Classification Trees

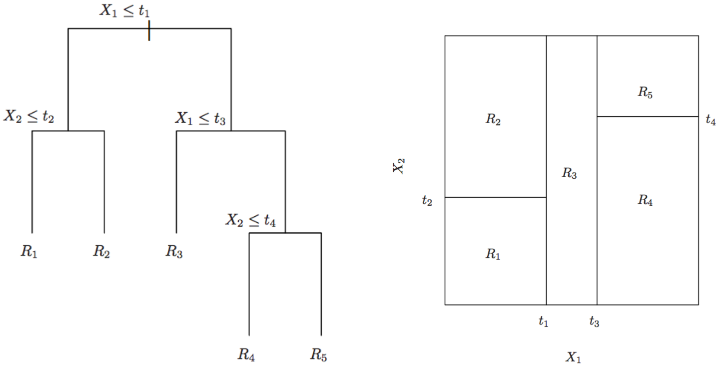
David Rosenberg

New York University

April 7, 2016

Binary Decision Tree on $\mathbf{R}^2$

- Consider a binary tree on $\{(X_1, X_2) \mid X_1, X_2 \in \mathbf{R}\}$



From *An Introduction to Statistical Learning, with applications in R* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

General Tree Building Procedure

- Choose a splitting variable and a split point
  - Splits input space $\mathcal{X}$ into R_1 and R_2
- We need to modify
  - criteria for splitting nodes
  - method for pruning tree

Classification Trees

- Consider classification case: $\mathcal{Y} = \{1, 2, \ldots, K\}$.
- We need to modify
  - criteria for splitting nodes
  - method for pruning tree

Root Node, Continuous Variables

- Let $x = (x_1, \dots, x_d) \in \mathbf{R}^d$.
- **Splitting variable** $j \in \{1, \dots, d\}$.
- **Split point** $s \in \mathbf{R}$.
- Partition based on j and s:

$$R_1(j,s) = \{x \mid x_j \leqslant s\}$$
$$R_2(j,s) = \{x \mid x_j > s\}$$

Classification Trees

- Let node m represent region R_m, with N_m observations
- Denote proportion of observations in R_m with class k by

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{\{i : x_i \in R_m\}} 1(y_i = k).$$

- **Predicted classification** for node m is

$$k(m) = \underset{k}{\arg\max}\, \hat{p}_{mk}.$$

- **Predicted class probability distribution** is $(\hat{p}_{m1}, \ldots, \hat{p}_{mK})$.

Misclassification Error

- Consider node m representing region R_m, with N_m observations
- Suppose we predict

$$k(m) = \arg\max_k \hat{p}_{mk}$$

as the class for all inputs in region R_m.
- What is the misclassification rate on the training data?
- It's just

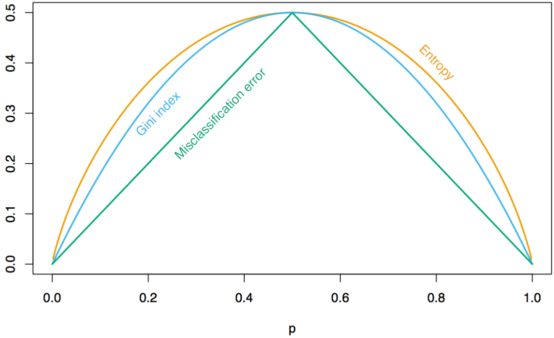
$$1 - \hat{p}_{mk(m)}.$$

Classification Trees: Node Impurity Measures

- Consider node m representing region R_m, with N_m observations
- How can we generalize from squared error to classification?
- We will introduce some different measures of **node impurity**.
  - We want **pure** leaf nodes (i.e. as close to a single class as possible)
- We'll find splitting variables and split point **minimizing node impurity**.

Two-Class Node Impurity Measures

- Consider binary classification
- Let p be the relative frequency of class 1.
- Here are three node impurity measures as a function of p

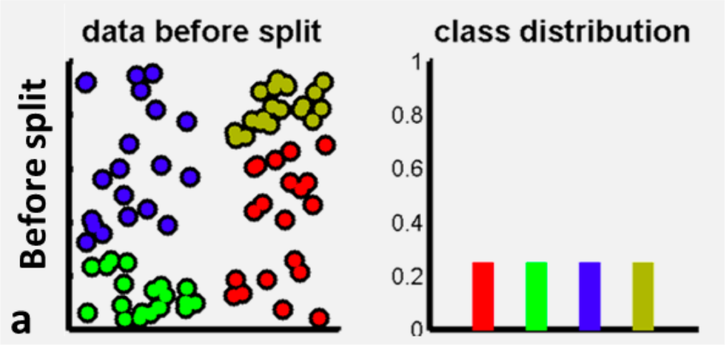


HTF Figure 9.3

Classification Trees: Node Impurity Measures

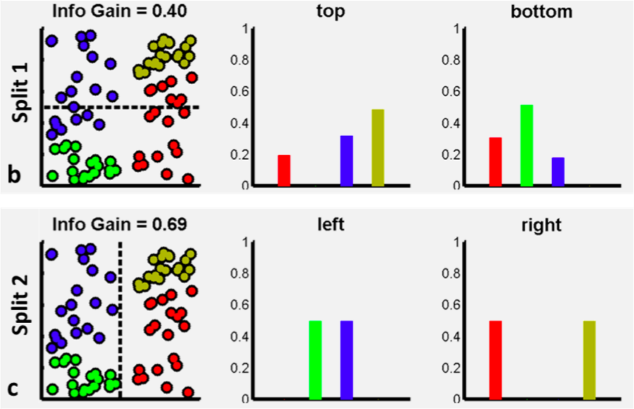
- Consider leaf node m representing region R_m, with N_m observations
- Three measures $Q_m(T)$ of **node impurity** for leaf node m:
  - Misclassification error:
$$1 - \hat{p}_{mk(m)}.$$
  - Gini index:
$$\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$
  - Entropy or deviance:
$$-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}.$$

Class Distributions: Pre-split



From Criminisi et al. MSR-TR-2011-114, 28 October 2011.

Class Distributions: Split Search



- (Maximizing information gain is equivalent to minimizing entropy)

From Criminisi et al. MSR-TR-2011-114, 28 October 2011.

Classification Trees: How exactly do we do this?

- Let R_L and R_R be regions corresponding to a potential node split.
- Suppose we have N_L points in R_L and N_R points in R_R.
- Let $Q(R_L)$ and $Q(R_R)$ be the node impurity measures.
- The we search for a split that minimizes

$$N_L Q(R_L) + N_R Q(R_R)$$

Classification Trees: Node Impurity Measures

- For building the tree, Gini and Entropy are more effective.
  - They push for more pure nodes, not just misclassification rate
- For pruning the tree, use misclassification error – closer to risk estimate.

Comments about Trees

- Trees make no use of **geometry**
  - No inner products or distances
  - called a "nonmetric" method
  - **Feature scale irrelevant**
- Predictions are not continuous
  - not so bad for classification
  - may not be desirable for regression