classification and regression trees

tree-based models consist of one or more nested ‘if-then’ statements for the predictors that partition the data

within these partitions, a model is used to predict the outcome

classification and regression trees

grow a binary tree

at each node, “split” the data in to two “daughter” nodes

bottom nodes are “terminal” nodes

for regression the predicted value at a node is the *average* response variable for all observations in the node

for classification the predicted class is the *most common class* in the node (majority vote)

pioneers

Morgan and Sonquist (1963)

Breiman, Friedman, Olshen, Stone (1984). *CART*

Quinlan (1993)

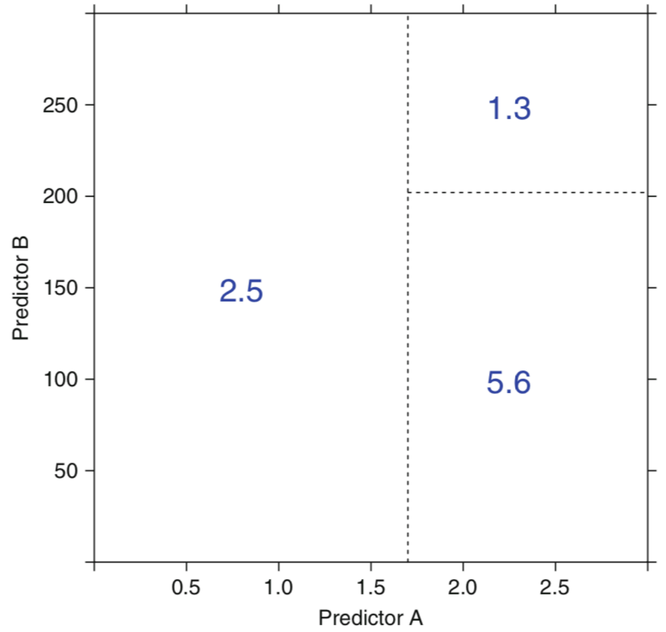
the if-then statements generated by a tree define a unique route to one terminal node for any sample

a rule is a set of if-then condi- tions (possibly created by a tree) that have been collapsed into independent conditions.

if predictor A >= 1.7 and predictor B >= 202.1 then outcome = 1.3

if predictor A >= 1.7 and predictor B < 202.1 then outcome = 5.6

if predictor A < 1.7 then outcome = 2.5



tree-based and rule-based models are popular modeling tools for a number of reasons.

they generate a set of conditions that are highly interpretable and are easy to implement

they can effectively handle many types of predictors (sparse, skewed, continuous, categorical, etc.) without the need to preprocess them

these models do not require the user to specify the form of the predictors' relationship to the response like, for example, a linear regression model requires

these models can effectively handle missing data and implicitly conduct feature selection, characteristics that are desirable for many real-life modeling problems

models based on single trees or rules, however, do have particular weaknesses

two well-known weaknesses are

(1) model instability (i.e., slight changes in the data can drastically change the structure of the tree or rules and, hence, the interpretation)

(2) less-than-optimal predictive performance.

the latter is due to the fact that these models define rectangular regions that contain more homogeneous outcome values. If the relationship between predictors and the response cannot be adequately defined by rectangular subspaces of the predictors, then tree-based or rule-based models will have larger prediction error than other kinds of models.

Basic regression trees partition the data into smaller groups that are more homogenous with respect to the response. To achieve outcome homogeneity, regression trees determine:

The predictor to split on and value of the split

The depth or complexity of the tree

The prediction equation in the terminal nodes