Chapter 5. Tree-based Methods

Wei Pan

Division of Biostatistics, School of Public Health, University of Minnesota,
Minneapolis, MN 55455
Email: panxx014@umn.edu

PubH 7475/8475

Regression And Classification Tree (CART)

- §9.2: Breiman et al (1984).
  $\approx$ C4.5 (Quinlan 1993).
- Main idea: approximate any $f(x)$ by a piece-wise constant $\hat{f}(x)$.
- Use recursive partitioning: Fig 9.2,
  1) Partition the x space into two regions R_1 and R_2 by $x_j < c_j$;
  2) Partition R_1, R_2;
  3) Then their sub-regions, ... until the model fits data well.
- $\hat{f}(x) = \sum_m c_m I(x \in R_m)$.
  can be represented as a (decision) tree.

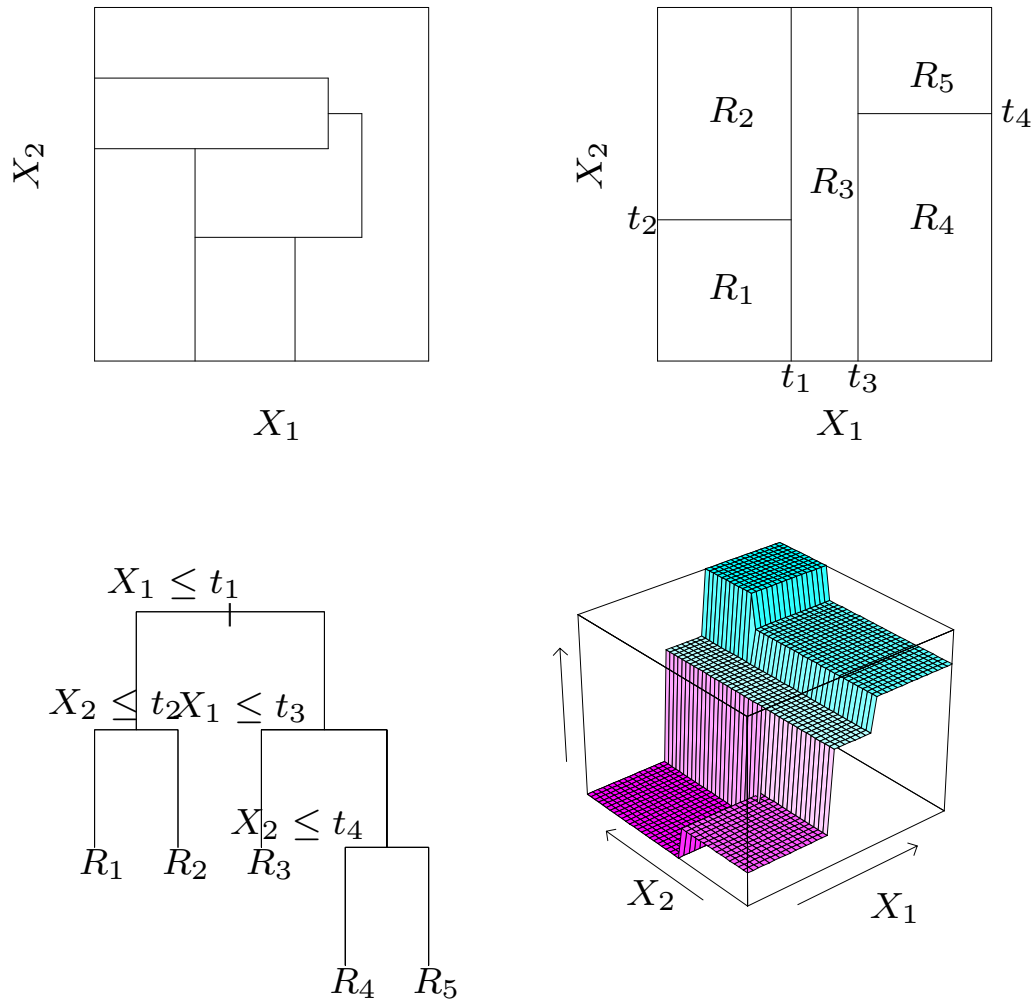

FIGURE 9.2. *Partitions and CART. Top right panel shows a partition of a two-dimensional feature space by recursive binary splitting, as used in CART, applied to some fake data. Top left panel shows a general partition that cannot be obtained from recursive binary splitting*

Regression Tree

- Y: continuous.
- Key: 1) determin splitting variables and split points (e.g. $x_j < t_j$); $\Longrightarrow$ R_1, R_2, ...;
  2) determine c_m in each R_m.
- in 1), use a sequential or greedy searchfor each j and s: find $x_j < s$ s.t.
  $R_1(j,s) = \{x | x_j < s\}$, $R_2(j,s) = \{x | x_j \geq s\}$,
  $\min_{j,s}[\min_{c_1} \sum_{X_i \in R_1(j,s)} (Y_i - c_1)^2 + \min_{c_2} \sum_{X_i \in R_2(j,s)} (Y_i - c_2)^2]$.
- in 2), given R_1 and R_2,
  $\hat{c}_k = \text{Ave}(Y_i | X_i \in R_k\}$ for $k = 1, 2$.
- Repeat the process on R_1 and R_2 respectively, ...
- When to stop?
  Have to stop when having all equal or too few Y_i's in R_m;
  Tree size gives a model complexity!

- A strategy: first grow a large tree, then prune it.
- Cost-complexity criterion for tree T:

$$C_\alpha(T) = RSS(T) + \alpha|T| = \sum_m \sum_{X_i \in R_m} (Y_i - \hat{c}_m)^2 + \alpha|T|,$$

where $|T|$ is # of terminal nodes (leaves) and $\alpha > 0$ is a tuning parameter to be determined by CV.

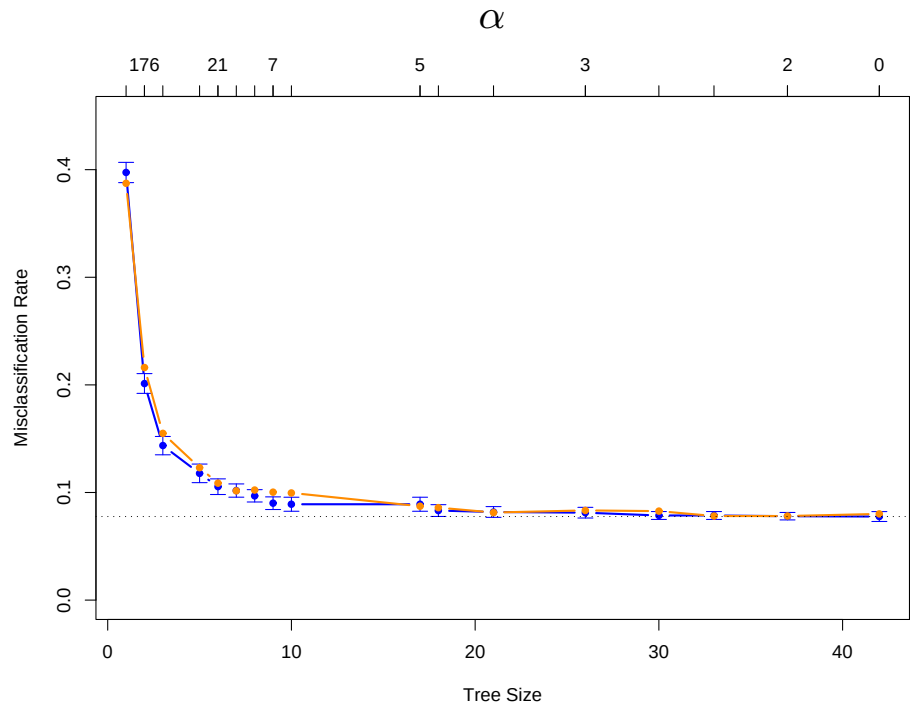

FIGURE 9.4. *Results for* `spam` *example. The blue curve is the* 10*-fold cross-validation estimate of misclassification rate as a function of tree size, with standard error bars. The minimum occurs at a tree size with about* 17 *terminal nodes (using the "one-standard--error" rule). The orange curve is the test error, which tracks the CV error quite closely. The cross-validation is indexed by values of* α*, shown above. The tree sizes shown below refer to* $|T_\alpha|$*, the size of the original tree indexed by* α*.*

Classification Tree

- $Y_i \in \{1, 2, ..., K\}$.
- Classify obs's in node m to the majority class:
  $\hat{p}_{mk} = \sum_{X_i \in R_m} I(Y_i = k)/n_m$,
  $k(m) = \arg\max_k \hat{p}_{mk}$.
- Impurity measure $Q_m(T)$:
  Used squarted error in regression trees.
  1. Misclassification error:
  $\frac{1}{n_m} \sum_{X_i \in R_m} I(Y_i \neq k(m)) = 1 - \hat{p}_{m,k(m)}$.
  2. Gini index: $\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$.
  3. Cross-entropy or deviance: $\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$.
- For $K = 2$, 1-3 reduce to $1 - \max(\hat{p}, 1 - \hat{p})$, $2\hat{p}(1 - \hat{p})$, $-\hat{p} \log \hat{p} - (1 - \hat{p}) \log(1 - \hat{p})$.
  Look similar; see Fig 9.3.
- Example: ex5.1.r

- Advantages:
  1. Easy to incorporate unequal losses of misclassifications: $\frac{1}{n_m}\sum_{X_i \in R_m} w_i I(Y_i \neq k(m))$ with $w_i = C_k$ if $Y_i = k$.
  2. Handling missing data: use a surrogate splitting var/value at each node (to best approximate the selected one).
- Extensions:
  1. May use non-binary splits;
  2. A linear combination of multiple var's as a splitting var.

  more flexible, but better?
- +: easy interpretation –decision trees!

  -: unstable due to greedy search and discontinuity; predicting performance not best.
- R packages `tree`, `rpart`; commercial CART.
- Other implementations: C4.5/C5.0;
  FIRM by Prof Hawkins (U of M): to detect interactions;
  by Prof Loh's group (UW-Madison): for count, survival, ... data; regression in each terminal node; ...

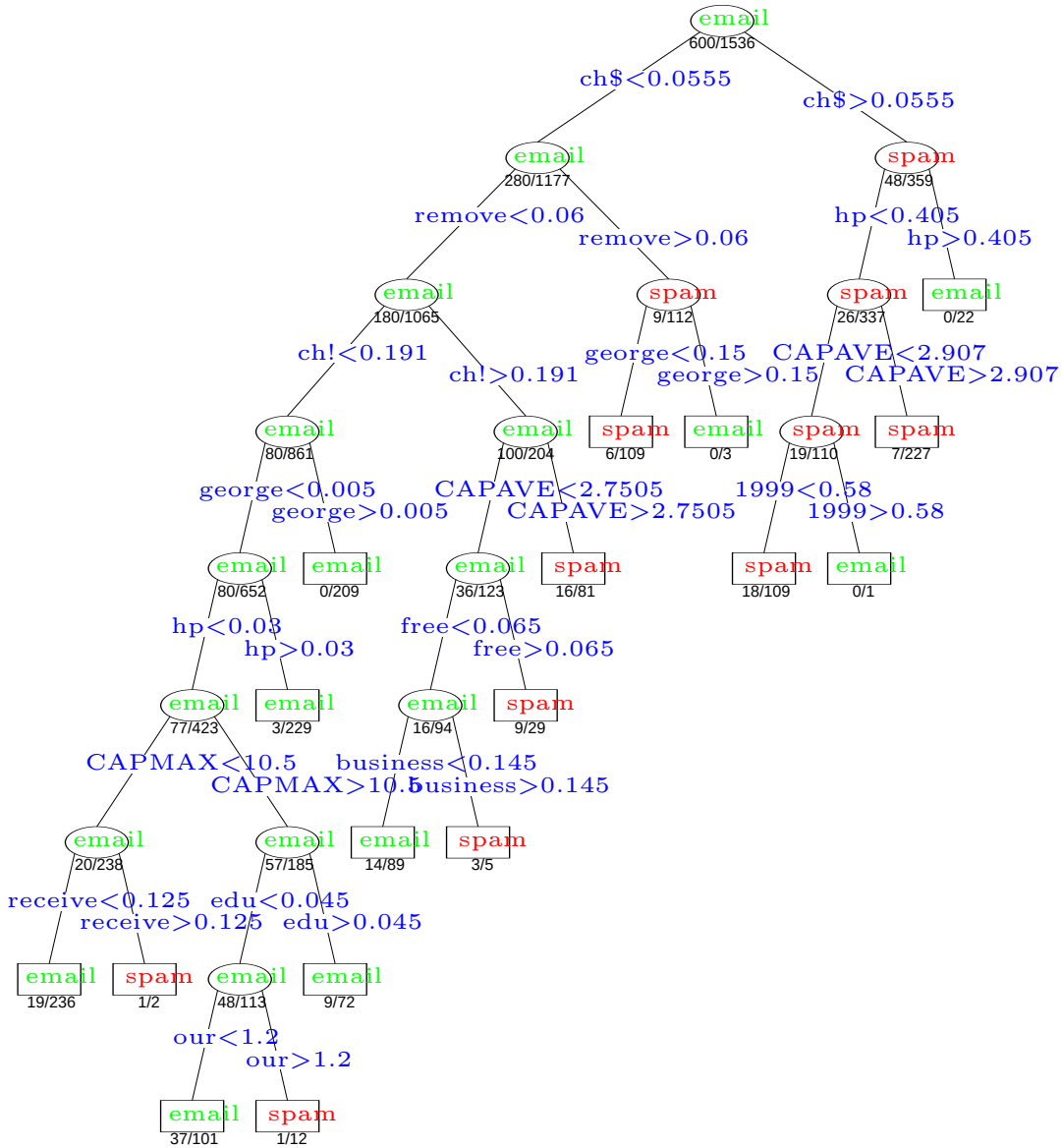

FIGURE 9.5. *The pruned tree for the* `spam` *example.*

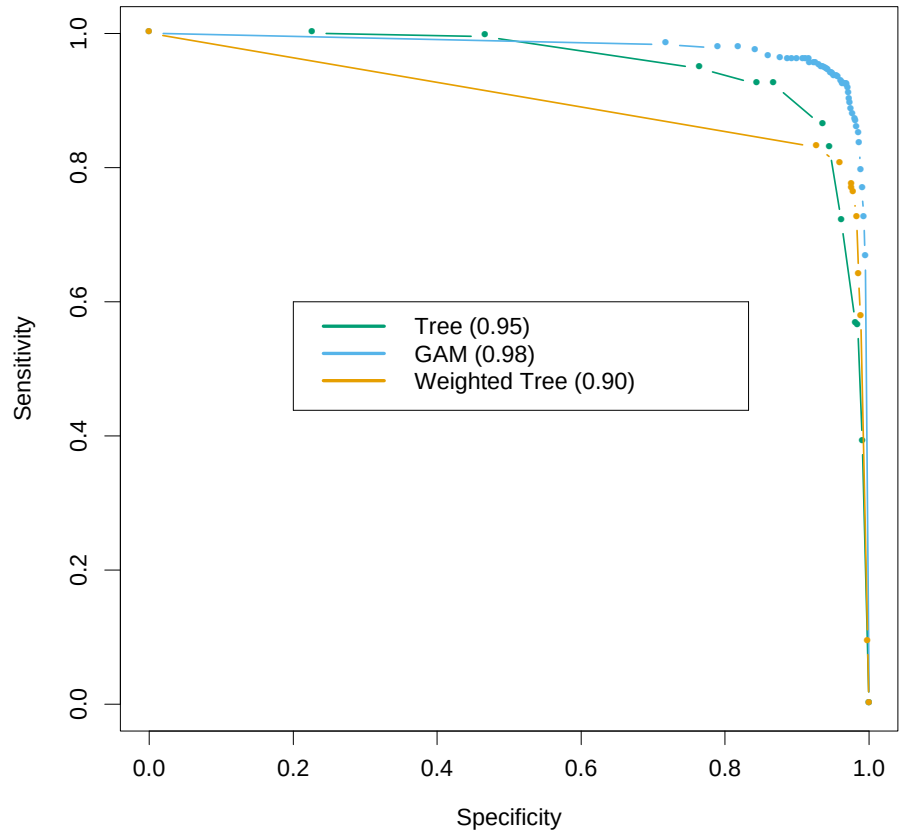

FIGURE 9.6. *ROC curves for the classification rules fit to the* `spam` *data. Curves that are closer to the north-east corner represent better classifiers. In this case the GAM classifier dominates the trees. The weighted tree achieves better sensitivity for higher specificity than the unweighted tree. The numbers in the legend represent the area under the curve.*

Application: personalized medicine

- Also called subgroup analysis (or Precision Medicine): to identify subgroups of patients that would be **most** benefit from a treatment.
- Statistical problem: detect (qualitative) trt-predictor interaction!
  quantitative interactions: differ in magnitudes but in teh same direction;
  qualitative interactions: differ in directions.
- Many approaches ... one of them is to use trees.
- Prof Loh's GUIDE:
  `http://www.stat.wisc.edu/~loh/guide.html`
- An example:
  `http://onlinelibrary.wiley.com/doi/10.1002/sim.6454/abstract`
- Another example:
  `https://www.ncbi.nlm.nih.gov/pubmed/24983709`

(8000) Causal inference

- Causal tree: inference.
  Athey & Imbens (2016). Recursive partitioning for heterogeneous causal effects. *PNAS*.