An Introduction to Recursive Partitioning Using the RPART Routines

Terry M. Therneau
Elizabeth J. Atkinson
Mayo Foundation

December 5, 2023

Contents

1 Introduction

This document is a modification of a technical report from the Mayo Clinic Division of Biostatistics [6], which was itself an expansion of an earlier Stanford report [5]. It is intended to give a short overview of the methods found in the `rpart` routines, which implement many of the ideas found in the CART (Classification and Regression Trees) book and programs of Breiman, Friedman, Olshen and Stone [1]. Because CART is the trademarked name of a particular software implementation of these ideas, and *tree* has been used for the S-Plus routines of Clark and Pregibon [2] a different acronym — Recursive PARTitioning or rpart — was chosen. It is somewhat humorous that this label "rpart" has now become more common than the original and more descriptive "cart", a testament to the influence of freely available software.

The `rpart` programs build classification or regression models of a very general structure using a two stage procedure; the resulting models can be represented as binary trees. An example is some preliminary data gathered at Stanford on revival of cardiac arrest patients by paramedics. The goal is to predict which patients can be successfully revived in the field

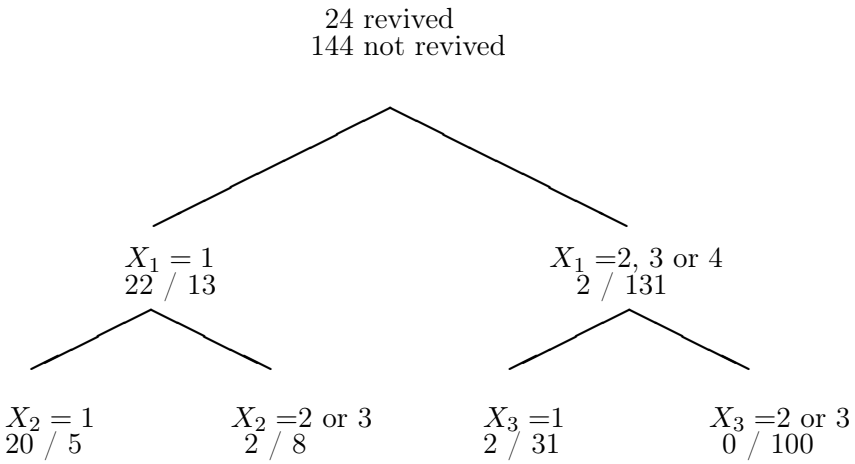


Figure 1: Revival data

on the basis of fourteen variables known at or near the time of paramedic arrival, e.g., sex, age, time from attack to first care, etc. Since some patients who are not revived on site are later successfully resuscitated at the hospital, early identification of these "recalcitrant" cases is of considerable clinical interest.

The resultant model separated the patients into four groups as shown in figure 1, where

$X1$ = initial heart rhythm
1= VF/VT 2=EMD 3=Asystole 4=Other

X_2 = initial response to defibrillation
1=Improved 2=No change 3=Worse

X_3 = initial response to drugs
1=Improved 2=No change 3=Worse

The other 11 variables did not appear in the final model. This procedure seems to work especially well for variables such as X_1 , where there is a definite ordering, but spacings are not necessarily equal.

The tree is built by the following process: first the single variable is found which best splits the data into two groups ('best' will be defined later). The data is separated, and then this process is applied *separately* to each sub-group, and so on recursively until the subgroups either reach a minimum size (5 for this data) or until no improvement can be made.

The resultant model is, with a certainty, too complex, and the question arises as it does with all stepwise procedures of when to stop. The second stage of the procedure consists of using cross-validation to trim back the full tree. In the medical example above the full tree

had ten terminal regions. A cross validated estimate of risk was computed for a nested set of sub trees; this final model was that sub tree with the lowest estimate of risk.

2 Notation

The partitioning method can be applied to many different kinds of data. We will start by looking at the classification problem, which is one of the more instructive cases (but also has the most complex equations). The sample population consists of n observations from C classes. A given model will break these observations into k terminal groups; to each of these groups is assigned a predicted class. In an actual application, most parameters will be estimated from the data, such estimates are given by $\approx$ formulae.

π_i	$i = 1, 2, ..., C$ prior probabilities of each class
$L(i, j)$	$i = 1, 2, ..., C$ Loss matrix for incorrectly classifying an i as a j. $L(i, i) \equiv 0$
A	some node of the tree Note that A represents both a set of individuals in the sample data, and, via the tree that produced it, a classification rule for future data.
$\tau(x)$	true class of an observation x, where x is the vector of predictor variables
$\tau(A)$	the class assigned to A, if A were to be taken as a final node
n_i, n_A	number of observations in the sample that are class i, number of obs in node A
$P(A)$	probability of A (for future observations) $= \sum_{i=1}^{C} \pi_i \, P\{x \in A \mid \tau(x) = i\}$ $\approx \sum_{i=1}^{C} \pi_i n_{iA}/n_i$
$p(i\|A)$	$P\{\tau(x) = i \mid x \in A\}$ (for future observations) $= \pi_i P\{x \in A \mid \tau(x) = i\}/P\{x \in A\}$ $\approx \pi_i(n_{iA}/n_i)/\sum \pi_i(n_{iA}/n_i\ )$

$R(A)$ risk of A
$= \sum_{i=1}^{C} p(i|A)L(i, \tau(A))$
where $\tau(A)$ is chosen to minimize this risk

$R(T)$ risk of a model (or tree) T
$= \sum_{j=1}^{k} P(A_j)R(A_j)$
where A_j are the terminal nodes of the tree

If $L(i, j) = 1$ for all $i \neq j$, and we set the prior probabilities π equal to the observed class frequencies in the sample then $p(i|A) = n_{iA}/n_A$ and $R(T)$ is the proportion misclassified.

3 Building the tree

3.1 Splitting criteria

If we split a node A into two sons A_L and A_R (left and right sons), we will have

$$P(A_L)r(A_L) + P(A_R)r(A_R) \leq P(A)r(A)$$

(this is proven in [1]). Using this, one obvious way to build a tree is to choose that split which maximizes Δr, the decrease in risk. There are defects with this, however, as the following example shows:

> Suppose losses are equal and that the data is 80% class 1's, and that some trial split results in A_L being 54% class 1's and A_R being 100% class 1's. Since the minimum risk prediction for both the left and right son is $\tau(A_L) = \tau(A_R) = 1$, this split will have $\Delta r = 0$, yet scientifically this is a very informative division of the sample. In real data with such a majority, the first few splits very often can do no better than this.

A more serious defect is that the risk reduction is essentially linear. If there were two competing splits, one separating the data into groups of 85% and 50% purity respectively, and the other into 70%-70%, we would usually prefer the former, if for no other reason than because it better sets things up for the next splits.

One way around both of these problems is to use look-ahead rules; but these are computationally very expensive. Instead `rpart` uses one of several measures of impurity, or diversity, of a node. Let f be some impurity function and define the impurity of a node A as

$$I(A) = \sum_{i=1}^{C} f(p_{iA})$$

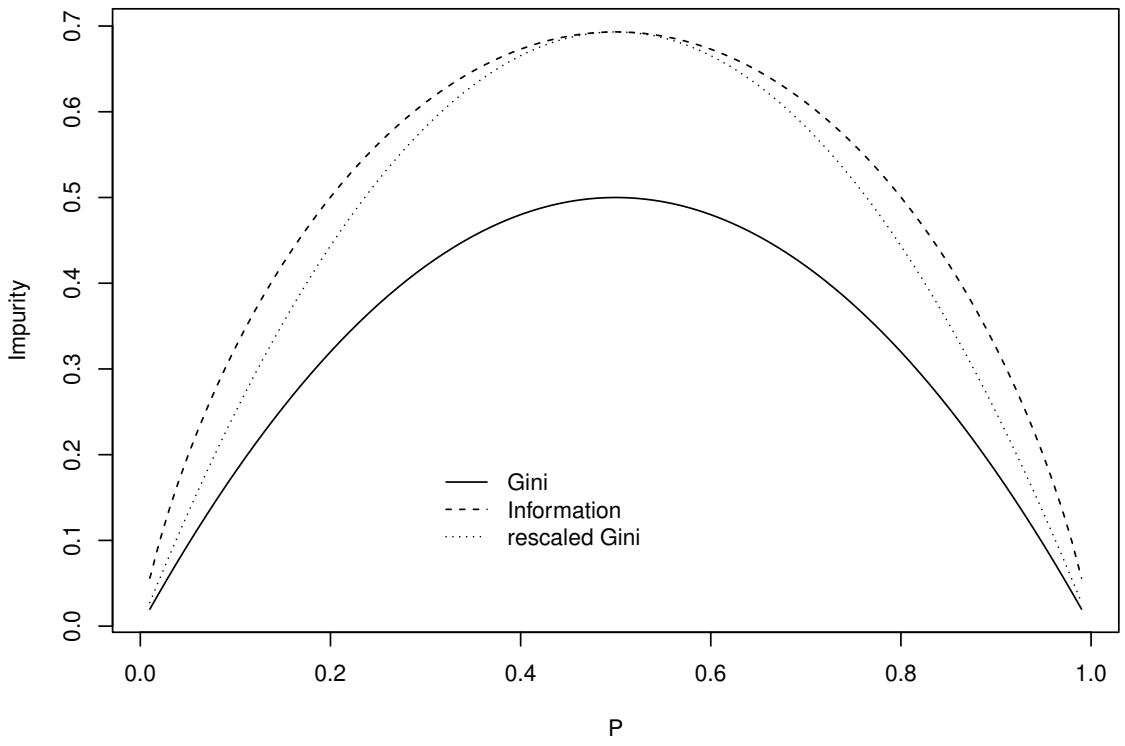


Figure 2: Comparison of Gini and Information impurity for two groups.

where p_{iA} is the proportion of those in A that belong to class i for future samples. Since we would like $I(A) = 0$ when A is pure, f must be concave with $f(0) = f(1) = 0$.

Two candidates for f are the information index $f(p) = -p\log(p)$ and the Gini index $f(p) = p(1-p)$. We then use that split with maximal impurity reduction

$$\Delta I = p(A)I(A) - p(A_L)I(A_L) - p(A_R)I(A_R)$$

The two impurity functions are plotted in figure (2), along with a rescaled version of the Gini measure. For the two class problem the measures differ only slightly, and will nearly always choose the same split point.

Another convex criteria not quite of the above class is twoing for which

$$I(A) = \min_{C_1 C_2}[f(p_{C_1}) + f(p_{C_2})]$$

where C_1,C_2 is some partition of the C classes into two disjoint sets. If $C = 2$ twoing is equivalent to the usual impurity index for f. Surprisingly, twoing can be calculated almost as efficiently as the usual impurity index. One potential advantage of twoing is that the

output may give the user additional insight concerning the structure of the data. It can be viewed as the partition of C into two superclasses which are in some sense the most dissimilar for those observations in A. For certain problems there may be a natural ordering of the response categories (e.g. level of education), in which case ordered twoing can be naturally defined, by restricting C_1 to be an interval $[1, 2, \ldots, k]$ of classes. Twoing is not part of `rpart`.

3.2 Incorporating losses

One salutatory aspect of the risk reduction criteria not found in the impurity measures is inclusion of the loss function. Two different ways of extending the impurity criteria to also include losses are implemented in CART, the generalized Gini index and altered priors. The `rpart` software implements only the altered priors method.

3.2.1 Generalized Gini index

The Gini index has the following interesting interpretation. Suppose an object is selected at random from one of C classes according to the probabilities $(p_1, p_2, ..., p_C)$ and is randomly assigned to a class using the same distribution. The probability of misclassification is

$$\sum_i \sum_{j \neq i} p_i p_j = \sum_i \sum_j p_i p_j - \sum_i p_i^2 = \sum_i 1 - p_i^2 = \text{Gini index for } p$$

Let $L(i, j)$ be the loss of assigning class j to an object which actually belongs to class i. The expected cost of misclassification is $\sum \sum L(i, j) p_i p_j$. This suggests defining a *generalized Gini* index of impurity by

$$G(p) = \sum_i \sum_j L(i, j) p_i p_j$$

The corresponding splitting criterion appears to be promising for applications involving variable misclassification costs. But there are several reasonable objections to it. First, $G(p)$ is not necessarily a concave function of p, which was the motivating factor behind impurity measures. More seriously, G symmetrizes the loss matrix before using it. To see this note that

$$G(p) = (1/2) \sum \sum [L(i, j) + L(j, i)] \, p_i p_j$$

In particular, for two-class problems, G in effect ignores the loss matrix.

3.2.2 Altered priors

Remember the definition of $R(A)$

$$\begin{aligned} R(A) &\equiv \sum_{i=1}^{C} p_{iA} L(i, \tau(A)) \\ &= \sum_{i=1}^{C} \pi_i L(i, \tau(A))(n_{iA}/n_i)(n/n_A) \end{aligned}$$

Assume there exists $\tilde{\pi}$ and $\tilde{L}$ be such that

$$\tilde{\pi}_i \tilde{L}(i, j) = \pi_i L(i, j) \qquad \forall i, j \in C$$

Then $R(A)$ is unchanged under the new losses and priors. If $\tilde{L}$ is proportional to the zero-one loss matrix then the priors $\tilde{\pi}$ should be used in the splitting criteria. This is possible only if L is of the form

$$L(i, j) = \begin{cases} L_i & i \neq j \\ 0 & i = j \end{cases}$$

in which case

$$\tilde{\pi}_i = \frac{\pi_i L_i}{\sum_j \pi_j L_j}$$

This is always possible when $C = 2$, and hence altered priors are exact for the two class problem. For arbitrary loss matrix of dimension $C > 2$, `rpart` uses the above formula with $L_i = \sum_j L(i, j)$.

A second justification for altered priors is this. An impurity index $I(A) = \sum f(p_i)$ has its maximum at $p_1 = p_2 = \ldots = p_C = 1/C$. If a problem had, for instance, a misclassification loss for class 1 which was twice the loss for a class 2 or 3 observation, one would wish I(A) to have its maximum at p_1 =1/5, $p_2 = p_3$ =2/5, since this is the worst possible set of proportions on which to decide a node's class. The altered priors technique does exactly this, by shifting the p_i.

Two final notes

- When altered priors are used, they affect only the choice of split. The ordinary losses and priors are used to compute the risk of the node. The altered priors simply help the impurity rule choose splits that are likely to be "good" in terms of the risk.
- The argument for altered priors is valid for both the Gini and information splitting rules.

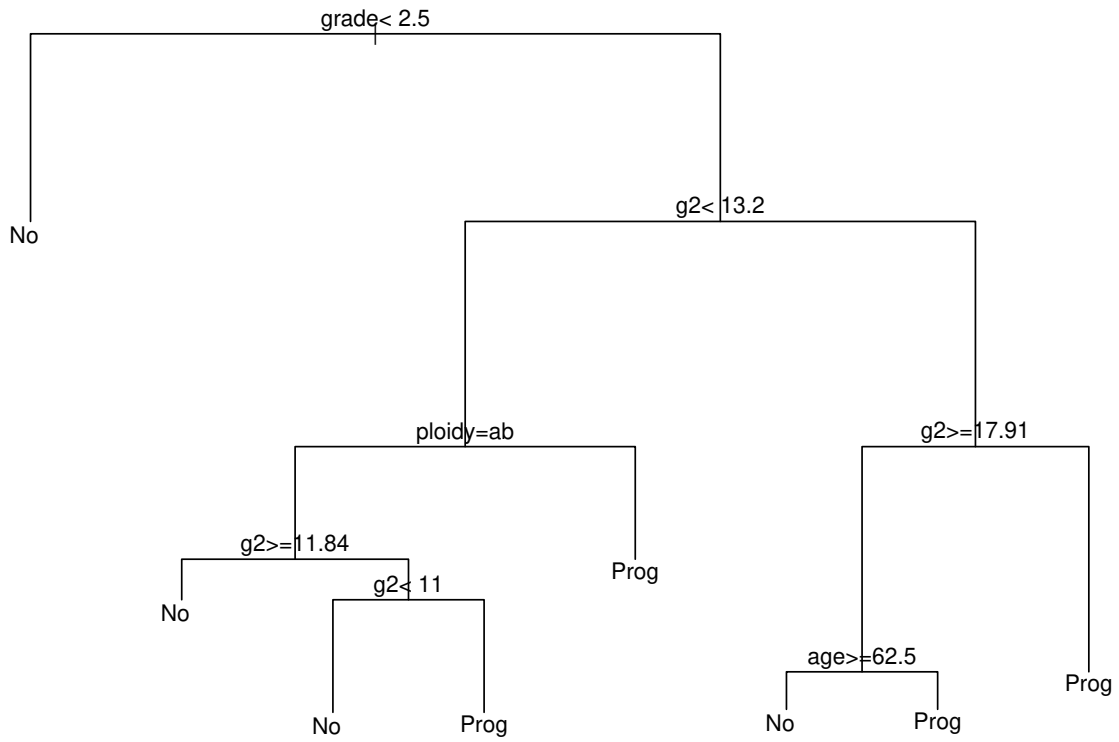


Figure 3: Classification tree for the Stage C data

3.3 Example: Stage C prostate cancer (`class` method)

This first example is based on a data set of 146 stage C prostate cancer patients [4]. The main clinical endpoint of interest is whether the disease recurs after initial surgical removal of the prostate, and the time interval to that progression (if any). The endpoint in this example is `status`, which takes on the value 1 if the disease has progressed and 0 if not. Later we'll analyze the data using the `exp` method, which will take into account time to progression. A short description of each of the variables is listed below. The main predictor variable of interest in this study was DNA ploidy, as determined by flow cytometry. For diploid and tetraploid tumors, the flow cytometry method was also able to estimate the percent of tumor cells in a G_2 (growth) stage of their cell cycle; G_2% is systematically missing for most aneuploid tumors.

The variables in the data set are

pgtime	time to progression, or last follow-up free of progression
pgstat	status at last follow-up (1=progressed, 0=censored)
age	age at diagnosis
eet	early endocrine therapy (1=no, 0=yes)
ploidy	diploid/tetraploid/aneuploid DNA pattern
g2	% of cells in G_2 phase
grade	tumor grade (1-4)
gleason	Gleason grade (3-10)

The model is fit by using the `rpart` function. The first argument of the function is a model formula, with the $\sim$ symbol standing for "is modeled as". The `print` function gives an abbreviated output, as for other R models. The `plot` and `text` command plot the tree and then label the plot, the result is shown in figure 3.

```
> progstat <- factor(stagec$pgstat, levels = 0:1, labels = c("No", "Prog"))
> cfit  <- rpart(progstat ~  age + eet + g2 + grade + gleason + ploidy,
                 data = stagec, method = 'class')
> print(cfit)
n= 146

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 146 54 No (0.6301370 0.3698630)
   2) grade< 2.5 61  9 No (0.8524590 0.1475410) *
   3) grade>=2.5 85 40 Prog (0.4705882 0.5294118)
     6) g2< 13.2 40 17 No (0.5750000 0.4250000)
      12) ploidy=diploid,tetraploid 31 11 No (0.6451613 0.3548387)
```

```
        24) g2>=11.845 7  1 No (0.8571429 0.1428571) *
        25) g2< 11.845 24 10 No (0.5833333 0.4166667)
          50) g2< 11.005 17  5 No (0.7058824 0.2941176) *
          51) g2>=11.005 7  2 Prog (0.2857143 0.7142857) *
      13) ploidy=aneuploid 9  3 Prog (0.3333333 0.6666667) *
     7) g2>=13.2 45 17 Prog (0.3777778 0.6222222)
      14) g2>=17.91 22  8 No (0.6363636 0.3636364)
        28) age>=62.5 15  4 No (0.7333333 0.2666667) *
        29) age< 62.5 7  3 Prog (0.4285714 0.5714286) *
      15) g2< 17.91 23  3 Prog (0.1304348 0.8695652) *
> par(mar = rep(0.1, 4))
> plot(cfit)
> text(cfit)
```

- The creation of a labeled factor variable as the response improves the labeling of the printout.
- We have explicitly directed the routine to treat `progstat` as a categorical variable by asking for `method='class'`. (Since `progstat` is a factor this would have been the default choice). Since no optional classification parameters are specified the routine will use the Gini rule for splitting, prior probabilities that are proportional to the observed data frequencies, and 0/1 losses.
- The child nodes of node x are always $2x$ and $2x+1$, to help in navigating the printout (compare the printout to figure 3).
- Other items in the list are the definition of the split used to create a node, the number of subjects at the node, the loss or error at the node (for this example, with proportional priors and unit losses this will be the number misclassified), and the predicted class for the node.
- * indicates that the node is terminal.
- Grades 1 and 2 go to the left, grades 3 and 4 go to the right. The tree is arranged so that the branches with the largest "average class" go to the right.

3.4 Variable importance

The long form of the printout for the stage C data, obtained with the summary function, contains further information on the surrogates. The `cp` option of the summary function instructs it to prune the printout, but it does not prune the tree. For each node up to 5

surrogate splits (default) will be printed, but only those whose utility is greater than the baseline "go with the majority" surrogate. The first surrogate for the first split is based on the following table:

```
> temp <- with(stagec, table(cut(grade, c(0, 2.5, 4)),
                             cut(gleason, c(2, 5.5, 10)),
                             exclude = NULL))
> temp
          (2,5.5] (5.5,10] <NA>
  (0,2.5]      42       16    3
  (2.5,4]       1       84    0
```

The surrogate sends 126 of the 146 observations the correct direction for an agreement of 0.863. The majority rule gets 85 correct, and the adjusted agreement is (126 - 85)/ (146 - 85).

A variable may appear in the tree many times, either as a primary or a surrogate variable. An overall measure of variable importance is the sum of the goodness of split measures for each split for which it was the primary variable, plus goodness * (adjusted agreement) for all splits in which it was a surrogate. In the printout these are scaled to sum to 100 and the rounded values are shown, omitting any variable whose proportion is less than 1%. Imagine two variables which were essentially duplicates of each other; if we did not count surrogates they would split the importance with neither showing up as strongly as it should.

4 Pruning the tree

4.1 Definitions

We have built a complete tree, possibly quite large and/or complex, and must now decide how much of that model to retain. In stepwise regression, for instance, this issue is addressed sequentially and the fit is stopped when the F test fails to achieve some level α.

Let T_1, T_2,....,T_k be the terminal nodes of a tree T. Define

$|T|$ = number of terminal nodes
risk of $T = R(T) = \sum_{i=1}^{k} P(T_i)R(T_i)$

In comparison to regression, $|T|$ is analogous to the degrees of freedom and $R(T)$ to the residual sum of squares.

Now let α be some number between 0 and ∞ which measures the 'cost' of adding another variable to the model; α will be called a complexity parameter. Let $R(T_0)$ be the risk for the zero split tree. Define

$$R_\alpha(T) = R(T) + \alpha|T|$$

to be the cost for the tree, and define T_α to be that sub tree of the full model which has minimal cost. Obviously T_0 = the full model and T_∞ = the model with no splits at all. The following results are shown in [1].

1. If T_1 and T_2 are sub trees of T with $R_\alpha(T_1) = R_\alpha(T_2)$, then either T_1 is a sub tree of T_2 or T_2 is a sub tree of T_1; hence either $|T_1| < |T_2|$ or $|T_2| < |T_1|$.

2. If $\alpha > \beta$ then either $T_\alpha = T_\beta$ or T_α is a strict sub tree of T_β.

3. Given some set of numbers $\alpha_1, \alpha_2, \ldots, \alpha_m$; both $T_{\alpha_1}, \ldots, T_{\alpha_m}$ and $R(T_{\alpha_1}), \ldots, R(T_{\alpha_m})$ can be computed efficiently.

Using the first result, we can uniquely define T_α as the smallest tree T for which $R_\alpha(T)$ is minimized.

Since any sequence of nested trees based on T has at most $|T|$ members, result 2 implies that all possible values of α can be grouped into m intervals, $m \leq |T|$

$$\begin{aligned} I_1 &= [0, \alpha_1] \\ I_2 &= (\alpha_1, \alpha_2] \\ &\vdots \\ I_m &= (\alpha_{m-1}, \infty] \end{aligned}$$

where all $\alpha \in I_i$ share the same minimizing sub tree.

4.2 Cross-validation

Cross-validation is used to choose a best value for α by the following steps:

1. Fit the full model on the data set
   compute $I_1, I_2, ..., I_m$
   set $\beta_1 = 0$
   $\beta_2 = \sqrt{\alpha_1 \alpha_2}$
   $\beta_3 = \sqrt{\alpha_2 \alpha_3}$
   $\vdots$
   $\beta_{m-1} = \sqrt{\alpha_{m-2} \alpha_{m-1}}$
   $\beta_m = \infty$
   each β_i is a 'typical value' for its I_i

2. Divide the data set into s groups $G_1, G_2, ..., G_s$ each of size s/n, and for each group separately:

- fit a full model on the data set 'everyone except G_i' and determine $T_{\beta_1}, T_{\beta_2}, ..., T_{\beta_m}$ for this reduced data set,
- compute the predicted class for each observation in G_i, under each of the models T_{β_j} for $1 \leq j \leq m$,
- from this compute the risk for each subject.

3. Sum over the G_i to get an estimate of risk for each β_j. For that β (complexity parameter) with smallest risk compute T_β for the full data set, this is chosen as the best trimmed tree.

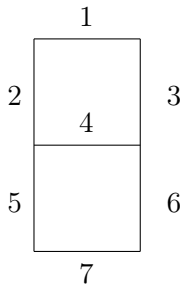
In actual practice, we may use instead the 1-SE rule. A plot of β versus risk often has an initial sharp drop followed by a relatively flat plateau and then a slow rise. The choice of β among those models on the plateau can be essentially random. To avoid this, both an estimate of the risk and its standard error are computed during the cross-validation. Any risk within one standard error of the achieved minimum is marked as being equivalent to the minimum (i.e. considered to be part of the flat plateau). Then the simplest model, among all those "tied" on the plateau, is chosen.

In the usual definition of cross-validation we would have taken $s = n$ above, i.e., each of the G_i would contain exactly one observation, but for moderate n this is computationally prohibitive. A value of $s = 10$ has been found to be sufficient, but users can vary this if they wish.

In Monte-Carlo trials, this method of pruning has proven very reliable for screening out 'pure noise' variables in the data set.

4.3 Example: The Stochastic Digit Recognition Problem

This example is found in section 2.6 of [1], and used as a running example throughout much of their book. Consider the segments of an unreliable digital readout



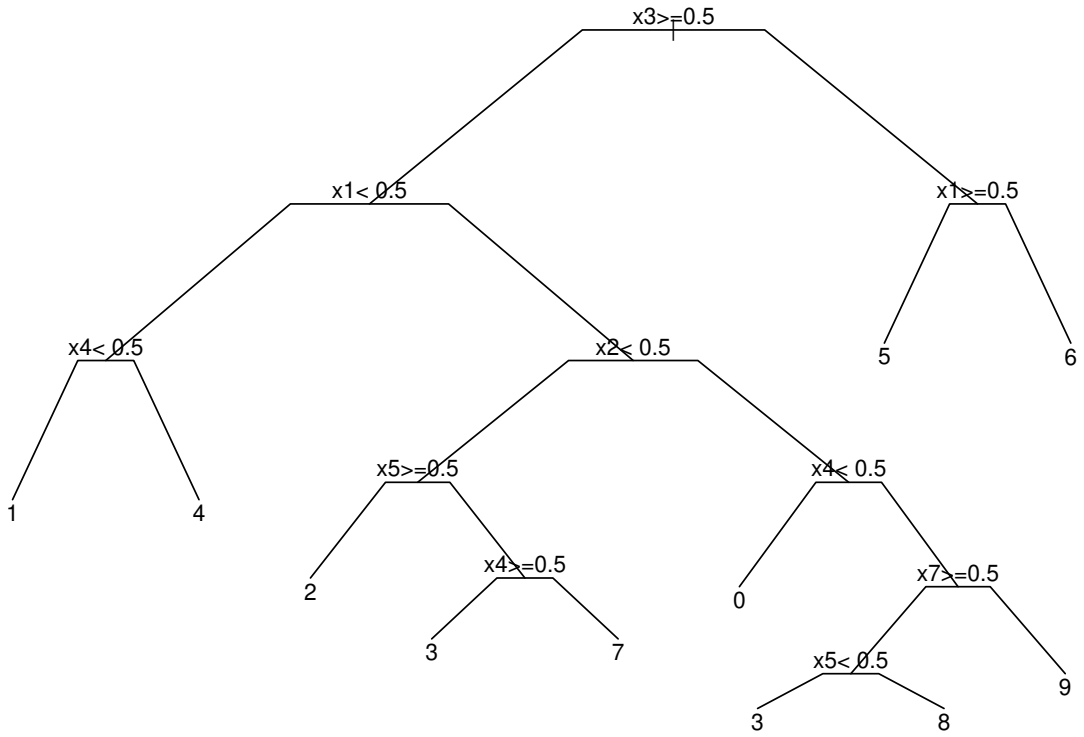


Figure 4: Optimally pruned tree for the stochastic digit recognition data

where each light is correct with probability 0.9, e.g., if the true digit is a 2, the lights 1, 3, 4, 5, and 7 are on with probability 0.9 and lights 2 and 6 are on with probability 0.1. Construct test data where $Y \in \{0, 1, ..., 9\}$, each with proportion $1/10$ and the X_i, $i = 1, \ldots, 7$ are i.i.d. Bernoulli variables with parameter depending on Y. $X_8 - X_{24}$ are generated as i.i.d Bernoulli $P\{X_i = 1\} = .5$, and are independent of Y. They correspond to embedding the readout in a larger rectangle of random lights.

A sample of size 200 was generated accordingly and the procedure applied using the Gini index (see 3.2.1) to build the tree. The R code to compute the simulated data and the fit are shown below.

```
> set.seed(1953)  # An auspicious year
> n <- 200
> y <- rep(0:9, length = 200)
> temp <- c(1,1,1,0,1,1,1,
            0,0,1,0,0,1,0,
            1,0,1,1,1,0,1,
```

```
            1,0,1,1,0,1,1,
            0,1,1,1,0,1,0,
            1,1,0,1,0,1,1,
            0,1,0,1,1,1,1,
            1,0,1,0,0,1,0,
            1,1,1,1,1,1,1,
            1,1,1,1,0,1,0)
> lights <- matrix(temp, 10, 7, byrow = TRUE)    # The true light pattern 0-9
> temp1 <- matrix(rbinom(n*7, 1, 0.9), n, 7) # Noisy lights
> temp1 <- ifelse(lights[y+1, ] == 1, temp1, 1-temp1)
> temp2 <- matrix(rbinom(n*17, 1, 0.5), n, 17) # Random lights
> x <- cbind(temp1, temp2)
> dfit <- rpart(y ~ x, method='class',
                control = rpart.control(xval = 10, minbucket = 2, cp = 0))
> printcp(dfit)
Classification tree:
rpart(formula = y ~ x, method = "class", control = rpart.control(xval = 10,
    minbucket = 2, cp = 0))

Variables actually used in tree construction:
 [1] x1  x10 x13 x14 x17 x19 x2  x21 x3  x4  x5  x6  x7  x8
[15] x9

Root node error: 180/200 = 0.9

n= 200

          CP nsplit rel error  xerror      xstd
1  0.1111111      0   1.00000 1.10556 0.0055416
2  0.1000000      1   0.88889 1.04444 0.0186587
3  0.0888889      2   0.78889 0.95000 0.0276637
4  0.0777778      4   0.61111 0.85000 0.0333125
5  0.0666667      5   0.53333 0.68889 0.0381356
6  0.0611111      6   0.46667 0.61667 0.0390453
7  0.0555556      7   0.40556 0.58889 0.0392129
8  0.0388889      8   0.35000 0.50000 0.0390868
9  0.0222222      9   0.31111 0.38889 0.0374743
10 0.0166667     10   0.28889 0.39444 0.0375956
11 0.0083333     11   0.27222 0.38333 0.0373485
```

```
12 0.0055556     17   0.22222 0.37222 0.0370831
13 0.0027778     25   0.17778 0.36111 0.0367990
14 0.0000000     27   0.17222 0.36111 0.0367990
> fit9 <- prune(dfit, cp = 0.02)
> par(mar = rep(0.1, 4))
> plot(fit9, branch = 0.3, compress = TRUE)
> text(fit9)
```

This table differs from that in section 3.5 of [1] in several ways, the last two of which are substantive.

- The actual values are different, of course, because of different random number generators in the two runs.
- The complexity table is printed from the smallest tree (no splits) to the largest one (28 splits). We find it easier to compare one tree to another when they start at the same place.
- The number of splits is listed, rather than the number of nodes. The number of nodes is always 1 + the number of splits.
- For easier reading, the error columns have been scaled so that the first node has an error of 1. Since in this example the model with no splits must make 180/200 misclassifications, multiply columns 3-5 by 180 to get a result in terms of absolute error. (Computations are done on the absolute error scale, and printed on relative scale).
- The complexity parameter column has been similarly scaled.

Looking at the table, we see that the best tree has 10 terminal nodes (9 splits), based on cross-validation. This sub tree is extracted with a call to `prune` and saved in `fit9`. The pruned tree is shown in figure 4. Two options have been used in the plot. The `compress` option tries to narrow the printout by vertically overlapping portions of the plot. The `branch` option controls the shape of the branches that connect a node to its children.

The largest tree, with 35 terminal nodes, correctly classifies 170/200 = 85% of the observations, but uses several of the random predictors in doing so and seriously over fits the data. If the number of observations per terminal node (minbucket) had been set to 1 instead of 2, then every observation would be classified correctly in the final model, many in terminal nodes of size 1.

5 Missing data

5.1 Choosing the split

Missing values are one of the curses of statistical models and analysis. Most procedures deal with them by refusing to deal with them – incomplete observations are tossed out. `Rpart` is somewhat more ambitious. Any observation with values for the dependent variable and at least one independent variable will participate in the modeling.

The quantity to be maximized is still

$$\Delta I = p(A)I(A) - p(A_L)I(A_L) - p(A_R)I(A_R)$$

The leading term is the same for all variables and splits irrespective of missing data, but the right two terms are somewhat modified. Firstly, the impurity indices $I(A_R)$ and $I(A_L)$ are calculated only over the observations which are not missing a particular predictor. Secondly, the two probabilities $p(A_L)$ and $p(A_R)$ are also calculated only over the relevant observations, but they are then adjusted so that they sum to $p(A)$. This entails some extra bookkeeping as the tree is built, but ensures that the terminal node probabilities sum to 1.

In the extreme case of a variable for which only 2 observations are non-missing, the impurity of the two sons will both be zero when splitting on that variable. Hence ΔI will be maximal, and this 'almost all missing' coordinate is guaranteed to be chosen as best; the method is certainly flawed in this extreme case. It is difficult to say whether this bias toward missing coordinates carries through to the non-extreme cases, however, since a more complete variable also affords for itself more possible values at which to split.

5.2 Surrogate variables

Once a splitting variable and a split point for it have been decided, what *is* to be done with observations missing that variable? One approach is to estimate the missing datum using the other independent variables; `rpart` uses a variation of this to define *surrogate* variables.

As an example, assume that the split (age $<$40, age $\geq$40) has been chosen. The surrogate variables are found by re-applying the partitioning algorithm (without recursion) to predict the two categories 'age $<$40' vs. 'age $\geq$40' using the other independent variables.

For each predictor an optimal split point and a misclassification error are computed. (Losses and priors do not enter in — none are defined for the age groups — so the risk is simply #misclassified / n.) Also evaluated is the blind rule 'go with the majority' which has misclassification error $\min(p, 1-p)$ where

$$p = (\#\text{ in A with age} < 40) \ / \ n_A.$$

The surrogates are ranked, and any variables which do no better than the blind rule are discarded from the list.

Assume that the majority of subjects have age ≤ 40 and that there is another variable x which is uncorrelated to age; however, the subject with the largest value of x is also over 40 years of age. Then the surrogate variable $x <$ max versus $x \geq$ max will have one less error that the blind rule, sending 1 subject to the right and $n - 1$ to the left. A continuous variable that is completely unrelated to age has probability $1 - p^2$ of generating such a trim-one-end surrogate by chance alone. For this reason the `rpart` routines impose one more constraint during the construction of the surrogates: a candidate split must send at least 2 observations to the left and at least 2 to the right.

Any observation which is missing the split variable is then classified using the first surrogate variable, or if missing that, the second surrogate is used, and etc. If an observation is missing all the surrogates the blind rule is used. Other strategies for these 'missing everything' observations can be convincingly argued, but there should be few or no observations of this type (we hope).

5.3 Example: Stage C prostate cancer (cont.)

Let us return to the stage C prostate cancer data of the earlier example. For a more detailed listing of the rpart object, we use the `summary` function. It includes the information from the CP table (not repeated below), plus information about each node. It is easy to print a sub tree based on a different cp value using the `cp` option. Any value between 0.0555 and 0.1049 would produce the same result as is listed below, that is, the tree with 3 splits. Because the printout is long, the `file` option of `summary.rpart` is often useful.

```
> printcp(cfit)
Classification tree:
rpart(formula = progstat ~ age + eet + g2 + grade + gleason +
    ploidy, data = stagec, method = "class")

Variables actually used in tree construction:
[1] age    g2     grade  ploidy

Root node error: 54/146 = 0.36986

n= 146

        CP nsplit rel error  xerror    xstd
1 0.104938      0   1.00000 1.00000 0.10802
2 0.055556      3   0.68519 1.07407 0.10949
3 0.027778      4   0.62963 0.96296 0.10715
```

```
4 0.018519      6   0.57407 0.98148 0.10760
5 0.010000      7   0.55556 1.00000 0.10802
> summary(cfit, cp = 0.06)
Call:
rpart(formula = progstat ~ age + eet + g2 + grade + gleason +
    ploidy, data = stagec, method = "class")
  n= 146

          CP nsplit rel error    xerror      xstd
1 0.10493827      0 1.0000000 1.0000000 0.1080241
2 0.05555556      3 0.6851852 1.0740741 0.1094927
3 0.02777778      4 0.6296296 0.9629630 0.1071508
4 0.01851852      6 0.5740741 0.9814815 0.1075992
5 0.01000000      7 0.5555556 1.0000000 0.1080241

Variable importance
     g2   grade gleason  ploidy     age     eet
     30      28      20      13       7       2

Node number 1: 146 observations,    complexity param=0.1049383
  predicted class=No    expected loss=0.369863  P(node) =1
    class counts:    92    54
   probabilities: 0.630 0.370
  left son=2 (61 obs) right son=3 (85 obs)
  Primary splits:
      grade   < 2.5    to the left,  improve=10.357590, (0 missing)
      gleason < 5.5    to the left,  improve= 8.399574, (3 missing)
      ploidy  splits as  LRR,        improve= 7.656533, (0 missing)
      g2      < 13.2   to the left,  improve= 7.186766, (7 missing)
      age     < 58.5   to the right, improve= 1.388128, (0 missing)
  Surrogate splits:
      gleason < 5.5    to the left,  agree=0.863, adj=0.672, (0 split)
      ploidy  splits as  LRR,        agree=0.644, adj=0.148, (0 split)
      g2      < 9.945  to the left,  agree=0.630, adj=0.115, (0 split)
      age     < 66.5   to the right, agree=0.589, adj=0.016, (0 split)

Node number 2: 61 observations
  predicted class=No    expected loss=0.147541  P(node) =0.4178082
    class counts:    52     9
```

```
   probabilities: 0.852 0.148

Node number 3: 85 observations,    complexity param=0.1049383
  predicted class=Prog  expected loss=0.4705882  P(node) =0.5821918
    class counts:    40    45
   probabilities: 0.471 0.529
  left son=6 (40 obs) right son=7 (45 obs)
  Primary splits:
      g2      < 13.2   to the left,  improve=2.1781360, (6 missing)
      ploidy  splits as  LRR,        improve=1.9834830, (0 missing)
      age     < 56.5   to the right, improve=1.6596080, (0 missing)
      gleason < 8.5    to the left,  improve=1.6386550, (0 missing)
      eet     < 1.5    to the right, improve=0.1086108, (1 missing)
  Surrogate splits:
      ploidy  splits as  LRL,        agree=0.962, adj=0.914, (6 split)
      age     < 68.5   to the right, agree=0.608, adj=0.114, (0 split)
      gleason < 6.5    to the left,  agree=0.582, adj=0.057, (0 split)

Node number 6: 40 observations
  predicted class=No    expected loss=0.425  P(node) =0.2739726
    class counts:    23    17
   probabilities: 0.575 0.425

Node number 7: 45 observations,    complexity param=0.1049383
  predicted class=Prog  expected loss=0.3777778  P(node) =0.3082192
    class counts:    17    28
   probabilities: 0.378 0.622
  left son=14 (22 obs) right son=15 (23 obs)
  Primary splits:
      g2      < 17.91  to the right, improve=5.2414830, (1 missing)
      age     < 62.5   to the right, improve=0.8640576, (0 missing)
      gleason < 7.5    to the left,  improve=0.2115900, (0 missing)
      eet     < 1.5    to the right, improve=0.1280042, (1 missing)
  Surrogate splits:
      age     < 61.5   to the right, agree=0.614, adj=0.190, (1 split)
      eet     < 1.5    to the right, agree=0.591, adj=0.143, (0 split)
      grade   < 3.5    to the right, agree=0.545, adj=0.048, (0 split)
      gleason < 6.5    to the right, agree=0.545, adj=0.048, (0 split)

Node number 14: 22 observations
```

```
  predicted class=No    expected loss=0.3636364  P(node) =0.1506849
    class counts:    14     8
   probabilities: 0.636 0.364

Node number 15: 23 observations
  predicted class=Prog  expected loss=0.1304348  P(node) =0.1575342
    class counts:     3    20
   probabilities: 0.130 0.870
```

- There are 54 progressions (class 1) and 94 non-progressions, so the first node has an expected loss of $54/146 \approx 0.37$. (The computation is this simple only for the default priors and losses).
- Grades 1 and 2 go to the left, grades 3 and 4 to the right. The tree is arranged so that the "more severe" nodes go to the right.
- The improvement is n times the change in impurity index. In this instance, the largest improvement is for the variable `grade`, with an improvement of 10.36. The next best choice is Gleason score, with an improvement of 8.4. The actual values of the improvement are not so important, but their relative size gives an indication of the comparative utility of the variables.
- Ploidy is a categorical variable, with values of diploid, tetraploid, and aneuploid, in that order. (To check the order, type `table(stagec$ploidy)`). All three possible splits were attempted: aneuploid+diploid vs. tetraploid, aneuploid+tetraploid vs. diploid, and aneuploid vs. diploid + tetraploid. The best split sends diploid to the right and the others to the left.
- The 2 by 2 table of diploid/non-diploid vs grade= 1-2/3-4 has 64% of the observations on the diagonal.

	1-2	3-4
diploid	38	29
tetraploid	22	46
aneuploid	1	10

- For node 3, the primary split variable is missing on 6 subjects. All 6 are split based on the first surrogate, ploidy. Diploid and aneuploid tumors are sent to the left, tetraploid to the right. As a surrogate, eet was no better than 45/85 (go with the majority), and

was not retained.

	g2 < 13.2	g2 > 13.2
Diploid/aneuploid	71	1
Tetraploid	3	64

6 Further options

6.1 Program options

The central fitting function is `rpart`, whose main arguments are

- `formula`: the model formula, as in `lm` and other R model fitting functions. The right hand side may contain both continuous and categorical (factor) terms. If the outcome y has more than two levels, then categorical predictors must be fit by exhaustive enumeration, which can take a very long time.
- `data, weights, subset`: as for other R models.
- `method`: the type of splitting rule to use. Options at this point are classification, anova, Poisson, and exponential.
- `parms`: a list of method specific optional parameters. For classification, the list can contain any of: the vector of prior probabilities (component `prior`), the loss matrix (component `loss`) or the splitting index (component `split`). The priors must be positive and sum to 1. The loss matrix must have zeros on the diagonal and positive off-diagonal elements. The splitting index can be `"gini"` or `"information"`.
- `na.action`: the action for missing values. The default action for rpart is `na.rpart`, this default is not overridden by the `options(na.action)` global option. The default action removes only those rows for which either the response y or *all* of the predictors are missing. This ability to retain partially missing observations is perhaps the single most useful feature of `rpart` models.
- `control`: a list of control parameters, usually the result of the `rpart.control` function. The list must contain
  - `minsplit`: The minimum number of observations in a node for which the routine will even try to compute a split. The default is 20. This parameter can save computation time, since smaller nodes are almost always pruned away by cross-validation.
  - `minbucket`: The minimum number of observations in a terminal node. This defaults to `minsplit/3`.

- `maxcompete`: It is often useful in the printout to see not only the variable that gave the best split at a node, but also the second, third, etc best. This parameter controls the number that will be printed. It has no effect on computational time, and a small effect on the amount of memory used. The default is 4.
- `xval`: The number of cross-validations to be done. Usually set to zero during exploratory phases of the analysis. A value of 10, for instance, increases the compute time to 11-fold over a value of 0.
- `maxsurrogate`: The maximum number of surrogate variables to retain at each node. (No surrogate that does worse than "go with the majority" is printed or used). Setting this to zero will cut the computation time in half, and set `usesurrogate` to zero. The default is 5. Surrogates give different information than competitor splits. The competitor list asks "which other splits would have as many correct classifications", surrogates ask "which other splits would classify the same subjects in the same way", which is a harsher criteria.
- `usesurrogate`: A value of usesurrogate=2, the default, splits subjects in the way described previously. This is similar to CART. If the value is 0, then a subject who is missing the primary split variable does not progress further down the tree. A value of 1 is intermediate: all surrogate variables except "go with the majority" are used to send a case further down the tree.
- `cp`: The threshold complexity parameter.

The complexity parameter `cp` is, like `minsplit`, an advisory parameter, but is considerably more useful. It is specified according to the formula

$$R_{cp}(T) \equiv R(T) + cp * |T| * R(T_1)$$

where T_1 is the tree with no splits, $|T|$ is the number of splits for a tree, and R is the risk. This scaled version is much more user friendly than the original CART formula (4.1) since it is unit less. A value of `cp = 1` will always result in a tree with no splits. For regression models (see next section) the scaled cp has a very direct interpretation: if any split does not increase the overall R^2 of the model by at least cp (where R^2 is the usual linear-models definition) then that split is decreed to be, a priori, not worth pursuing. The program does not split said branch any further, and saves considerable computational effort. The default value of .01 has been reasonably successful at 'pre-pruning' trees so that the cross-validation step need only remove 1 or 2 layers, but it sometimes over prunes, particularly for large data sets.

6.2 Example: Consumer Report Auto Data

A second example using the `class` method demonstrates the outcome for a response with multiple (> 2) categories. We also explore the difference between Gini and information

splitting rules. The dataset cu.summary contains a collection of variables from the April, 1990 Consumer Reports summary on 117 cars. For our purposes, car Reliability will be treated as the response. The variables are:

Reliability	an ordered factor (contains NAs): Much worse < worse < average < better < Much Better
Price	numeric: list price in dollars, with standard equipment
Country	factor: country where car manufactured
Mileage	numeric: gas mileage in miles/gallon, contains NAs
Type	factor: Small, Sporty, Compact, Medium, Large, Van

In the analysis we are treating reliability as an unordered outcome. Nodes potentially can be classified as Much worse, worse, average, better, or Much better, though there are none that are labeled as just "better". The 32 cars with missing response (listed as NA) were not used in the analysis. Two fits are made, one using the Gini index and the other the information index.

```
> fit1 <- rpart(Reliability ~ Price + Country + Mileage + Type,
                  data = cu.summary, parms = list(split = 'gini'))
> fit2 <- rpart(Reliability ~ Price + Country + Mileage + Type,
                  data = cu.summary, parms = list(split = 'information'))
> par(mfrow = c(1,2), mar = rep(0.1, 4))
> plot(fit1, margin = 0.05);  text(fit1, use.n = TRUE, cex = 0.8)
> plot(fit2, margin = 0.05);  text(fit2, use.n = TRUE, cex = 0.8)
```

Due to the wide labels at the bottom, we had to increase the figure space slightly and decrease the character size in order to avoid truncation at the left and right edges. Details for the first two splits in the Gini tree are

```
> summary(fit1, cp = 0.06)
Call:
rpart(formula = Reliability ~ Price + Country + Mileage + Type,
    data = cu.summary, parms = list(split = "gini"))
  n=85 (32 observations deleted due to missingness)

          CP nsplit rel error    xerror       xstd
1 0.30508475      0 1.0000000 1.0000000 0.07200310
2 0.08474576      1 0.6949153 0.6949153 0.07808305
3 0.05084746      2 0.6101695 0.7288136 0.07812633
4 0.03389831      3 0.5593220 0.7118644 0.07812633
5 0.01000000      4 0.5254237 0.7796610 0.07786634
```

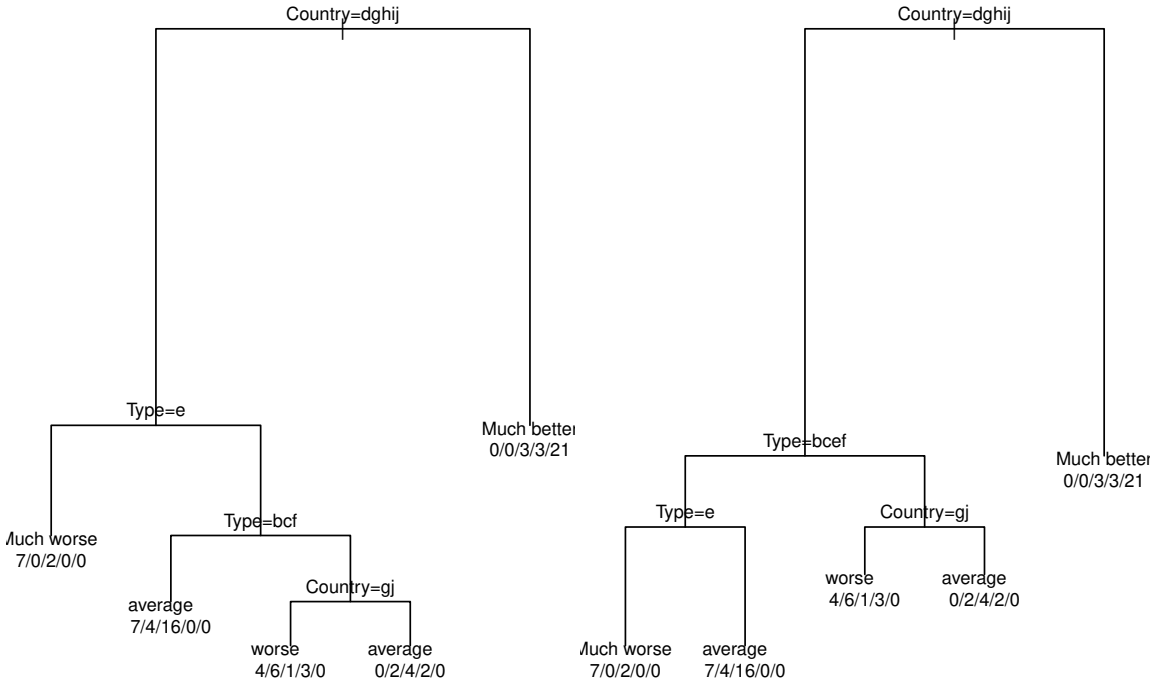


Figure 5: Displays the rpart-based model relating automobile Reliability to car type, price, and country of origin. The figure on the left uses the `gini` splitting index and the figure on the right uses the `information` splitting index.

```
Variable importance
Country    Type   Price
     66      22      12

Node number 1: 85 observations,    complexity param=0.3050847
  predicted class=average      expected loss=0.6941176  P(node) =1
    class counts:    18    12    26     8    21
   probabilities: 0.212 0.141 0.306 0.094 0.247
  left son=2 (58 obs) right son=3 (27 obs)
  Primary splits:
      Country splits as  ---LRRLLLL,  improve=15.220690, (0 missing)
      Type    splits as  RLLRLL,      improve= 4.288063, (0 missing)
      Price   < 11972.5 to the right, improve= 3.200000, (0 missing)
```

```
      Mileage < 24.5    to the left,  improve= 2.476190, (36 missing)

Node number 2: 58 observations,    complexity param=0.08474576
  predicted class=average      expected loss=0.6034483  P(node) =0.6823529
    class counts:    18    12    23     5     0
   probabilities: 0.310 0.207 0.397 0.086 0.000
  left son=4 (9 obs) right son=5 (49 obs)
  Primary splits:
      Type    splits as  RRRRLR,      improve=3.186567, (0 missing)
      Price   < 11232.5 to the left,  improve=2.563521, (0 missing)
      Mileage < 24.5    to the left,  improve=1.801587, (30 missing)
      Country splits as  ---L--RLRL,  improve=1.329310, (0 missing)

Node number 3: 27 observations
  predicted class=Much better  expected loss=0.2222222  P(node) =0.3176471
    class counts:     0     0     3     3    21
   probabilities: 0.000 0.000 0.111 0.111 0.778

Node number 4: 9 observations
  predicted class=Much worse   expected loss=0.2222222  P(node) =0.1058824
    class counts:     7     0     2     0     0
   probabilities: 0.778 0.000 0.222 0.000 0.000

Node number 5: 49 observations
  predicted class=average      expected loss=0.5714286  P(node) =0.5764706
    class counts:    11    12    21     5     0
   probabilities: 0.224 0.245 0.429 0.102 0.000
```

And for the information splitting the first split is

```
> fit3 <- rpart(Reliability ~ Price + Country + Mileage + Type,
                  data=cu.summary, parms=list(split='information'),
                maxdepth=2)
> summary(fit3)
Call:
rpart(formula = Reliability ~ Price + Country + Mileage + Type,
    data = cu.summary, parms = list(split = "information"), maxdepth = 2)
  n=85 (32 observations deleted due to missingness)

          CP nsplit rel error    xerror       xstd
```

```
1 0.30508475      0 1.0000000 1.0000000 0.07200310
2 0.05084746      1 0.6949153 0.7118644 0.07812633
3 0.01000000      2 0.6440678 0.7627119 0.07799644

Variable importance
Country    Type   Price
     73      16      11

Node number 1: 85 observations,    complexity param=0.3050847
  predicted class=average      expected loss=0.6941176  P(node) =1
    class counts:    18    12    26     8    21
   probabilities: 0.212 0.141 0.306 0.094 0.247
  left son=2 (58 obs) right son=3 (27 obs)
  Primary splits:
      Country splits as  ---LRRLLLL,  improve=38.541250, (0 missing)
      Type    splits as  RLLRLL,      improve=11.333260, (0 missing)
      Price   < 11972.5 to the right, improve= 6.241277, (0 missing)
      Mileage < 24.5    to the left,  improve= 5.548285, (36 missing)

Node number 2: 58 observations,    complexity param=0.05084746
  predicted class=average      expected loss=0.6034483  P(node) =0.6823529
    class counts:    18    12    23     5     0
   probabilities: 0.310 0.207 0.397 0.086 0.000
  left son=4 (36 obs) right son=5 (22 obs)
  Primary splits:
      Type    splits as  RLLRLL,      improve=9.280711, (0 missing)
      Price   < 11232.5 to the left,  improve=5.608640, (0 missing)
      Mileage < 24.5    to the left,  improve=5.593989, (30 missing)
      Country splits as  ---L--RRRL,  improve=2.891017, (0 missing)
  Surrogate splits:
      Price   < 10970   to the right, agree=0.879, adj=0.682, (0 split)
      Country splits as  ---R--RRRL,  agree=0.793, adj=0.455, (0 split)

Node number 3: 27 observations
  predicted class=Much better  expected loss=0.2222222  P(node) =0.3176471
    class counts:     0     0     3     3    21
   probabilities: 0.000 0.000 0.111 0.111 0.778

Node number 4: 36 observations
  predicted class=average      expected loss=0.5  P(node) =0.4235294
```

```
      class counts:    14     4    18     0     0
    probabilities: 0.389 0.111 0.500 0.000 0.000

Node number 5: 22 observations
  predicted class=worse        expected loss=0.6363636  P(node) =0.2588235
      class counts:     4     8     5     5     0
    probabilities: 0.182 0.364 0.227 0.227 0.000
```

The first 3 countries (Brazil, England, France) had only one or two cars in the listing, all of which were missing the reliability variable. There are no entries for these countries in the first node, leading to the $-$ symbol for the rule. The information measure has larger "improvements", consistent with the difference in scaling between the information and Gini criteria shown in figure 2, but the relative merits of different splits are fairly stable.

The two rules do not choose the same primary split at node 2. The data at this point are

	Compact	Large	Medium	Small	Sporty	Van
Much worse	2	2	4	2	7	1
worse	5	0	4	3	0	0
average	3	5	8	2	2	3
better	2	0	0	3	0	0
Much better	0	0	0	0	0	0

Since there are 6 different categories, all $2^5 = 32$ different combinations were explored, and as it turns out there are several with a nearly identical improvement. The Gini and information criteria make different "random" choices from this set of near ties. For the Gini index, *Sporty vs others* and *Compact/Small vs others* have improvements of 3.19 and 3.12, respectively. For the information index, the improvements are 6.21 versus 9.28. Thus the Gini index picks the first rule and information the second. Interestingly, the two splitting criteria arrive at exactly the same final nodes, for the full tree, although by different paths. (Compare the class counts of the terminal nodes).

We have said that for a categorical predictor with m levels, all 2^{m-1} different possible splits are tested.. When there are a large number of categories for the predictor, the computational burden of evaluating all of these subsets can become large. For instance, the call `rpart(Reliability  ., data=car90)` does not return for a *long,* long time: one of the predictors in that data set is an unordered factor with 30 levels! Luckily, for any ordered outcome there is a computational shortcut that allows the program to find the best split using only $m - 1$ comparisons. This includes the classification method when there are only two categories, along with the anova and Poisson methods to be introduced later.

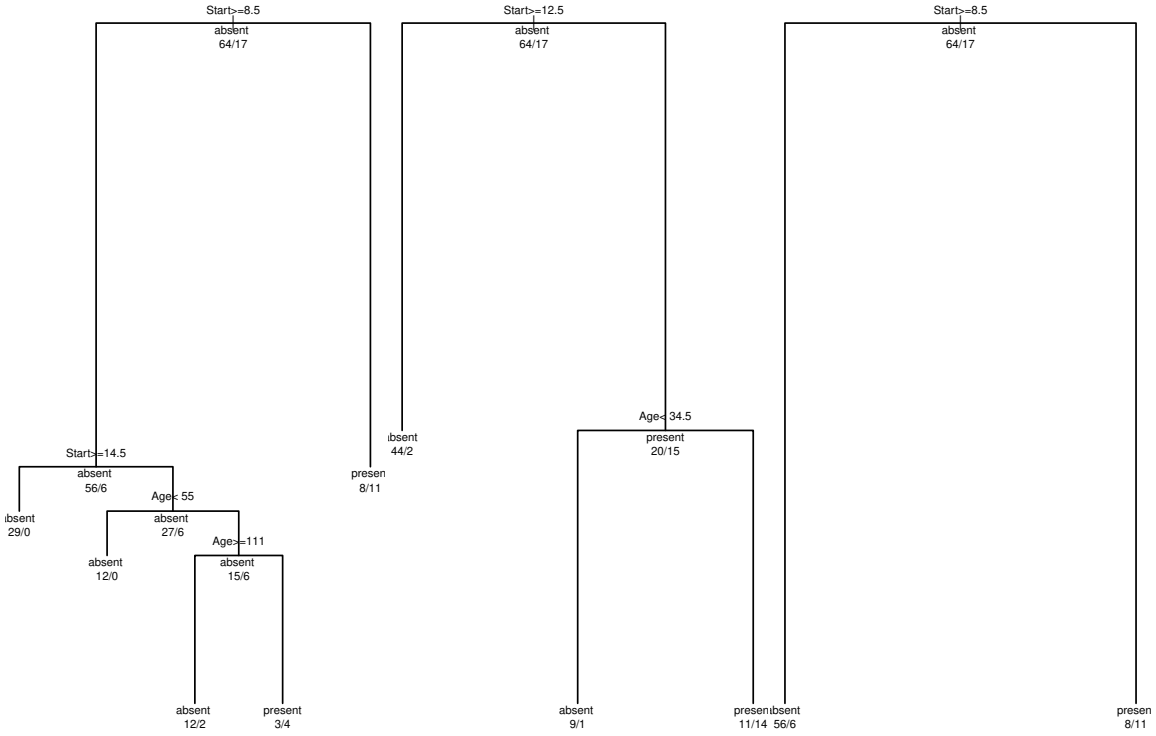


Figure 6: Displays the rpart-based models for the presence/absence of kyphosis. The figure on the left uses the default prior (0.79,0.21) and loss; the middle figure uses the user-defined prior (0.65,0.35) and default loss; and the third figure uses the default prior and the user-defined loss $L(1,2) = 3$, $L(2,1) = 2$).

6.3 Example: Kyphosis data

A third `class` method example explores the parameters `prior` and `loss`. The dataset kyphosis has 81 rows representing data on 81 children who have had corrective spinal surgery. The variables are:

Kyphosis	factor: postoperative deformity is present/absent
Age	numeric: age of child in months
Number	numeric: number of vertebrae involved in operation
Start	numeric: beginning of the range of vertebrae involved

```
> lmat <- matrix(c(0,3, 4,0), nrow = 2, ncol = 2, byrow = FALSE)
> fit1 <- rpart(Kyphosis ~  Age + Number + Start, data = kyphosis)
> fit2 <- rpart(Kyphosis ~  Age + Number + Start, data = kyphosis,
```

```
                    parms = list(prior = c(0.65, 0.35)))
  > fit3 <- rpart(Kyphosis ~  Age + Number + Start, data = kyphosis,
                    parms = list(loss = lmat))
  > par(mfrow = c(1, 3), mar = rep(0.1, 4))
  > plot(fit1);  text(fit1, use.n = TRUE, all = TRUE, cex = 0.8)
  > plot(fit2);  text(fit2, use.n = TRUE, all = TRUE, cex = 0.8)
  > plot(fit3);  text(fit3, use.n = TRUE, all = TRUE, cex = 0.8)
```

This example shows how even the initial split changes depending on the prior and loss that are specified. The first and third fits have the same initial split (Start < 8.5), but the improvement differs. The second fit splits Start at 12.5 which moves 46 people to the left instead of 62.

Looking at the leftmost tree, we see that the sequence of splits on the left hand branch yields only a single node classified as *present*. For any loss greater than 4 to 3, the routine will instead classify this node as *absent*, and the entire left side of the tree collapses, as seen in the right hand figure. This is not unusual — the most common effect of alternate loss matrices is to change the amount of pruning in the tree, more pruning in some branches and less in others, rather than to change the choice of splits.

The first node from the default tree is

```
 Node number 1: 81 observations,    complexity param=0.1765
  predicted class= absent  expected loss= 0.2099
     class counts:  64 17
    probabilities:  0.7901 0.2099
  left son=2 (62 obs) right son=3 (19 obs)
  Primary splits:
      Start  < 8.5  to the right, improve=6.762, (0 missing)
      Number < 5.5  to the left,  improve=2.867, (0 missing)
      Age    < 39.5 to the left,  improve=2.250, (0 missing)
  Surrogate splits:
      Number < 6.5  to the left,  agree=0.8025, (0 split)
```

The fit using the prior (0.65,0.35) is

```
Node number 1: 81 observations,    complexity param=0.302
  predicted class= absent  expected loss= 0.35
     class counts:  64 17
    probabilities:  0.65 0.35
  left son=2 (46 obs) right son=3 (35 obs)
  Primary splits:
      Start  < 12.5 to the right, improve=10.900, (0 missing)
```

```
      Number < 4.5  to the left,  improve= 5.087, (0 missing)
      Age    < 39.5 to the left,  improve= 4.635, (0 missing)
  Surrogate splits:
      Number < 3.5  to the left,  agree=0.6667, (0 split)
```

And first split under 4/3 losses is

```
Node number 1: 81 observations,    complexity param=0.01961
  predicted class= absent  expected loss= 0.6296
     class counts:  64 17
    probabilities:  0.7901 0.2099
  left son=2 (62 obs) right son=3 (19 obs)
  Primary splits:
      Start  < 8.5  to the right, improve=5.077, (0 missing)
      Number < 5.5  to the left,  improve=2.165, (0 missing)
      Age    < 39.5 to the left,  improve=1.535, (0 missing)
  Surrogate splits:
      Number < 6.5  to the left,  agree=0.8025, (0 split)
```

7 Regression

7.1 Definition

Up to this point the classification problem has been used to define and motivate our formulae. However, the partitioning procedure is quite general and can be extended by specifying 5 "ingredients":

- A splitting criterion, which is used to decide which variable gives the best split. For classification this was either the Gini or log-likelihood function. In the `anova` method the splitting criteria is $SS_T - (SS_L + SS_R)$, where $SS_T = \sum(y_i - \bar{y})^2$ is the sum of squares for the node, and SS_R, SS_L are the sums of squares for the right and left son, respectively. This is equivalent to choosing the split to maximize the between-groups sum-of-squares in a simple analysis of variance. This rule is identical to the regression option for `tree`.
- A summary statistic or vector, which is used to describe a node. The first element of the vector is considered to be the fitted value. For the anova method this is the mean of the node; for classification the response is the predicted class followed by the vector of class probabilities.
- The error of a node. This will be the variance of y for anova, and the predicted loss for classification.

- The prediction error for a new observation, assigned to the node. For anova this is $(y_{new} - \bar{y})$.
- Any necessary initialization.

The `anova` method leads to regression trees; it is the default method if y a simple numeric vector, i.e., not a factor, matrix, or survival object.

7.2 Example: Consumer Report car data

The dataset `car90` contains a collection of variables from the April, 1990 Consumer Reports; it has 34 variables on 111 cars. We've excluded 3 variables: tire size and model name because they are factors with a very large number of levels whose printout does not fit well in this report's page size, and rim size because it is too good a predictor of price and leads to a less interesting illustration. (Tiny cars are cheaper and have small rims.)

```
> cars <- car90[, -match(c("Rim", "Tires", "Model2"), names(car90))]
> carfit <- rpart(Price/1000 ~ ., data=cars)
> carfit
n=105 (6 observations deleted due to missingness)

node), split, n, deviance, yval
      * denotes terminal node

 1) root 105 7118.26700 15.805220
   2) Disp< 156 70 1491.62800 11.855870
     4) Country=Brazil,Japan,Japan/USA,Korea,Mexico,USA 58  421.21470 10.318470
       8) Type=Small 21   50.30983  7.629048 *
       9) Type=Compact,Medium,Sporty,Van 37  132.80330 11.844890 *
     5) Country=France,Germany,Sweden 12  270.72330 19.286670 *
   3) Disp>=156 35 2351.19600 23.703910
     6) HP.revs< 5550 24  980.27790 20.388710
      12) Disp< 267.5 16  395.99670 17.820060 *
      13) Disp>=267.5 8  267.58000 25.526000 *
     7) HP.revs>=5550 11  531.63680 30.937090 *
> printcp(carfit)
Regression tree:
rpart(formula = Price/1000 ~ ., data = cars)

Variables actually used in tree construction:
[1] Country Disp    HP.revs Type
```

```
Root node error: 7118.3/105 = 67.793

n=105 (6 observations deleted due to missingness)

        CP nsplit rel error  xerror    xstd
1 0.460146      0   1.00000 1.00939 0.16190
2 0.117905      1   0.53985 0.74001 0.11094
3 0.112343      2   0.42195 0.68477 0.10619
4 0.044491      3   0.30961 0.52755 0.10917
5 0.033449      4   0.26511 0.49119 0.10856
6 0.010000      5   0.23166 0.48643 0.11352
```

- The relative error is $1 - R^2$, similar to linear regression. The xerror is related to the PRESS statistic. The first split appears to improve the fit the most. The last split adds little improvement to the apparent error, and increases the cross-validated error.
- The 1-SE rule would choose a tree with 3 splits.
- This is a case where the default cp value of .01 may have over pruned the tree, since the cross-validated error is barely at a minimum. A rerun with the cp threshold at .001 gave little change, however.
- For any CP value between Sexprround(temp[3,1],2) and 0.12 the best model has one split; for any CP value between 0.04 and 0.11 the best model is with 2 splits; and so on.

The `print` and `summary` commands also recognizes the `cp` option, which allows the user to look at only the top few splits.

```
> summary(carfit, cp = 0.1)
Call:
rpart(formula = Price/1000 ~ ., data = cars)
  n=105 (6 observations deleted due to missingness)

           CP nsplit rel error    xerror      xstd
1 0.46014608      0 1.0000000 1.0093895 0.1618977
2 0.11790530      1 0.5398539 0.7400119 0.1109404
3 0.11234341      2 0.4219486 0.6847730 0.1061948
4 0.04449133      3 0.3096052 0.5275489 0.1091709
```

```
5 0.03344936      4 0.2651139 0.4911900 0.1085584
6 0.01000000      5 0.2316645 0.4864269 0.1135207

Variable importance
      Disp      Disp2     Weight       Tank         HP
        19         18         13         12         11
Wheel.base    Country    HP.revs      Gear2   Front.Hd
         9          6          4          3          1
      Type     Length   Steering      Width
         1          1          1          1

Node number 1: 105 observations,    complexity param=0.4601461
  mean=15.80522, MSE=67.79302
  left son=2 (70 obs) right son=3 (35 obs)
  Primary splits:
      Disp   < 156    to the left,  improve=0.4601461, (0 missing)
      Disp2  < 2.55   to the left,  improve=0.4601461, (0 missing)
      HP     < 154    to the left,  improve=0.4548845, (0 missing)
      Tank   < 17.8   to the left,  improve=0.4431325, (0 missing)
      Weight < 2890   to the left,  improve=0.3912428, (0 missing)
  Surrogate splits:
      Disp2      < 2.55   to the left,  agree=1.000, adj=1.000, (0 split)
      Weight     < 3095   to the left,  agree=0.914, adj=0.743, (0 split)
      HP         < 139    to the left,  agree=0.895, adj=0.686, (0 split)
      Tank       < 17.95  to the left,  agree=0.895, adj=0.686, (0 split)
      Wheel.base < 105.5  to the left,  agree=0.857, adj=0.571, (0 split)

Node number 2: 70 observations,    complexity param=0.1123434
  mean=11.85587, MSE=21.30898
  left son=4 (58 obs) right son=5 (12 obs)
  Primary splits:
      Country splits as  L-RRLLLLRL, improve=0.5361191, (0 missing)
      Tank    < 15.65  to the left,  improve=0.3805426, (0 missing)
      Weight  < 2567.5 to the left,  improve=0.3691043, (0 missing)
      Type    splits as  R-RLRR,     improve=0.3649737, (0 missing)
      HP      < 105.5  to the left,  improve=0.3577873, (0 missing)
  Surrogate splits:
      Tank         < 17.8   to the left,  agree=0.857, adj=0.167, (0 split)
      Rear.Seating < 28.75  to the left,  agree=0.843, adj=0.083, (0 split)
```

```
Node number 3: 35 observations,    complexity param=0.1179053
  mean=23.70391, MSE=67.17703
  left son=6 (24 obs) right son=7 (11 obs)
  Primary splits:
      HP.revs  < 5550   to the left,  improve=0.3569594, (0 missing)
      HP       < 173.5  to the left,  improve=0.3146835, (0 missing)
      Frt.Shld < 57.75  to the right, improve=0.2554028, (0 missing)
      Front.Hd < 3.25   to the right, improve=0.2278477, (0 missing)
      Type     splits as  RLR-RL,     improve=0.2178764, (0 missing)
  Surrogate splits:
      Country  splits as  -R-RR----L, agree=0.886, adj=0.636, (0 split)
      Gear2    < 2.735  to the left,  agree=0.886, adj=0.636, (0 split)
      Disp     < 172.5  to the right, agree=0.800, adj=0.364, (0 split)
      Disp2    < 2.85   to the right, agree=0.800, adj=0.364, (0 split)
      Front.Hd < 3.25   to the right, agree=0.800, adj=0.364, (0 split)

Node number 4: 58 observations
  mean=10.31847, MSE=7.262322

Node number 5: 12 observations
  mean=19.28667, MSE=22.56027

Node number 6: 24 observations
  mean=20.38871, MSE=40.84491

Node number 7: 11 observations
  mean=30.93709, MSE=48.33062
```

The first split on displacement partitions the 105 observations into groups of 70 and 58 (nodes 2 and 3) with mean prices of 12 and 10

- The improvement listed is the percent change in SS for this split, i.e., $1 - (SS_{right} + SS_{left})/SS_{parent}$, which is the gain in R^2 for the fit.
- The data set has displacement of the engine in both cubic inches (Disp) and liters (Disp2). The second is a perfect surrogate split for the first, obviously.
- The weight and displacement are very closely related, as shown by the surrogate split agreement of 91%.
- Not all the countries are represented in node 3, e.g., there are no larger cars from Brazil. This is indicated by a - in the list of split directions.

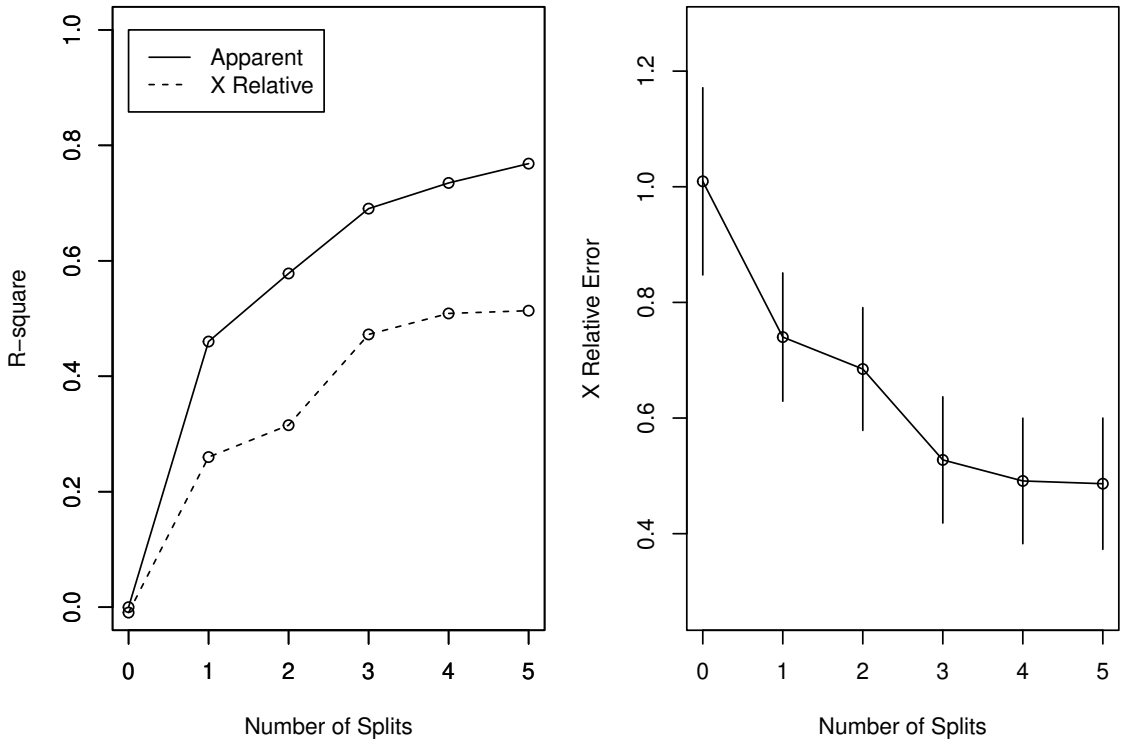


Figure 7: Both plots were obtained using the function `rsq.rpart(fit3)`. The figure on the left shows that the first split offers the most information. The figure on the right suggests that the tree should be pruned to include only 1 or 2 splits.

```
Regression tree:
rpart(formula = Price/1000 ~ ., data = cars)

Variables actually used in tree construction:
[1] Country Disp    HP.revs Type

Root node error: 7118.3/105 = 67.793

n=105 (6 observations deleted due to missingness)

        CP nsplit rel error  xerror    xstd
1 0.460146      0   1.00000 1.00939 0.16190
2 0.117905      1   0.53985 0.74001 0.11094
3 0.112343      2   0.42195 0.68477 0.10619
```

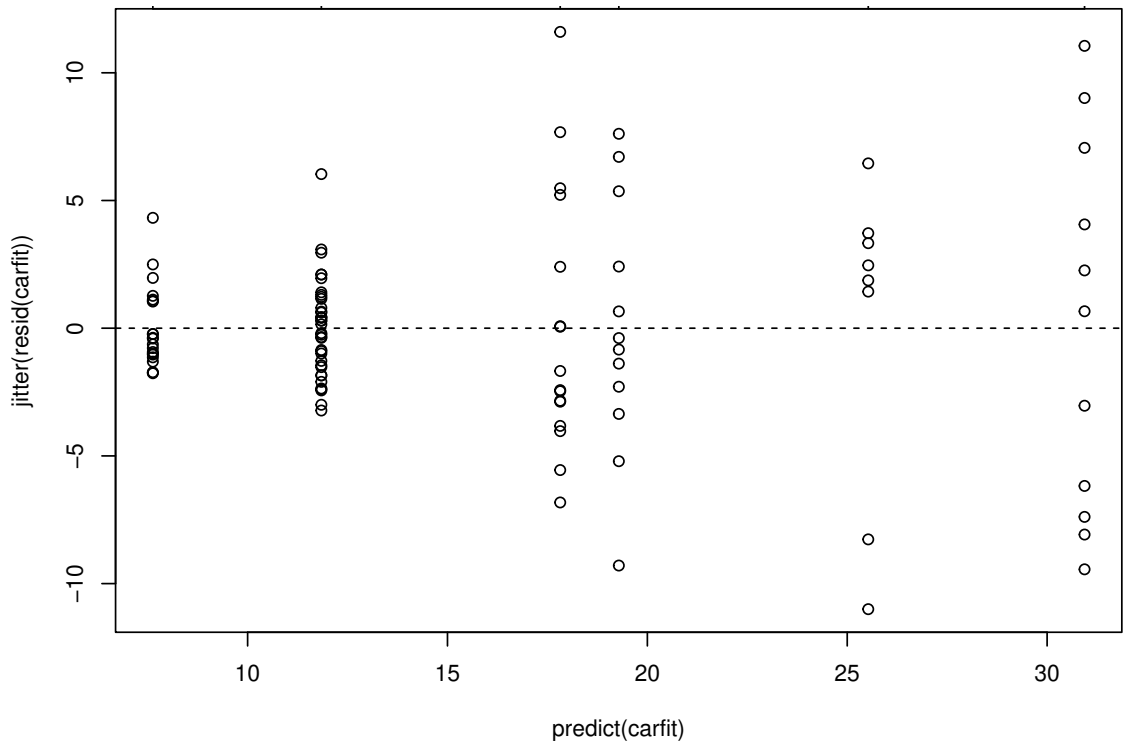


Figure 8: This plot shows the (observed-expected) cost of cars versus the predicted cost of cars based on the nodes/leaves in which the cars landed. There appears to be more variability in node 7 than in some of the other leaves.

```
4 0.044491      3   0.30961 0.52755 0.10917
5 0.033449      4   0.26511 0.49119 0.10856
6 0.010000      5   0.23166 0.48643 0.11352
```

Other plots can be used to help determine the best cp value for this model. The function `rsq.rpart` plots the jackknifed error versus the number of splits. Of interest is the smallest error, but any number of splits within the "error bars" (1-SE rule) are considered a reasonable number of splits (in this case, 1 or 2 splits seem to be sufficient). As is often true with modeling, simpler is often better. Another useful plot is the R^2 versus number of splits. The (1 - apparent error) and (1 - relative error) show how much is gained with additional splits. This plot highlights the differences between the R^2 values.

Finally, it is possible to look at the residuals from this model, just as with a regular linear regression fit, as shown in figure 8 produced by the following.

```
> plot(predict(carfit), jitter(resid(carfit)))
> temp <- carfit$frame[carfit$frame$var == '<leaf>',]
> axis(3, at = temp$yval, as.character(row.names(temp)))
> mtext('leaf number', side = 3, line = 3)
> abline(h = 0, lty = 2)
```

7.3 Example: Stage C data (anova method)

The stage C prostate cancer data of the earlier section can also be fit using the anova method, by treating the status variable as though it were continuous.

```
> cfit2 <- rpart(pgstat ~ age + eet + g2 + grade + gleason + ploidy,
                 data = stagec)
> printcp(cfit2)
Regression tree:
rpart(formula = pgstat ~ age + eet + g2 + grade + gleason + ploidy,
    data = stagec)

Variables actually used in tree construction:
[1] age    g2     grade  ploidy

Root node error: 34.027/146 = 0.23306

n= 146

        CP nsplit rel error  xerror     xstd
1 0.152195      0   1.00000 1.00922 0.045167
2 0.054395      1   0.84781 0.98806 0.069742
3 0.032487      3   0.73901 1.03223 0.083042
4 0.019932      4   0.70653 1.02073 0.084044
5 0.018833      5   0.68660 1.00406 0.086101
6 0.017493      7   0.64893 0.99749 0.085544
7 0.013027      8   0.63144 1.01708 0.090030
8 0.010000      9   0.61841 1.02463 0.094926
> print(cfit2, cp = 0.03)
n= 146

node), split, n, deviance, yval
      * denotes terminal node
```

```
1) root 146 34.027400 0.3698630
  2) grade< 2.5 61  7.672131 0.1475410
    4) g2< 13.19 40  1.900000 0.0500000 *
    5) g2>=13.19 21  4.666667 0.3333333 *
  3) grade>=2.5 85 21.176470 0.5294118
    6) g2< 13.2 40  9.775000 0.4250000 *
    7) g2>=13.2 45 10.577780 0.6222222
     14) g2>=17.91 22  5.090909 0.3636364 *
     15) g2< 17.91 23  2.608696 0.8695652 *
```

If this tree is compared to the earlier results, we see that it has chosen exactly the same variables and split points as before. The only addition is further splitting of node 2, the upper left "No" of figure 3. This is no accident, for the two class case the Gini splitting rule reduces to $2p(1-p)$, which is the variance of a node.

The two methods differ in their evaluation and pruning, however. Note that nodes 4 and 5, the two children of node 2, contain 2/40 and 7/21 progressions, respectively. For classification purposes both nodes have the same predicted value (No) and the split will be discarded since the error (# of misclassifications) with and without the split is identical. In the regression context the two predicted values of .05 and .33 *are* different — the split has identified a nearly pure subgroup of significant size.

This setup is known as *odds regression*, and may be a more sensible way to evaluate a split when the emphasis of the model is on understanding/explanation rather than on prediction error per se. Extension of this rule to the multiple class problem is appealing, but has not yet been implemented in `rpart`.

8 Poisson regression

8.1 Definition

The Poisson splitting method attempts to extend `rpart` models to event rate data. The model in this case is

$$\lambda = f(x)$$

where λ is an event rate and x is some set of predictors. As an example consider hip fracture rates. For each county in the United States we can obtain

- number of fractures in patients age 65 or greater (from Medicare files)
- population of the county (US census data)
- potential predictors such as

- socio-economic indicators
- number of days below freezing
- ethnic mix
- physicians/1000 population
- etc.

Such data would usually be approached by using Poisson regression; can we find a tree based analogue? In adding criteria for rates regression to this ensemble, the guiding principle was the following: the between groups sum-of-squares is not a very robust measure, yet tree based regression works fairly well for this data. So do the simplest (statistically valid) thing possible.

Let c_i be the observed event count for observation i, t_i be the observation time, and $x_{ij}, j = 1, \ldots, p$ be the predictors. The y variable for the program will be a 2 column matrix.

Splitting criterion: The likelihood ratio test for two Poisson groups

$$D_{\text{parent}} - \left(D_{\text{left son}} + D_{\text{right son}}\right)$$

Summary statistics: The observed event rate and the number of events.

$$\hat{\lambda} \quad = \quad \frac{\#\text{ events}}{\text{total time}} = \frac{\sum c_i}{\sum t_i}$$

Error of a node: The within node deviance.

$$D = \sum \left[c_i \log\left(\frac{c_i}{\hat{\lambda} t_i}\right) - (c_i - \hat{\lambda} t_i) \right]$$

Prediction error: The deviance contribution for a new observation, using $\hat{\lambda}$ of the node as the predicted rate.

8.2 Improving the method

There is a problem with the criterion just proposed, however: cross-validation of a model often produces an infinite value for the deviance. The simplest case where this occurs is easy to understand. Assume that some terminal node of the tree has 20 subjects, but only 1 of the 20 has experienced any events. The cross-validated error (deviance) estimate for that node will have one subset — the one where the subject with an event is left out — which has $\hat{\lambda} = 0$. When we use the prediction for the 10% of subjects who were set aside, the deviance contribution of the subject with an event is

$$\ldots + c_i \log(c_i/0) + \ldots$$

which is infinite since $c_i > 0$. The problem is that when $\hat{\lambda} = 0$ the occurrence of an event is infinitely improbable, and, using the deviance measure, the corresponding model is then infinitely bad.

One might expect this phenomenon to be fairly rare, but unfortunately it is not so. One given of tree-based modeling is that a right-sized model is arrived at by purposely over-fitting the data and then pruning back the branches. A program that aborts due to a numeric exception during the first stage is uninformative to say the least. Of more concern is that this edge effect does not seem to be limited to the pathological case detailed above. Any near approach to the boundary value $\lambda = 0$ leads to large values of the deviance, and the procedure tends to discourage any final node with a small number of events.

An ad hoc solution is to use the revised estimate

$$\hat{\hat{\lambda}} = \max\left(\hat{\lambda}, \frac{k}{\sum t_i}\right)$$

where k is 1/2 or 1/6. That is, pure nodes are given a partial event. (This is similar to the starting estimates used in the GLM program for a Poisson regression.) This is unsatisfying, however, and we propose instead using a shrinkage estimate.

Assume that the true rates λ_j for the leaves of the tree are random values from a Gamma(μ, σ) distribution. Set μ to the observed overall event rate $\sum c_i / \sum t_i$, and let the user choose as a prior the coefficient of variation $k = \sigma/\mu$. A value of $k = 0$ represents extreme pessimism ("the leaf nodes will all give the same result"), whereas $k = \infty$ represents extreme optimism. The Bayes estimate of the event rate for a node works out to be

$$\hat{\lambda}_k = \frac{\alpha + \sum c_i}{\beta + \sum t_i},$$

where $\alpha = 1/k^2$ and $\beta = \alpha/\hat{\lambda}$.

This estimate is scale invariant, has a simple interpretation, and shrinks least those nodes with a large amount of information. In practice, a value of $k = 10$ does essentially no shrinkage. For `method='poisson'`, the optional parameters list is the single number k, with a default value of 1.

8.3 Example: solder.balance data

The solder.balance data frame, as explained in the R help file, is a dataset with 900 observations which are the results of an experiment varying 5 factors relevant to the wave-soldering procedure for mounting components on printed circuit boards. The full version of the data (unbalanced) is available in the `survival` package. The response variable, skips, is a count of how many solder skips appeared to a visual inspection. The other variables are listed below:

Opening	factor: amount of clearance around the mounting pad (S < M < L)
Solder	factor: amount of solder used (Thin < Thick)
Mask	factor: Type of solder mask used (5 possible)
PadType	factor: Mounting pad used (10 possible)
Panel	factor: panel (1, 2 or 3) on board being counted

In this call, the `rpart.control` options are modified: `maxcompete = 2` means that only 2 other competing splits are listed (default is 4); `cp = .05` means that a smaller tree will be built initially (default is .01). The y variable for Poisson partitioning may be a two column matrix containing the observation time in column 1 and the number of events in column 2, or it may be a vector of event counts alone.

```
> sfit <- rpart(skips ~ Opening + Solder + Mask + PadType + Panel,
                data = solder.balance, method = 'poisson',
                control = rpart.control(cp = 0.05, maxcompete = 2))
> sfit
n= 720

node), split, n, deviance, yval
      * denotes terminal node

 1) root 720 6855.6900  4.965278
   2) Opening=L,M 480 1803.1630  1.913780
     4) Mask=A1.5,A3,B3 360  718.0687  1.038308 *
     5) Mask=B6 120  599.6246  4.542376 *
   3) Opening=S 240 2543.1750 11.065710
     6) Mask=A1.5,A3 120  596.4945  4.550696 *
     7) Mask=B3,B6 120  962.4191 17.570510
      14) Solder=Thick 60  256.9253 10.398430 *
      15) Solder=Thin 60  343.8841 24.700420 *
```

- The response value is the expected event rate (with a time variable), or in this case the expected number of skips. The values are shrunk towards the global estimate of 5.53 skips/observation.
- The deviance is the same as the null deviance (sometimes called the residual deviance) that you'd get when calculating a Poisson glm model for the given subset of data.

```
> summary(sfit, cp = 0.1)
Call:
rpart(formula = skips ~ Opening + Solder + Mask + PadType + Panel,
    data = solder.balance, method = "poisson", control = rpart.control(cp = 0.05,
        maxcompete = 2))
  n= 720

          CP nsplit rel error    xerror       xstd
1 0.36602468      0 1.0000000 1.0018492 0.06421145
2 0.14356853      1 0.6339753 0.6386714 0.03709864
3 0.07081271      2 0.4904068 0.4950817 0.02692013
4 0.05274593      3 0.4195941 0.4590287 0.02547590
5 0.05000000      4 0.3668481 0.4012848 0.02171765

Variable importance
Opening    Mask  Solder
     58      34       8

Node number 1: 720 observations,    complexity param=0.3660247
  events=3575,  estimated rate=4.965278 , mean deviance=9.521792
  left son=2 (480 obs) right son=3 (240 obs)
  Primary splits:
      Opening splits as  LLR,  improve=2509.3530, (0 missing)
      Mask    splits as  LLRR, improve=1323.3680, (0 missing)
      Solder  splits as  LR,   improve= 936.9548, (0 missing)

Node number 2: 480 observations
  events=918,  estimated rate=1.91378 , mean deviance=3.75659

Node number 3: 240 observations,    complexity param=0.1435685
  events=2657,  estimated rate=11.06571 , mean deviance=10.59656
  left son=6 (120 obs) right son=7 (120 obs)
  Primary splits:
      Mask    splits as  LLRR,       improve=984.2639, (0 missing)
      Solder  splits as  LR,         improve=631.5271, (0 missing)
      PadType splits as  RRRRLLRLRL, improve=244.9255, (0 missing)

Node number 6: 120 observations
  events=546,  estimated rate=4.550696 , mean deviance=4.970788
```

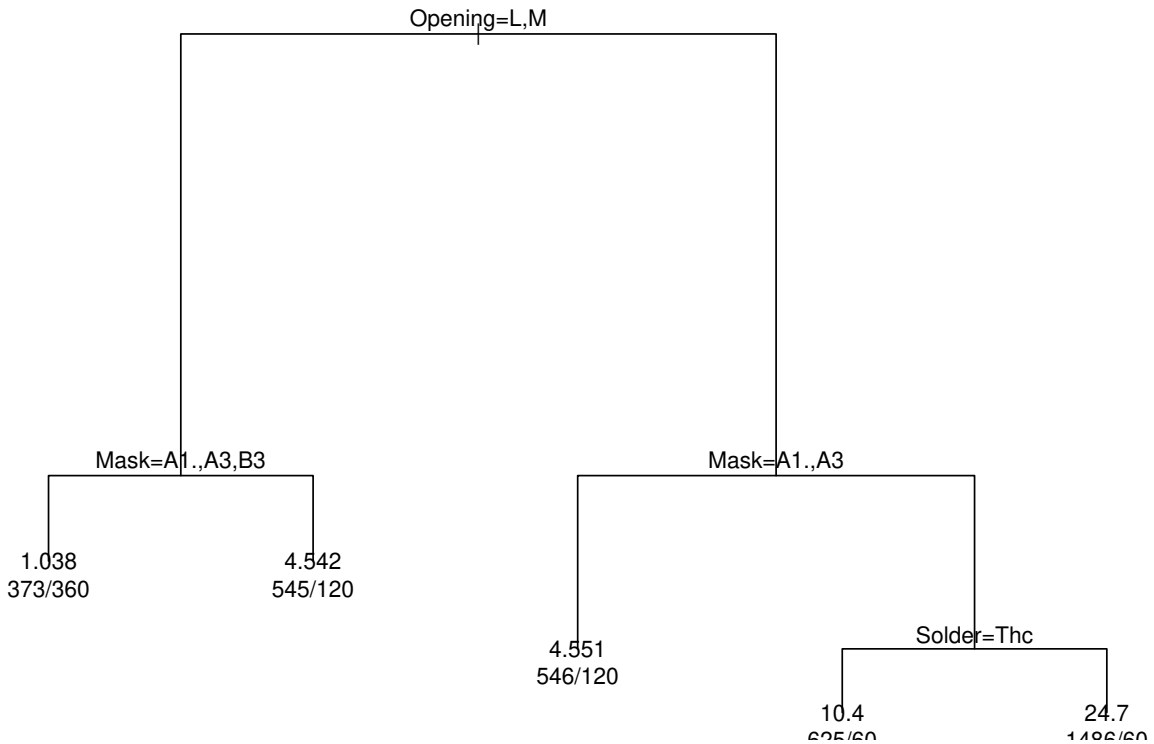


Figure 9: The first figure shows the solder.balance data, fit with the `poisson` method, using a `cp` value of 0.05. The second figure shows the same fit, but with a `cp` value of 0.10. The function `prune.rpart` was used to produce the smaller tree.

```
Node number 7: 120 observations
  events=2111,  estimated rate=17.57051 , mean deviance=8.020159
```

- The improvement is $\text{Deviance}_{\text{parent}} - (\text{Deviance}_{\text{left}} + \text{Deviance}_{\text{right}})$, which is the likelihood ratio test for comparing two Poisson samples.
- The cross-validated error has been found to be overly pessimistic when describing how much the error is improved by each split. This is likely an effect of the boundary effect mentioned earlier, but more research is needed.
- The variation xstd is not as useful, given the bias of xerror.

```
> par(mar = rep(0.1, 4))
> plot(sfit)
> text(sfit, use.n = TRUE, min = 3)
> fit.prune <- prune(sfit, cp = 0.10)
> plot(fit.prune)
> text(fit.prune, use.n = TRUE, min = 2)
```

The `use.n = TRUE` option specifies that number of events / total N should be listed along with the predicted rate (number of events/person-years). The function `prune` trims the tree `fit` to the `cp` value 0.10. The same tree could have been created by specifying `cp = 0.10` in the original call to `rpart`.

8.4 Example: Stage C Prostate cancer, survival method

One special case of the Poisson model is of particular interest for medical consulting (such as the authors do). Assume that we have survival data, i.e., each subject has either 0 or 1 event. Further, assume that the time values have been pre-scaled so as to fit an exponential model. That is, stretch the time axis so that a Kaplan-Meier plot of the data will be a straight line when plotted on the logarithmic scale. An approximate way to do this is

```
> require(survival)
> temp <- coxph(Surv(pgtime, pgstat) ~ 1, stagec)
> newtime <- predict(temp, type = 'expected')
```

and then do the analysis using the `newtime` variable. (This replaces each time value by $\Lambda(t)$, where Λ is the cumulative hazard function).

A slightly more sophisticated version of this which we will call *exponential scaling* gives a straight line curve for log(survival) under a parametric exponential model. The only difference from the approximate scaling above is that a subject who is censored between observed death times will receive "credit" for the intervening interval, i.e., we assume the baseline hazard to be linear between observed deaths. If the data is pre-scaled in this way, then the Poisson model above is equivalent to the *local full likelihood* tree model of LeBlanc and Crowley [3]. They show that this model is more efficient than the earlier suggestion of Therneau et al. [7] to use the martingale residuals from a Cox model as input to a regression tree (anova method). Exponential scaling or method='exp' is the default if y is a `Surv` object.

Let us again return to the stage C cancer example. Besides the variables explained previously, we will use pgtime, which is time to tumor progression.

```
> require(survival)
> pfit <- rpart(Surv(pgtime, pgstat) ~ age + eet + g2 + grade +
```

```
                  gleason + ploidy, data = stagec)
> print(pfit)
n= 146

node), split, n, deviance, yval
      * denotes terminal node

 1) root 146 192.111100 1.0000000
   2) grade< 2.5 61  44.799010 0.3634439
     4) g2< 11.36 33   9.117405 0.1229835 *
     5) g2>=11.36 28  27.602190 0.7345610
      10) gleason< 5.5 20  14.297110 0.5304115 *
      11) gleason>=5.5 8  11.094650 1.3069940 *
   3) grade>=2.5 85 122.441500 1.6148600
     6) age>=56.5 75 103.062900 1.4255040
      12) gleason< 7.5 50  66.119800 1.1407320
        24) g2< 13.475 24  27.197170 0.8007306 *
        25) g2>=13.475 26  36.790960 1.4570210
          50) g2>=17.915 15  20.332740 0.9789825 *
          51) g2< 17.915 11  13.459010 2.1714480 *
      13) gleason>=7.5 25  33.487250 2.0307290
        26) g2>=15.29 10  11.588480 1.2156230 *
        27) g2< 15.29 15  18.939150 2.7053610 *
     7) age< 56.5 10  13.769010 3.1822320 *
> pfit2 <- prune(pfit, cp = 0.016)
> par(mar = rep(0.2, 4))
> plot(pfit2, uniform = TRUE, branch = 0.4, compress = TRUE)
> text(pfit2, use.n = TRUE)
```

Note that the primary split on grade is the same as when status was used as a dichotomous endpoint, but that the splits thereafter differ.

Suppose that we wish to simplify this tree, so that only four splits remain. Looking at the table of complexity parameters, we see that `prune(fit, cp = 0.016)` would give the desired result, which is shown in figure 10. From the figure node 4 (leftmost terminal) has only 1 event for 33 subjects, a relative death rate of .133 times the overall rate, and is defined by grade = 1–2 and g2 $<$ 11.36.

For a final summary of the model, it can be helpful to plot the probability of survival based on the final bins in which the subjects landed. To create new variables based on the rpart groupings, using the `where` component of the fit, as shown in figure 11. We'll further prune the tree down to four nodes by removing the split at node 6.

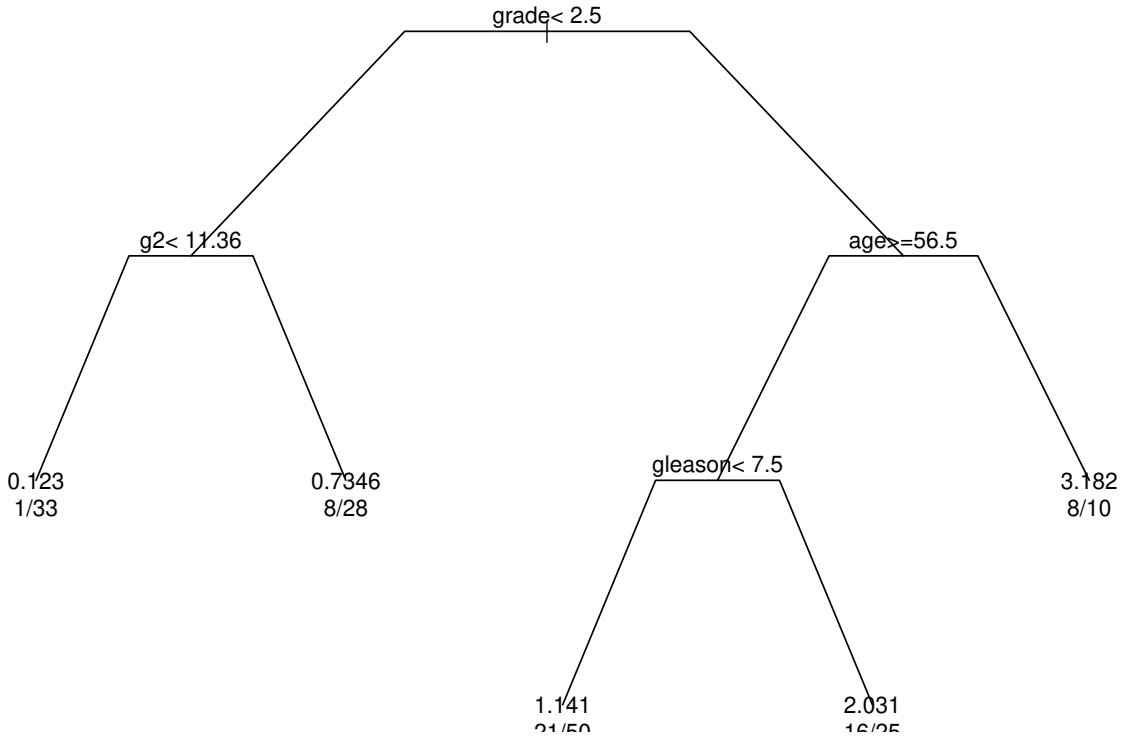


Figure 10: Survival fit to the stage C prostate data.

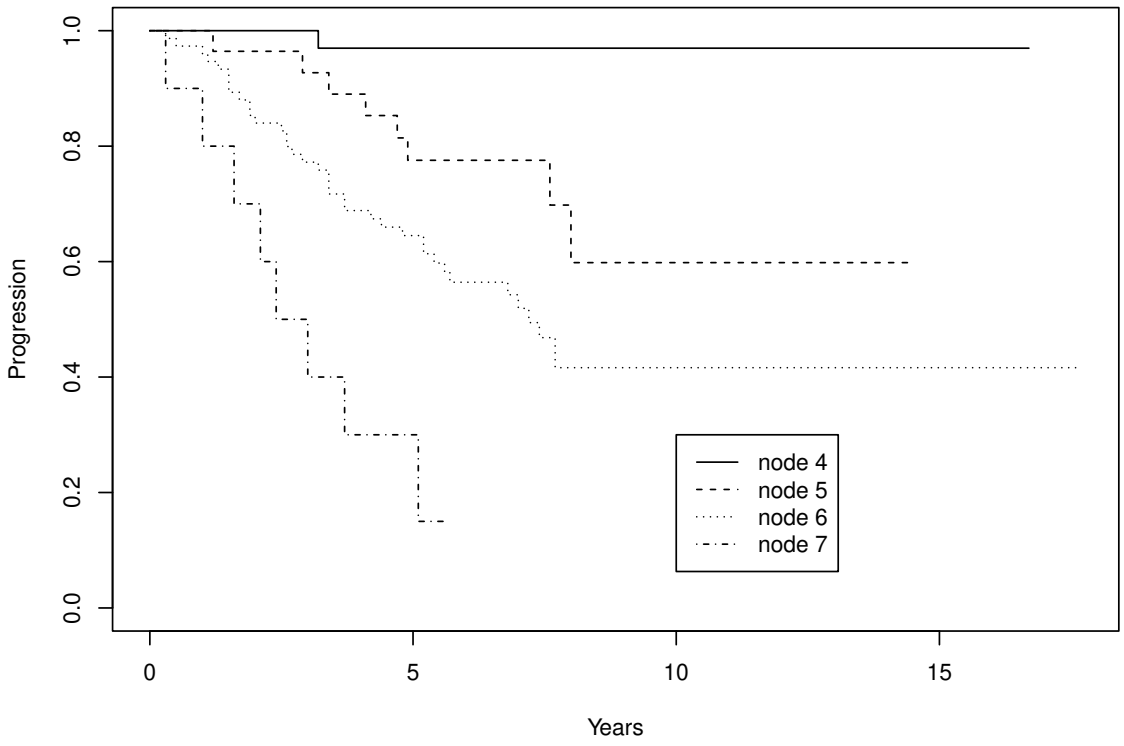


Figure 11: Survival plot based on snipped rpart object. The probability of tumor progression is greatest in node 8, which has patients who are older and have a higher initial tumor grade.

```
> temp <- snip.rpart(pfit2, 6)
> km <- survfit(Surv(pgtime, pgstat) ~ temp$where, stagec)
> plot(km, lty = 1:4, mark.time = FALSE,
       xlab = "Years", ylab = "Progression")
> legend(10, 0.3, paste('node', c(4,5,6,7)), lty = 1:4)
```

8.5 Open issues

The default value of the shrinkage parameter k is 1. This corresponds to prior coefficient of variation of 1 for the estimated λ_j. We have not nearly enough experience to decide if this is a good value. (It does stop the log(0) message though).

Cross-validation does not work very well. The procedure gives very conservative results, and quite often declares the no-split tree to be the best. This may be another artifact of the edge effect.

9 Plotting options

This section examines the various options that are available when plotting an rpart object. For simplicity, the same model (data from Example 1) will be used throughout. You have doubtless already noticed the use of `par(mar =)` in the examples. The plot function for rpart uses the general `plot` function to set up the plot region. By default, this leaves space for axes, legends or titles on the bottom, left, and top. Room for axes is not needed in general for rpart plots, and for this report we also do not have top titles. For the small plots in this report it was important to use all of the page, so we reset these for each plot. (Due to the way that Sweave works each plot is a separate environment, so the par() parameters do not endure from plot to plot.)

The simplest labeled plot is called by using `plot` and `text` without changing any of the defaults. This is useful for a first look, but sometimes you'll want more information about each of the nodes.

```
> fit <- rpart(pgstat ~  age + eet + g2 + grade + gleason + ploidy,
               stagec, control = rpart.control(cp = 0.025))
> par(mar = rep(0.2, 4))
> plot(fit)
> text(fit)
```

The next plot has uniform stem lengths (`uniform = TRUE`), has extra information (`use.n = TRUE`) specifying number of subjects at each node, and has labels on all the nodes, not just the terminal nodes (`all = TRUE`).

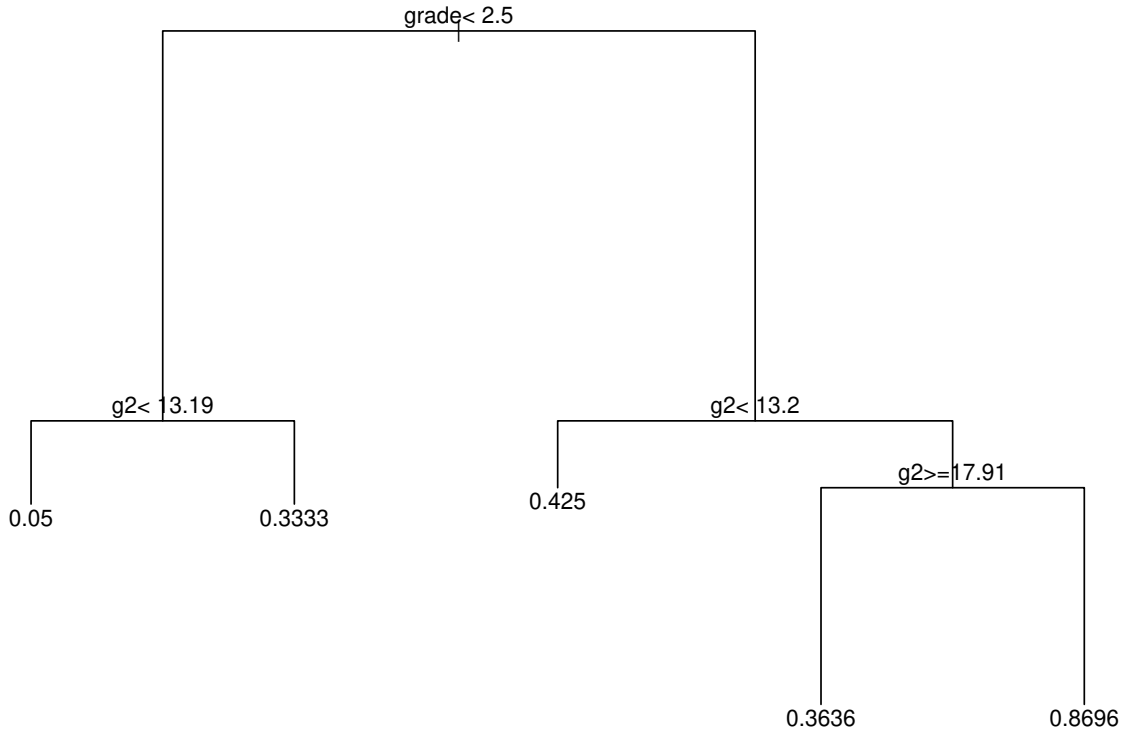


Figure 12: `plot(fit); text(fit)`

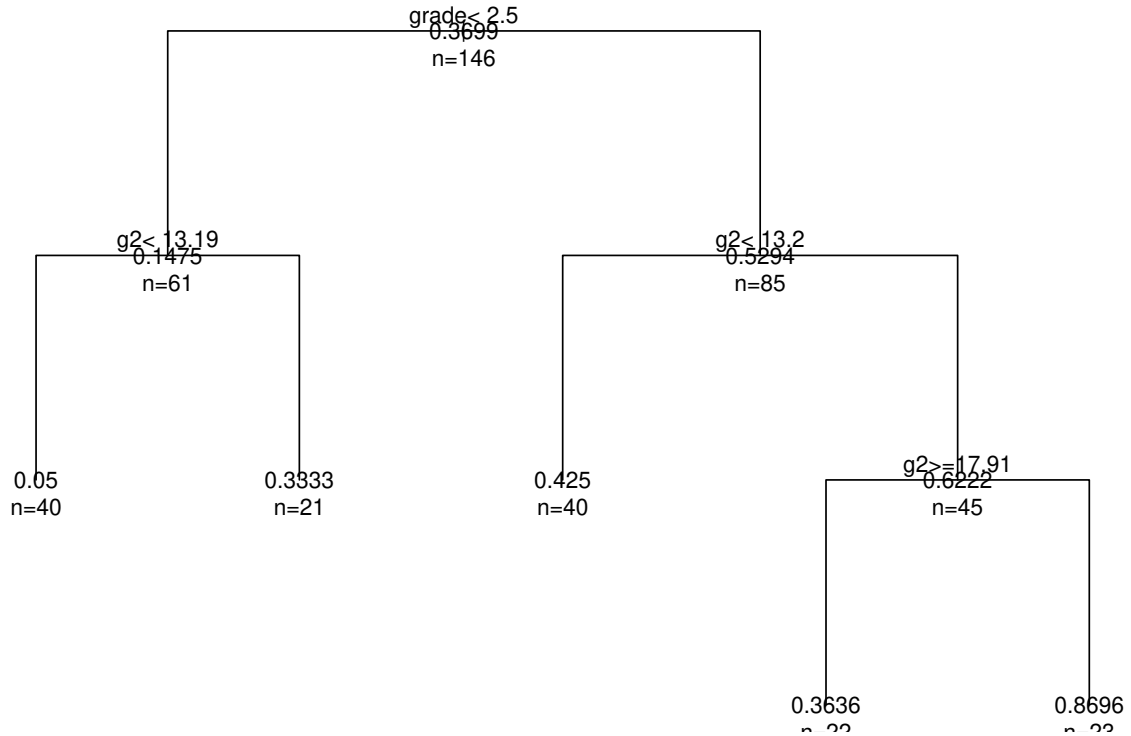


Figure 13: `plot(fit, uniform = TRUE); text(fit,use.n = TRUE,all = TRUE)`

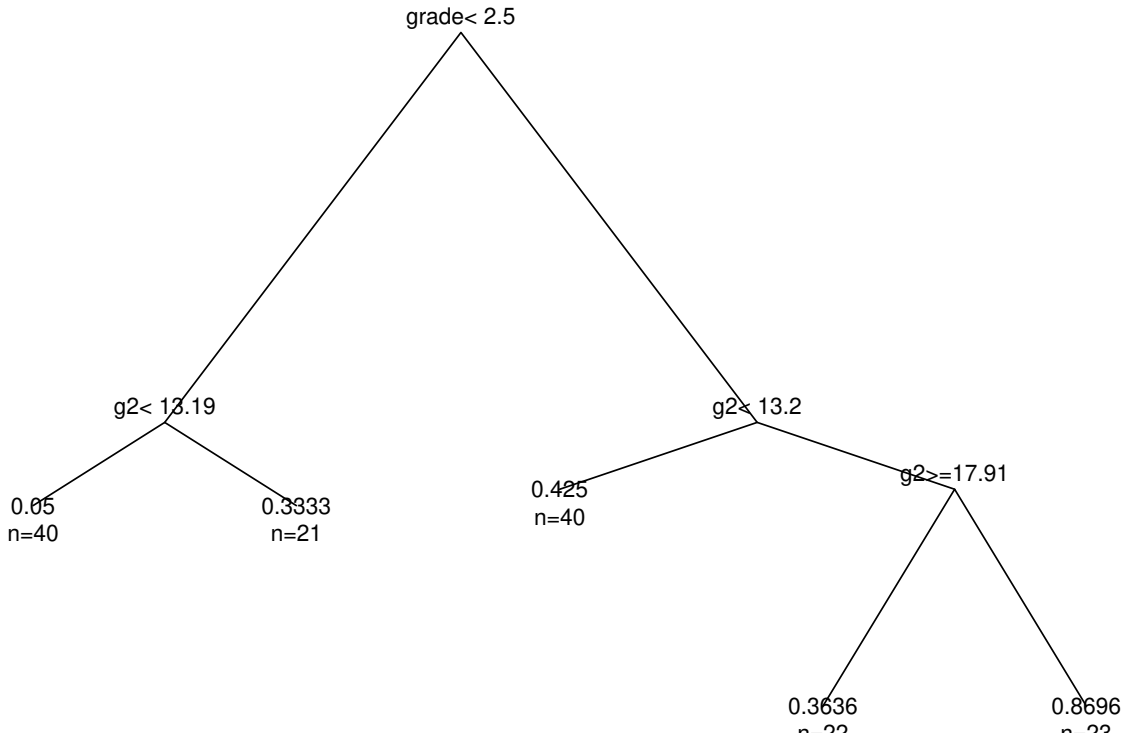


Figure 14: `plot(fit, branch=0); text(fit,use.n = TRUE)`

```
> par(mar = rep(0.2, 4))
> plot(fit, uniform = TRUE)
> text(fit, use.n = TRUE, all = TRUE)
```

Fancier plots can be created by modifying the `branch` option, which controls the shape of branches that connect a node to its children. The default for the plots is to have square shouldered trees (`branch = 1.0`). This can be taken to the other extreme with no shoulders at all (`branch=0`).

```
> par(mar = rep(0.2, 4))
> plot(fit, branch = 0)
> text(fit, use.n = TRUE)
```

These options can be combined with others to create the plot that fits your particular needs. The default plot may be inefficient in it's use of space: the terminal nodes will always lie at x-coordinates of 1,2,.... The `compress` option attempts to improve this by overlapping some nodes. It has little effect on figure 15, but in figure 4 it allows the lowest branch to

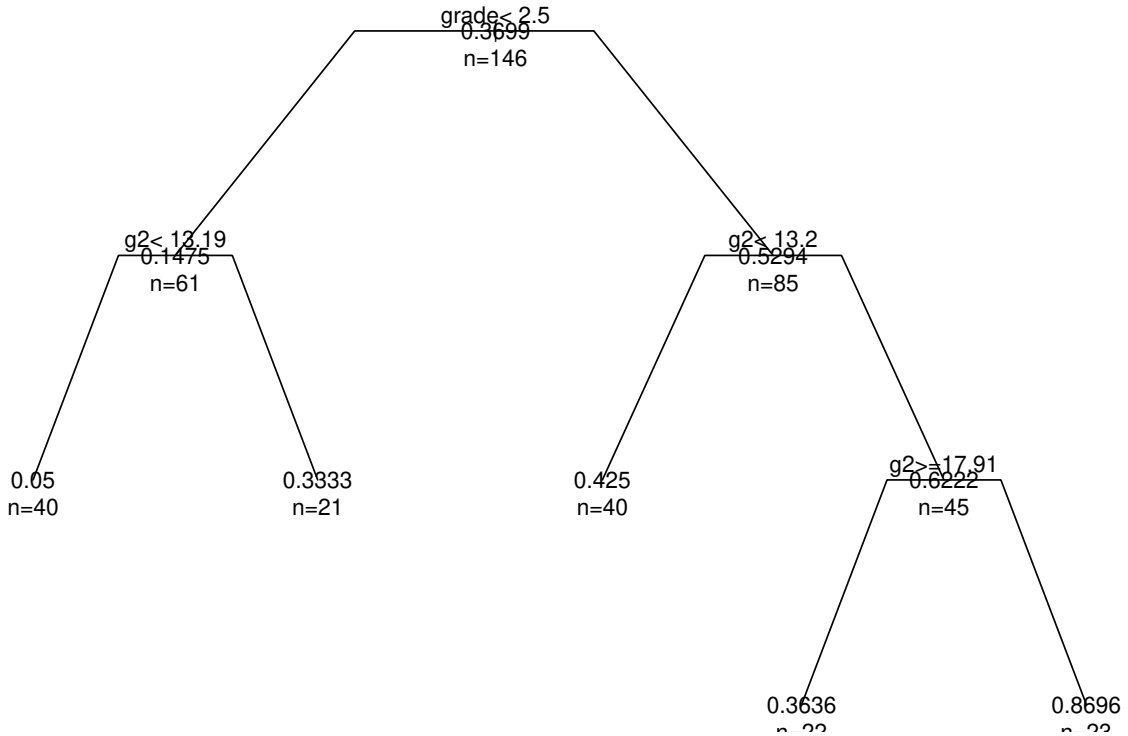


Figure 15: `plot(fit, branch = 0.4, uniform = TRUE, compress = TRUE)`

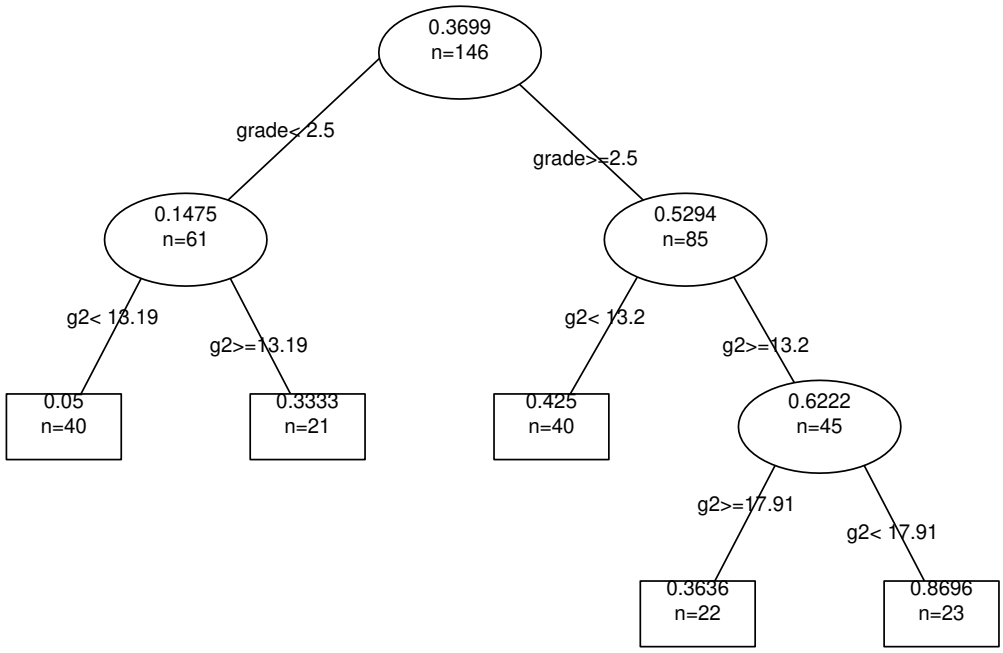


Figure 16: Fancier plot

"tuck under" those above. If you want to play around with the spacing with compress, try using `nspace` which regulates the space between a terminal node and a split.

```
> par(mar = rep(0.2, 4))
> plot(fit, branch = 0.4,uniform = TRUE, compress = TRUE)
> text(fit, all = TRUE, use.n = TRUE)
```

Several of these options were combined into a function called `post.rpart`, whose default action of creating a .ps file in the current directory is now rather dated. Identical results to the function can be obtained by the collection of options shown below, the result is displayed in figure 16. The code is essentially

```
> par(mar = rep(0.2, 4))
> plot(fit, uniform = TRUE, branch = 0.2, compress = TRUE, margin = 0.1)
> text(fit, all = TRUE, use.n = TRUE, fancy = TRUE, cex= 0.9)
```

The `fancy` option of `text` creates the ellipses and rectangles, and moves the splitting rule to the midpoints of the branches. `Margin` shrinks the plotting region slightly so that the

`text` boxes don't run over the edge of the plot. The `branch` option makes the lines exit the ellipse at a "good" angle. A separate package `rpart.plot` carries these ideas much further.

10 Other functions

A more general approach to cross-validation can be gained using the `xpred.rpart` function. Given an rpart fit, a vector of k complexity parameters, and the desired number of cross-validation groups, this function returns an n by k matrix containing the predicted value $\hat{y}_{(-i)}$ for each subject, from the model that was fit without that subject. The *cp* vector defaults to the geometric mean of the *cp* sequence for the pruned tree on the full data set.

Here is an example that uses the mean absolute deviation for continuous data rather than the usual mean square error.

```
> carfit <- rpart(Price/1000 ~ ., cars)
> carfit$cptable
          CP nsplit rel error    xerror      xstd
1 0.46014608      0 1.0000000 1.0245300 0.1630984
2 0.11790530      1 0.5398539 0.7169324 0.1120680
3 0.11234341      2 0.4219486 0.6667297 0.1060676
4 0.04449133      3 0.3096052 0.6403983 0.1155832
5 0.03344936      4 0.2651139 0.5617111 0.1141092
6 0.01000000      5 0.2316645 0.5846245 0.1318216
> price2 <-  cars$Price[!is.na(cars$Price)]/1000
> temp <- xpred.rpart(carfit)
> errmat <- price2 - temp
> abserr <- colMeans(abs(errmat))
> rbind(abserr, relative=abserr/mean(abs(price2-mean(price2))))
         0.73007304 0.23292416 0.11509076 0.07069871
abserr     6.570864  5.1601190  5.2037303  4.2955745
relative   1.003264  0.7878665  0.7945253  0.6558646
         0.03857728 0.01828917
abserr    4.0164478  3.6572999
relative  0.6132465  0.5584104
```

We see that on the absolute error scale the relative error improvement is not quite so large, though this could be expected given that the optimal split was not chosen to minimize absolute error.

y	1	2	3	1	2	3	1	2
x1	1	2	3	4	5	6	7	8
x2	1	2	3	4	5	6	1	2
x3	NA	22	38	12	NA	48	14	32

y	3	1	2	3	1	2	1
x1	9	10	11	12	13	14	15
x2	3	4	5	6	1	2	3
x3	40	NA	30	46	28	34	48

Table 1: Data set for the classification test case

11 Test Cases

11.1 Classification

The definitions for classification trees can get the most complex, especially with respect to priors and loss matrices. In this section we lay out a simple example, in great detail. (This was done to debug the R functions.)

Let $n = 15$, and the data be as given in table 1. The loss matrix is defined as

$$L = \begin{matrix} 0 & 2 & 2 \\ 2 & 0 & 6 \\ 1 & 1 & 0 \end{matrix}$$

where rows represent the true class and columns the assigned class. Thus the error in mistakenly calling a class 2 observation a 3 is quite large in this data set. The prior probabilities for the study are assumed to be $\pi = .2, .3, .5$, so class 1 is most prevalent in the input data (n_i=6, 5, and 4 observations, respectively), but class 3 the most prevalent in the external population.

Splits are chosen using the Gini index with altered priors, as defined in equation (4.15) of Breiman et al [1].

$$\begin{aligned} \tilde{\pi}_1 &= \pi_1(0+2+2)/\sum \tilde{\pi}_i = 4/21 \\ \tilde{\pi}_2 &= \pi_2(2+0+6)/\sum \tilde{\pi}_i = 12/21 \\ \tilde{\pi}_3 &= \pi_3(1+1+0)/\sum \tilde{\pi}_i = 5/21 \end{aligned}$$

For a given node T, the Gini impurity will be $\sum_j p(j|T)[1 - p(j|T)] = 1 - \sum p(j|T)^2$, where $p(j|T)$ is the expected proportion of class j in the node:

$$p(j|T) = \tilde{\pi}_j[n_i(T)/n_i]/\sum p(i|T)$$

.

Starting with the top node, for each possible predicted class we have the following loss

predicted class	E(loss)
1	.2*0 + .3*2 + .5*1 = 1.1
2	.2*2 + .3*0 + .5*1 = 0.9
3	.2*2 + .3*6 + .5*0 = 2.2

The best choice is class 2, with an expected loss of 0.9. The Gini impurity for the node, using altered priors, is $G = 1 - (16 + 144 + 25)/21^2 = 256/441 \approx .5805$.

Assume that variable $x1$ is split at 12.5, which is, as it turns out, the optimal split point for that variable under the constraint that at least 2 observations are in each terminal node. Then the right node will have class counts of (4,4,4) and the left node of (2,1,0). For the right node (node 3 in the tree)

$$\begin{aligned}
P(R) &= .2(4/6) + .3(4/5) + .5(4/4) = 131/150 \\
&= \text{probability of the node (in the population)} \\
p(i|R) &= (.2(4/6), .3(4/5), .5(4/4))/P(R) \\
\widetilde{P}(R) &= (4/21)(4/6) + (12/21)(4/5) + (5/21)(4/4) = 259/315 = 37/45 \\
\tilde{p}(i|R) &= [(4/21)(4/6), (12/21)(4/5), (5/21)(4/4)]\ (45/37) \\
G(R) &= 1 - \textstyle\sum \tilde{p}(i|R)^2 \approx .5832
\end{aligned}$$

For the left node (node 2)

$$\begin{aligned}
P(L) &= .2(2/6) + .3(1/5) + .5(0/4) = 19/150 \\
p(i|L) &= (.4/3, .3/5, 0)/\ P(L) = (10/19, 9/19, 0) \\
\widetilde{P}(L) &= 1 - \widetilde{P}(R) = 8/45 \\
\tilde{p}(i|L) &= [(4/21)(2/6), (12/21)(1/5), (5/21)(0/4)]/\ \widetilde{P}(L) \\
G(L) &= 1 - \textstyle\sum \tilde{p}(i|L)^2 \approx .459
\end{aligned}$$

The total improvement for the split involves the change in impurity between the parent and the two child nodes

$$n(G - [\widetilde{P}(L) * G(L) + \widetilde{P}(R) * G(R)]) \approx .2905$$

where $n = 15$ is the total sample size.

For variable $x2$ the best split occurs at 5.5, splitting the data into left and right nodes with class counts of (2,2,1) and (4,3,3), respectively. Computations just exactly like the above give an improvement of 1.912.

For variable $x3$ there are 3 missing values, and the computations are similar to what would be found for a node further down the tree with only 12/15 observations. The best split point is at 3.6, giving class counts of (3,4,0) and (1,0,4) in the left and right nodes, respectively.

Cutpoint	$P(L)$	$P(R)$	$G(L)$	$G(R)$	ΔI
1.3	0.03	0.97	0.00	0.56	0.55
1.8	0.06	0.94	0.00	0.53	1.14
2.5	0.18	0.82	0.46	0.57	0.45
2.9	0.21	0.79	0.50	0.53	0.73
3.1	0.32	0.68	0.42	0.56	1.01
3.3	0.44	0.56	0.34	0.52	1.96
3.6	0.55	0.45	0.29	0.21	3.99
3.9	0.61	0.39	0.41	0.26	2.64
4.3	0.67	0.33	0.48	0.33	1.56
4.7	0.73	0.27	0.53	0.45	0.74
4.8	0.79	0.21	0.56	0.00	0.55

Table 2: Cut points and statistics for variable $x3$, top node

For the right node (node 3 in the tree)

$$\begin{aligned}
\widetilde{P}(R) &= (4/21)(3/6) + (12/21)(4/5) + (5/21)(0/4) = 174/315 \\
\tilde{p}(i|R) &= [(4/21)(3/6), (12/21)(4/5), (5/21)(0/4)]\ (315/174) \\
&= (5/29, 24/29, 0) \\
G(R) &= 1 - (25 + 576)/29^2 = 240/841 approx .2854
\end{aligned}$$

For the left node (node 2)

$$\begin{aligned}
\widetilde{P}(L) &= (4/21)(1/6) + (12/21)(0/5) + (5/21)(4/4) = 85/315 \\
\tilde{p}(i|L) &= [(4/21)(1/6), (12/21)(0/5), (5/21)(4/4)]\ (315/85) \\
&= (2/17, 0, 15/17) \\
G(L) &= 1 - (4 + 225)/17^2 = 60/289 \approx .2076
\end{aligned}$$

The overall impurity for the node involves only 12 of the 15 observations, giving the following values for the top node:

$$\begin{aligned}
\widetilde{P}(T) &= 174/315 + 85/315 = 259/315 \\
\tilde{p}(i|T) &= [(4/21)(4/6), (12/21)(4/5), (5/21)(4/4)]\ (315/259) \\
&= (40/259, 144/259, 75/259) \\
G(T) &= 1 - (40^2 + 144^2 + 75^2)/259^2 = 39120/67081
\end{aligned}$$

The total improvement for the split involves the impurity G of all three nodes, weighted by the probabilities of the nodes under the alternate priors.

$$15 * \{(259/315)(39120/67081) - [(174/315)(240/841) + (85/315)(60/289)]\} \approx 3.9876$$

As is true in general with the rpart routines, variables with missing values are penalized with respect to choosing a split – a factor of 259/315 or about 12/15 in the case of $x3$ at the top node. Table 2 shows the statistics for all of the cutpoints of $x3$.

Because $x3$ has missing values, the next step is choice of a surrogate split. Priors and losses currently play no role in the computations for selecting surrogates. For all

the prior computations, the effect of priors is identical to that of adding case weights to the observations such that the apparent frequencies are equal to the chosen priors; since surrogate computations do account for case weights, one could argue that they should also then make use of priors. The argument has not yet been found compelling enough to add this to the code.

Note to me: the cp is not correct for usesurrogate=0. The error after a split is not (left error + right error) – it also needs to have a term (parent error for those obs that weren't split).

References

[1] L. Breiman, J.H. Friedman, R.A. Olshen, , and C.J Stone. *Classification and Regression Trees*. Wadsworth, Belmont, Ca, 1983.

[2] L.A. Clark and D. Pregibon. Tree-based models. In J.M. Chambers and T.J. Hastie, editors, *Statistical Models in S*, chapter 9. Wadsworth and Brooks/Cole, Pacific Grove, Ca, 1992.

[3] M. LeBlanc and J Crowley. Relative risk trees for censored survival data. *Biometrics*, 48:411–425, 1992.

[4] O. Nativ, Y. Raz, H.Z. Winkler, Y. Hosaka, E.T. Boyle, T.M. Therneau, G.M. Farrow, R.P. Meyers, H. Zincke, and M.M Lieber. Prognostic value of flow cytometric nuclear DNA analysis in stage C prostate carcinoma. *Surgical Forum*, pages 685–687, 1988.

[5] T.M. Therneau. A short introduction to recursive partitioning. Orion Technical Report 21, Stanford University, Department of Statistics, 1983.

[6] T.M Therneau and E.J Atkinson. An introduction to recursive partitioning using the rpart routines. Divsion of Biostatistics 61, Mayo Clinic, 1997.

[7] T.M. Therneau, Grambsch P.M., and T.R. Fleming. Martingale based residuals for survival models. *Biometrika*, 77:147–160, 1990.