and the so-called "cross entropy"

$$-\frac{1}{N}\sum_{m=1}^{l} N_m\left(\sum_{k=1}^{K}\widehat{p_{mk}}\ln\left(\widehat{p_{mk}}\right)\right)$$

Classification tree growing

The latter two criteria are average (across rectangles) measures of "purity" (near degeneracy) of training set response distributions in the rectangles. Upon adopting one of these forms and using it to replace SSE in the regression tree discussion, one has a classification tree-building methodology. HTF suggest using the Gini index or cross entropy for tree growing and any of the indices (but most typically the empirical mis-classification rate) for "tree pruning" according to cost-complexity (an idea to be discussed next).