Predictors Constant on Rectangles

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Predictors constant on rectangles

This begins something genuinely new to our discussion. That is the search for good predictors that are constant on p-dimensional "rectangles" in the input space, that is on subsets of $\Re^p$ of the form

$$R = \{\mathbf{x} \in \Re^p \,|\, a_1 < x_1 < b_1 \text{ and } a_2 < x_2 < b_2 \ \ldots \text{ and } a_p < x_p < b_p\}$$

for (possibly infinite) values $a_j < b_j$ for $j = 1, 2, \ldots, p$. The basic idea is that if the values a_j and b_j can be chosen so that y_is corresponding to vectors of inputs $\mathbf{x}_i$ in a training set in a particular rectangle are "homogeneous," then a corresponding SEL predictor using training set "rectangle mean responses" or a 0-1 loss classifier using training set "rectangle majority classes" might be approximately optimal. (This is essentially the same motivation provided for nearest neighbor rules.)

Invariance to monotone transform of inputs

The search for good predictors constant on rectangles is fundamentally an algorithmic matter, rather than something that will have a nice closed form representation (it is not like ridge regression for example). But (provided "fast" algorithms can be identified) it has things that make it very attractive.

For one thing, there is complete **invariance to monotone transformation of numerical features**. It is irrelevant to searches for good boundaries for rectangles whether a coordinate of the input $\mathbf{x}$ is expressed on an "original" scale or a log scale or on another (monotone transform of the original scale). The same predictor/predictions will result. This is a very attractive and powerful feature and is no doubt partly responsible for the popularity of rectangle-based predictors as building blocks for more complicated methods (like "boosting trees").

Intuitively simple predictors

The structure of predictors constant on rectangles is also an intuitively appealing one, easily explained and understood. This helps make them very popular with non-technical consumers of predictive analytics.

We will consider two rectangle-based prediction methods, the first (CART) using binary tree structures and the second (PRIM) employing a kind of "bump-hunting" logic.