The main differences between algorithms for tree construction are the pruning strategy used and the exact rule for splitting nodes. Many algorithms only allow binary splits, that is to divide a node into two; a few allow multi-way splits (for example by flower colour). Note that these are just algorithms; there are only very simple models and no deep theorems in this field.

There are two types of optimality to be considered. One is optimality of the partition of $\mathscr{X}$, which can be judged by the error rate achieved. In principle we could seek an optimal partition amongst all prescribed partitions of $\mathscr{X}$, for example those representable by a set of decision rules splitting on a single feature. This is a computationally infeasible procedure for all but the smallest problems, but the stepwise construction of a partition by a decision tree can be seen as an approximation to finding the optimal partition.

The other sense of optimality is to represent a partition by a tree in the best possible way. The most obvious criterion is to use the minimal expected number of tests. Hyafil & Rivest (1976) showed this particular problem to be NP-complete; Payne & Meisel (1977) give an algorithm to construct optimal trees with respect to fairly general cost functions.

There are a number of partial surveys of the literature. Dietterich (1990) covers 'recent developments in practical learning algorithms'. Safavian & Landgrebe (1991) is wide-ranging but shallow. Quinlan (1986, 1990, 1993) surveys the machine-learning approaches within his own school.

7.1 Splitting rules

In this section and the next we consider the component pieces of currently favoured tree-construction algorithms. Some historical alternatives are mentioned in Section 7.4. Note that the number of possible trees is vast, so there is no question of an exhaustive search over trees.

Consider first splitting a leaf. There is a set of features from which to construct splitting attributes. For binary features we will clearly consider the binary split on that feature. For categorical features with $L > 2$ levels we can either consider an L-way split, or consider binary splits dividing the levels into two groups. (There will be $2^{L-1} - 1$ nonempty pairs of groups, so this generates many attributes for large L.) For ordered features the natural splits are binary of the form $x \leqslant x_c$; this applies both to continuous measurements and to ordered categories. Some systems also consider linear combinations of continuous features and Boolean combinations of logical ones. (See Section 7.5.)

Each leaf will have a set of *attributes* A on which it might be split. How should we consider the value of the split? There have been many suggestions from several different viewpoints. Consider first a population viewpoint. That is, there is a known probability distribution over $\mathscr{X} \times \mathscr{C}$ of examples which would reach that leaf. This gives a marginal probability distribution p_k over $\mathscr{C}$. Consider splitting on attribute A which has levels $a_1, \dots, a_m$. There is then a probability distribution p_{ik} over attributes and classes, and the child leaf corresponding to $A = a_i$ would have probability distribution $p(k \mid a_i) = p_{ik}/p_{i\cdot}$ over classes k.

$\mathscr{C}$ is the set of classes.

A $\cdot$ denotes summation over that index.

We can then ask if the child nodes are on average 'purer' than their parent. A measure of impurity should according to Breiman *et al.* (1984, p. 24) be zero if p_j is concentrated on one class, and maximal if p_j is uniform. Two commonly used measures of impurity are the *entropy*

$$i(p) = -\sum_j p_j \log p_j$$

(where $0 \log 0 = 0$) and the *Gini index*

$$i(p) = \sum_{i \neq j} p_i p_j = 1 - \sum_j p_j^2.$$

One interpretation of the Gini index is the expected error rate if the label is chosen randomly from the class distribution at the node. (It may be better to use this than the error rate from the Bayes rule at the node since it gives an element of 'look ahead'. Quite often no feasible split reduces the error rate, yet after two or three splits large reductions in error rate emerge; see the right-hand branch of Figure 7.2.)

The decrease in average impurity on splitting by attribute A is then

$$i(p_c) - \sum_{i=1}^{m} p_{i\cdot} \times i(p(c \mid a_i)).$$

A common approach is to choose the split that maximizes this. Since this will in general favour many-valued attributes, Breiman *et al.* and many others confine attention to binary attributes. (See Section 7.4 for adjustments for multi-way splits.)

Breiman *et al.* preferred the Gini index. The entropy index has been used widely, for example by Sethi & Sarvarayudu (1982) and Quinlan (1983) in the engineering and machine learning literature respectively.

The premise of the following proposition holds for both the entropy and Gini measures of impurity. Part (ii) reduces the number of attributes which need consideration for two classes from $2^{L-1} - 1$ to $L - 1$, but it has no simple extension to three or more classes. (The result is due to Breiman *et al.*, but the very much shorter proof is original.)

Proposition 7.1 *Suppose $i(p)$ is strictly concave.*

(i) *The decrease in impurity is non-negative, and zero if and only if the the distributions are the same in all children.*

(ii) *Suppose there are two classes. For a categorical feature, order the levels in increasing $p(1 \mid x = x_i)$. Then a split of the form $\{x_1, \dots x_\ell\}, \{x_{\ell+1}, \dots, x_L\}$ maximizes the reduction in average impurity.*

Proof: (i) We have by Jensen's inequality

See the glossary.

$$\sum p_{i\cdot} i(p(c \mid a_i)) \leqslant i(\sum_i p_{i\cdot} p(c \mid a_i)) = i(p_c)$$

with equality if and only if $p(c \mid a_i) = p(c)$ for all i and c.

(ii) With just two classes we can regard $i(p)$ as a function of p_1 only; it remains strictly concave. Consider dividing into two groups by allocating to group 1 with probability a_i when $x = x_i$. Then

(a) the average impurity of the two groups is minimized by taking $a_i = 0$ or 1 by concavity, and

(b) the partial right derivative of the average impurity with respect to a_i (which exists by concavity) at $a_i = 0$ is of the form

$$p(X = x_i)[Ap(1 \mid x = x_i) + B]$$

for constants A and B, and so is positive (when the optimal solution is to allocate x_i to group 1) for all $i \leqslant \ell$ or all $i > \ell$ for some ℓ.

and both examples lead to the postulated form of split since which group is labelled 1 is arbitrary. □

Another way to look at this approach is to define the average impurity of the tree as

$$I(T) = \sum_{\text{leaves } t} q_t i(p(c \mid t))$$

where q_t is the probability an example reaches node t. The decrease in I on splitting the node is then q_t times the decrease in node impurity we considered before, so that strategy is equivalent to splitting the node to minimize the average tree impurity.