STATS 769

Tree-based Models

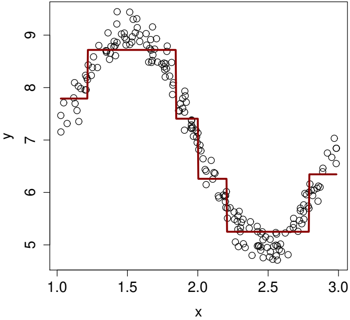
Yong Wang
Department of Statistics
The University of Auckland

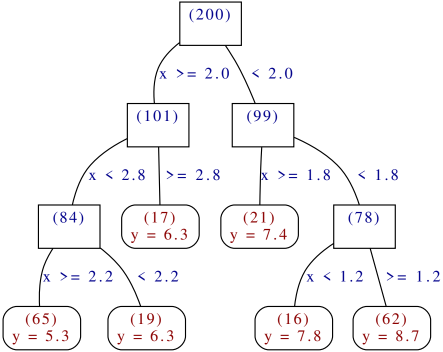
2021

Tree-based Models

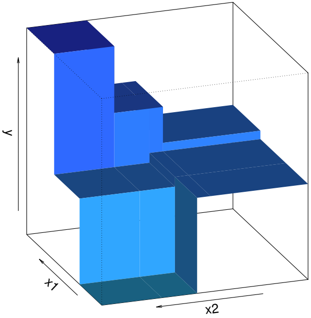
- Tree-based models are also known as tree-structured models.
- There are two major families:
  - Regression trees
  - Classification trees (or decision trees)
- There are many variants.
- In their basic form, they are essentially step functions.

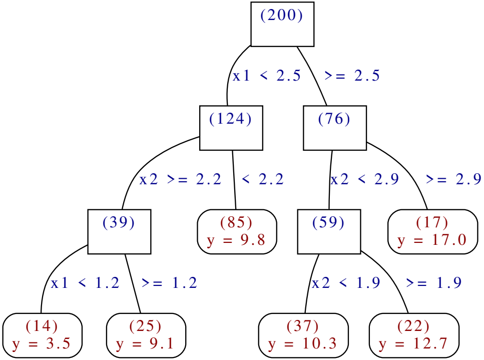
Regression Tree: Univariate X





Regression Tree: Bivariate X





Regression Trees

- It is tree-stuctured and always has a root node.
- Two types of nodes: internal vs. leaf/terminal nodes.
- Each internal node (of a binary tree) has two child nodes: left and right.
- At each internal node, there is a splitting criterion of the form $x_j < c$, which sends an observation to either the left or the right child node.
- Starting at the root, any observation will eventually be sent down to a leaf node, where the prediction for Y takes place (which is a constant value at each leaf node).

Building a Regression Tree

- Building a regression tree (and a classification tree) consists of two stages: tree growing *vs.* tree pruning.
- Tree growing is conducted in a top-down fashion.
  - Starting at the root, at each new node it looks for the optimal predictor variable x_j and the optimal cutoff point c.
  - The splitting criterion $x_j < c$ partitions observations into the left and right groups (child nodes).
  - The optimality means maximising the variation reduction

    $$\mathrm{VR} = v(\mathrm{current}) - v(\mathrm{left}) - v(\mathrm{right}),$$

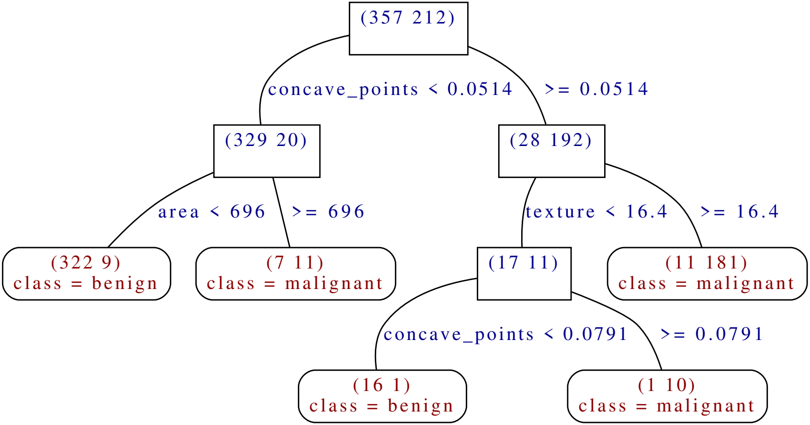
    where $v(\mathrm{node}) = \mathrm{TSS} \equiv \sum_i (y_i - \bar{y})^2$.
  - If deviance (i.e., -2 log-likelihood) is used in place of variation, then $v(\mathrm{node}) = n \log(\mathrm{TSS})$ (with some constant ignored).
  - For a categorical predictor, a splitting criterion may look like $x_j \in \{a, d, e\}$ *vs.* $x_j \in \{b, c\}$, where a, b, c, d, e are the levels of the variable.
  - Keep growing the tree until there is little variation or few observations remaining at a node.

Building a Regression Tree II

- Tree growing ends up with an unpruned tree, which is an overfitted model.
  - By taking the average of the response values of those reaching a node, each node has a constant model.
  - Each internal node also has a subtree model.
- Tree pruning is conducted in a bottom-up fashion.
  - By traversing all nodes and evaluating first those lower in the tree, it makes a choice at each internal node between its constant and subtree models.
  - If the constant model is chosen, then the subtree model is completely removed (or pruned).
  - The choice can be made using a model selection criterion (or cost-complexity criterion), and cross-validation can be used to determine the penalty (cost-complexity) parameter value.
- The final tree is known as a pruned tree.

Classification Trees

- Tumor diagnosis: `benign` *vs.* `malignant`



Building a Classification Tree

- Virtually the same as building a regression tree.
- For a subset data of size n, denote by n_j the number of observations in class j, and $p_j = n_j / n$.
- There are several commonly-used measures as the "variation" of the response for classification problems
  - Deviance/Entropy: $v(\mathrm{node}) = -2\sum_{j=1}^{J} n_j \log(p_j)$.
  - Gini index: $v(\mathrm{node}) = -\sum_{j=1}^{J} n_j p_j$.
- By taking the majority vote of the response values of those reaching the node, each node has one class label for prediction, along with the posterior probabilities for all classes.

Pros and Cons of Tree-based Models

Pros

- Little effort for data preparation
- Automatic variable selection
- Easy to deal with
  - Variables of different types
  - Irregular/nonlinear relationships
  - Missing values
- Interpretable results

Cons

- Accuracy may not be very high
  - Step functions are not even continuous
- Unstable
  - A small pertubation in the data can result in a completely different tree structure.

Recommended Readings

ISLv2 (basics):

- Sections 8.1
- Labs: Sections 8.3.1, 8.3.2

ESL (advanced):

- Section 9.2

Classical Book (CART):

- Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification and Regression Trees*. Wadsworth, Belmont CA.