How to Build a Tree?

Three elements:

1. Where to split?
2. When to stop?
3. How to predict at each leaf node?

Prediction at Leaf Nodes

Each leaf node (corresponds to region R_m) contains some samples.

Assign the prediction for a leaf node to be the average (of the response variable Y).

$$\hat{f}(X) = \sum_m c_m I\{X \in R_m\}.$$

$$\min_{c_m} \sum_{i=1,\ x_i \in R_m}^{n} (y_i - c_m)^2,$$

$$\Longrightarrow c_m = \text{ average of } y_i\text{'s whose } x_i \in R_m$$

Where to Split?

- A split is denoted by (j, s): split the data into two parts based on whether "var j < value s".
- For each split, define a split criterion $\Phi(j, s)$
  - deduction of RSS for regression
- Trees are built in a top-down greedy fashion. Start with the root: try all possible variables $j = 1 : p$ and all possible split values[a], and pick the best split, i.e., the split having the best Φ value. Now, data are divided into the left node and right node. Repeat this procedure in each node.

---

[a]For each variable j, sort the n values (from n samples), and choose s to be a middle point of two adjacent values. So at most $(n-1)$ possible values for s.

Goodness of Split $\Phi(j, s)$

For Regression tree, we look at the deduction of RSS if we split samples at node t into t_R and t_L:

$$\Phi(j, s) = \mathsf{RSS}(t) - \Big[\mathsf{RSS}(t_R) + \mathsf{RSS}(t_L)\Big],$$

where

$$\begin{aligned} \mathsf{RSS}(t) &= \sum_{x_i \in t} (y_i - c_t)^2, \\ c_t &= \mathsf{AVE}\{y_i : x_i \in t\}. \end{aligned}$$

Note that $\Phi(j, s)$ is always positive if we split the data into two groups (even randomly), unless the mean of the left node and the one of right node are the same.

Issues: Split Categorical Predictors

- For a categorical predictor with m levels, there are $2^{m-1} - 1$ possible partitions of the m labels into two groups.
- However, for regression with square error, the computation simplifies: order the m levels by their mean values of Y, and then split the categorical variable as if it were an ordered predictor — there are only $(m - 1)$ potential splits.

Issues: Missing Predictor Values

- Discard any observation with missing values $\longrightarrow$ serious depletion of the training set.
- Splitting criteria are evaluated on non-missing observations.
- Once a split (j, s) is determined, what to do with observations missing X_j?

– Find surrogate variables that can predict the binary outcome "$X_j < s$" and "$X_j \geq s$" using a one-split tree.

– Rank those surrogate variables along with the blind rule "go with majority".

– Any observation that is missing X_j is then classified with the first surrogate variable, or if missing that, the second surrogate variable (or the blind rule) is used, and etc.

When to Stop?

- A simple one : stop splitting at a node if the gain from any split is less than some pre-specified threshold.
- BUT, this is short-sighted.
- Another strategy: grow a large tree and then prune it (i.e., cut some branches).