MA8701 General Statistical Methods
Spring 2017

REGRESSION TREE AND PRUNING: AN EXAMPLE

Bo Lindqvist

Notation is as on page 308 in *Elements of Statistical Learning.* See also page 309 in *An Introduction to Statistical Learning.*

Figure 1 shows a regression tree T_0 grown from a given training set of data $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, N\}$ where the predictors are $\mathbf{x} = (x_1, x_2)$. The tree corresponds to the partition shown to the right. By convention the terminal nodes R_j are numbered from top to bottom in the tree, in the order that the splits are made.

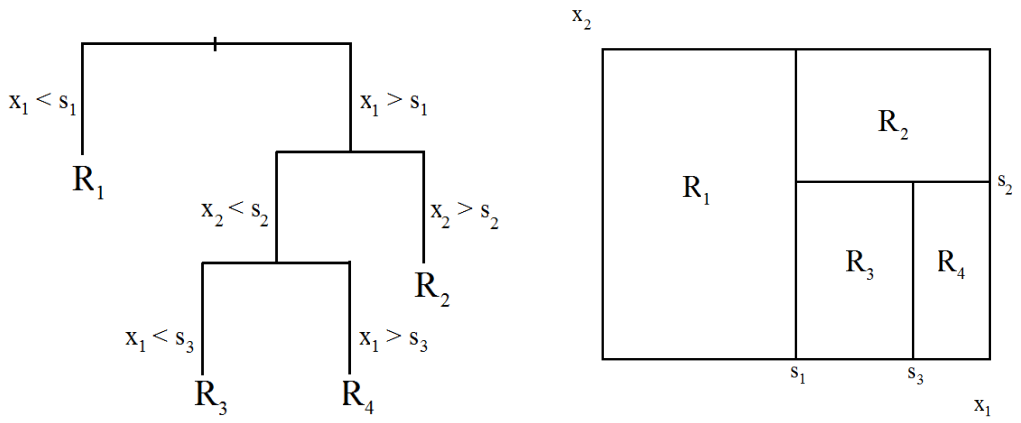

Figure 1: The full tree T_0 to the left, and the corresponding partition of the space of predictors (x_1, x_2) to the right.

Let $|T|$ mean the number of terminal nodes for a given tree T. Then define, as in (9.16) of *Elements*,

$$C_\alpha(T) = \sum_{m=1}^{|T|} \sum_{\mathbf{x}_i \in R_m} (y_i - \hat{c}_m)^2 + \alpha |T|$$

For simplicity, write

$$Q(T) = \sum_{m=1}^{|T|} \sum_{\mathbf{x}_i \in R_m} (y_i - \hat{c}_m)^2$$

The task is to find, for each $\alpha \geq 0$, the subtree of T_0 that minimizes $C_\alpha(T)$. This is done by successive pruning as described below.

The pruning process starts from the bottom of the tree, with the highest numbered terminal events R_3 and R_4. Pruning at the node corresponding to these two terminal nodes leads to the upper tree T_1 in Figure 2. The new terminal node R_{34} now corresponds to the union of the sets R_3 and R_4 to the right in Figure 1.

Next, the tree T_2 is produced by collapsing R_2 and R_{34} to R_{234}, corresponding to $R_2 \cup R_3 \cup R_4$ (see middle tree of Figure 2).

Finally, the tree T_3 (bottom of Figure 2) is the trivial one where all the R_i are collapsed into one set R_{1234}. This means that the same value of Y will be predicted for all values of the predictor $\mathbf{x}$.

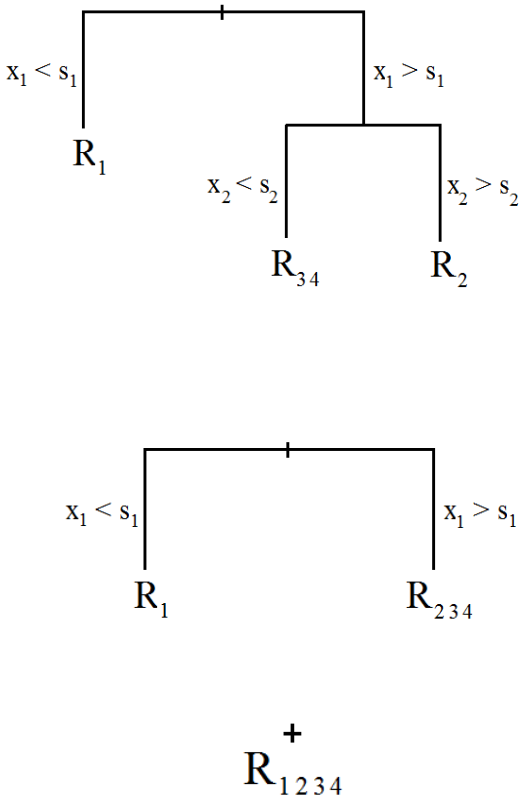

Figure 2: The trees T_i for $i = 1, 2, 3$ (from upper to bottom).

If we have the actual training data, it is straightforward to calculate the $Q(T_i)$ for the given trees. (How?)

Suppose the results are:

$$\begin{aligned} Q(T_0) &= 3.75 \\ Q(T_1) &= 4.00 \\ Q(T_2) &= 4.10 \\ Q(T_3) &= 5.00 \end{aligned}$$

(Why are the $Q(T_i)$ increasing with i?)

Taking $Q(T_0)$ as the starting point, calculate for each T_i, $(i = 1, 2, 3)$ the increase in $Q(\cdot)$ per decrease in number of terminal events:

$$\begin{aligned} T_1: \quad & \frac{Q(T_1) - Q(T_0)}{|T_0| - |T_1|} = \frac{4.00 - 3.75}{4 - 3} = 0.25 \\ T_2: \quad & \frac{Q(T_2) - Q(T_0)}{|T_0| - |T_2|} = \frac{4.10 - 3.75}{4 - 2} = 0.175 \\ T_3: \quad & \frac{Q(T_3) - Q(T_0)}{|T_0| - |T_3|} = \frac{5.00 - 3.75}{4 - 1} = 0.417 \end{aligned}$$

The minimum is hence for T_2. The sequence of trees so far is hence $T_0 \to T_2$. The algorithm continues by comparing the succeding trees T_i to the last found tree, which here was T_2. Since we have only one more tree, T_3, there is no choice here, but we still calculate

$$T_3: \quad \frac{Q(T_3) - Q(T_2)}{|T_2| - |T_3|} = \frac{5.00 - 4.10}{2 - 1} = 0.90$$

The full sequence of trees obtained is hence:

$$T_0 \to T_2 \to T_3$$

The theory of CART says that each of these trees equal $T(\alpha)$ for a certain range of α. It is in our example not difficult to verify that we have

$$T(\alpha) = \begin{cases} T_0 & ; \ 0 \leq \alpha < 0.175 \\ T_2 & ; \ 0.175 \leq \alpha < 0.90 \\ T_3 & ; \ 0.90 \leq \alpha \end{cases}$$

Cross-validation

In practice the optimal α is determined by cross-validation as follows.

First, divide the training set into K folds.

For each $k = 1, \ldots, K$:

1. Grow the tree T_0^{-k} in the same manner as T_0 was grown, using all but the kth fold of data.
2. Find the $T^{-k}(\alpha)$ for each each α in the same manner as describe in this note.
3. Calculate

$$q_k(\alpha) = \sum_{i \in k\text{th fold}} (y_i - \text{ predicted value from } \mathbf{x}_i \text{ using } T^{-k}(\alpha))^2$$

Then the cross-validated error measure for a given value of α is

$$CV(\alpha) = \frac{1}{N} \sum_{k=1}^{K} q_k(\alpha)$$