Regression trees

For quantitative response Y & p-variate features $\underline{X}$,

a <u>tree</u> is a model for $f(\underline{x})$ in $Y = f(\underline{x}) + \varepsilon$

such that

$$f(\underline{x}) = \sum_{i=1}^{M} c_i \, \mathbb{1}_{[\underline{x} \in R_i]}$$

where $R_1, \dots, R_M$ form a disjoint union of the feature space.

Algorithm

① Split featurespace into

$$R_1(m,s) = \{\underline{x} \mid x_m < s\} \qquad R_2(m,s) = \{\underline{x} \mid x_m \geq s\}$$

(m = variable over which we split; s = value of split)

(2) Choose m + s to minimize

$$\sum_{\underline{x}_i \in R_1(m,s)} (y_i - \hat{c}_1)^2 + \sum_{\underline{x}_i \in R_2(m,s)} (y_i - \hat{c}_2)^2$$

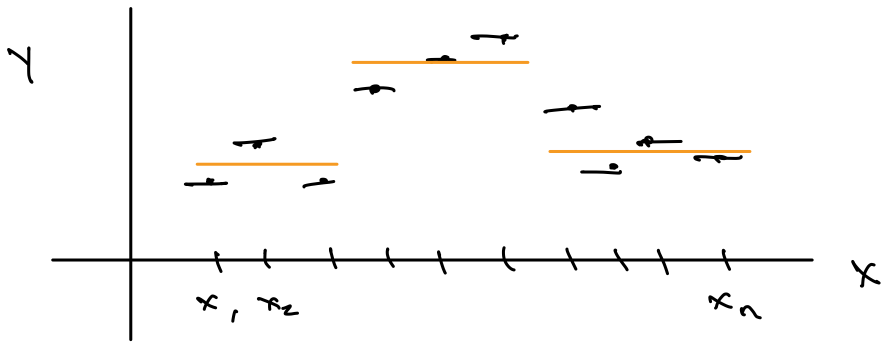
where $\hat{c}_1 = \text{ave}\{y_i \mid \underline{x}_i \in R_1(m,s)\}$

$\hat{c}_2 = \text{ave}\{y_i \mid \underline{x}_i \in R_2(m,s)\}$

(3) Within each of R_1 + R_2, repeat steps 1/2.

$\boxed{\text{Note}}$ This looks like a lot of work to go through all m's + s's, but the fit is just simple averages + squared errors for each trial value, so is pretty fast in practice.

Note



Trees can overfit data easily, so 1st trick is to have a stopping criterion, e.g. no fewer than 5 data points per terminal node.

Trees have little bias, but they suffer from high variance

One option: Grow a big/deep tree T_0, and prune to obtain a subtree. Can do this through cost-complexity pruning AKA weakest link pruning. Consider a sequence of trees indexed by $\alpha \geq 0$ where there is a unique

$T \subseteq T_0$ that minimizes

$$\sum_{m=1}^{|T|} \left(\sum_{i \mid \underline{x}_i \in R_m} (y_i - \hat{c}_i)^2 \right) + \alpha |T|$$

where $|T|$ = # of nodes in tree.

This is just another regularization idea where

$\alpha = 0 \Rightarrow$ fit the most complex tree

$\alpha \to \infty \Rightarrow$ trees become simple, eventually

$$Y = c + \varepsilon$$