Classification and Regression Trees (CART)

ACM

October 9th, 2023

Regression Trees

```
library(printr)
```

```
Registered S3 method overwritten by 'printr':
  method                from
  knit_print.data.frame rmarkdown
```

```
library(rpart)
bball <- read.csv("baseball.csv")
head(bball)
```

Salary	BatAvg	OBP	Runs	Hits	Doubles	Triples	HR	RBI	Walks	SO	SB	Err	FA	Prior_FA	Arb	Prior_Arb
3300	0.272	0.302	69	153	21	4	31	104	22	80	4	3	1	0	0	0
2600	0.269	0.335	58	111	17	2	18	66	39	69	0	3	1	1	0	0
2500	0.249	0.337	54	115	15	1	17	73	63	116	6	5	1	0	0	0
2475	0.260	0.292	59	128	22	7	12	50	23	64	21	21	0	0	1	0
2313	0.273	0.346	87	169	28	5	8	58	70	53	3	8	0	0	1	0
2175	0.291	0.379	104	170	32	2	26	100	87	89	22	4	1	0	0	0

```
`?`(rpart)
```

```
Recursive Partitioning and Regression Trees

Description:

     Fit a 'rpart' model

Usage:

     rpart(formula, data, weights, subset, na.action = na.rpart, method,
           model = FALSE, x = FALSE, y = TRUE, parms, control, cost, ...)

Arguments:

 formula: a formula, with a response but no interaction terms.  If this
          is a data frame, it is taken as the model frame (see
          'model.frame).'

    data: an optional data frame in which to interpret the variables
          named in the formula.
```

weights: optional case weights.

subset: optional expression saying that only a subset of the rows of
the data should be used in the fit.

na.action: the default action deletes all observations for which 'y' is
missing, but keeps those in which one or more predictors are
missing.

method: one of '"anova"', '"poisson"', '"class"' or '"exp"'. If
'method' is missing then the routine tries to make an
intelligent guess. If 'y' is a survival object, then 'method
= "exp"' is assumed, if 'y' has 2 columns then 'method =
"poisson"' is assumed, if 'y' is a factor then 'method =
"class"' is assumed, otherwise 'method = "anova"' is assumed.
It is wisest to specify the method directly, especially as
more criteria may added to the function in future.

Alternatively, 'method' can be a list of functions named
'init', 'split' and 'eval'. Examples are given in the file
'tests/usersplits.R' in the sources, and in the vignettes
'User Written Split Functions'.

model: if logical: keep a copy of the model frame in the result? If
the input value for 'model' is a model frame (likely from an
earlier call to the 'rpart' function), then this frame is
used rather than constructing new data.

x: keep a copy of the 'x' matrix in the result.

y: keep a copy of the dependent variable in the result. If
missing and 'model' is supplied this defaults to 'FALSE'.

parms: optional parameters for the splitting function.
Anova splitting has no parameters.
Poisson splitting has a single parameter, the coefficient of
variation of the prior distribution on the rates. The
default value is 1.
Exponential splitting has the same parameter as Poisson.
For classification splitting, 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'. The default priors are proportional
to the data counts, the losses default to 1, and the split
defaults to 'gini'.

control: a list of options that control details of the 'rpart'
algorithm. See 'rpart.control'.

cost: a vector of non-negative costs, one for each variable in the

```
          model. Defaults to one for all variables.  These are scalings
          to be applied when considering splits, so the improvement on
          splitting on a variable is divided by its cost in deciding
          which split to choose.

     ...: arguments to 'rpart.control' may also be specified in the
          call to 'rpart'.  They are checked against the list of valid
          arguments.
```

```
fit <- rpart(Salary ~ ., data = bball, method = "anova")
summary(fit)
```

```
Call:
rpart(formula = Salary ~ ., data = bball, method = "anova")
  n= 337

           CP nsplit rel error    xerror       xstd
1  0.33946196      0 1.0000000 1.0114105 0.09002609
2  0.14117583      1 0.6605380 0.8480921 0.07994572
3  0.08888501      2 0.5193622 0.6747731 0.06803844
4  0.05133621      3 0.4304772 0.5262333 0.05866293
5  0.04766668      4 0.3791410 0.5085223 0.05841257
6  0.03075163      5 0.3314743 0.4371061 0.05299548
7  0.02166442      6 0.3007227 0.4188152 0.05273598
8  0.02095516      7 0.2790583 0.4229586 0.05434003
9  0.01992146      8 0.2581031 0.4166549 0.05383359
10 0.01520262      9 0.2381816 0.3698494 0.04961054
11 0.01000000     10 0.2229790 0.3576907 0.04651626

Variable importance
    Runs      RBI     Hits  Doubles       SO    Walks       FA      Arb
      16       15       13       11       10        9        8        8
      HR       SB      Err   BatAvg Prior_FA      OBP
       4        2        1        1        1        1

Node number 1: 337 observations,    complexity param=0.339462
  mean=1248.528, MSE=1533070
  left son=2 (188 obs) right son=3 (149 obs)
  Primary splits:
      Runs < 46.5   to the left,  improve=0.3394620, (0 missing)
      RBI  < 64.5   to the left,  improve=0.3261317, (0 missing)
      FA   < 0.5    to the left,  improve=0.3187570, (0 missing)
      Hits < 94.5   to the left,  improve=0.3133598, (0 missing)
      HR   < 15.5   to the left,  improve=0.2858796, (0 missing)
  Surrogate splits:
      Hits    < 104.5  to the left,  agree=0.917, adj=0.812, (0 split)
      RBI     < 45.5   to the left,  agree=0.899, adj=0.772, (0 split)
      Doubles < 18.5   to the left,  agree=0.846, adj=0.651, (0 split)
      Walks   < 40.5   to the left,  agree=0.822, adj=0.597, (0 split)
      SO      < 57.5   to the left,  agree=0.801, adj=0.550, (0 split)

Node number 2: 188 observations,    complexity param=0.05133621
  mean=606.2979, MSE=450254.1
  left son=4 (135 obs) right son=5 (53 obs)
  Primary splits:
```

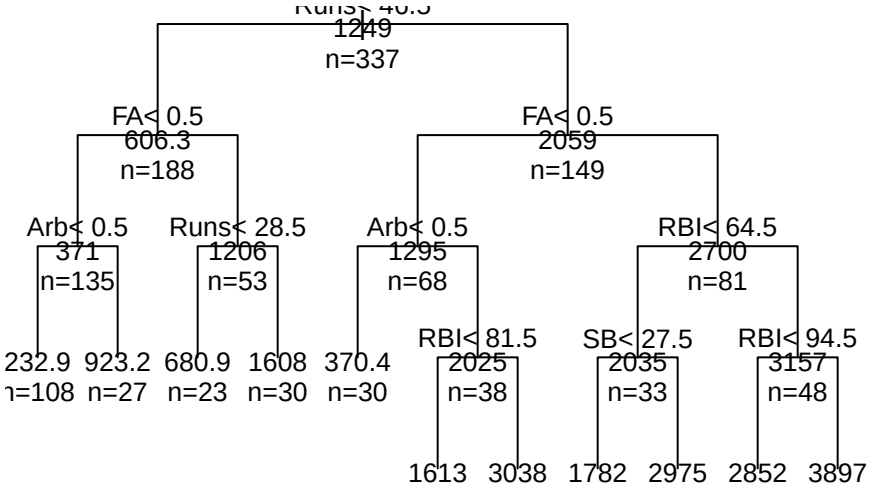
```
      FA    < 0.5    to the left,  improve=0.3133287, (0 missing)
      Hits  < 62.5   to the left,  improve=0.1875688, (0 missing)
      RBI   < 25.5   to the left,  improve=0.1744576, (0 missing)
      Runs  < 28.5   to the left,  improve=0.1609022, (0 missing)
      Walks < 32.5   to the left,  improve=0.1468637, (0 missing)
  Surrogate splits:
      Prior_FA < 0.5    to the left,  agree=0.840, adj=0.434, (0 split)
      Walks    < 46     to the left,  agree=0.750, adj=0.113, (0 split)
      SB       < 16.5   to the left,  agree=0.734, adj=0.057, (0 split)
      RBI      < 69.5   to the left,  agree=0.729, adj=0.038, (0 split)
      SO       < 75.5   to the left,  agree=0.729, adj=0.038, (0 split)

Node number 3: 149 observations,    complexity param=0.1411758
  mean=2058.859, MSE=1722253
  left son=6 (68 obs) right son=7 (81 obs)
  Primary splits:
      FA      < 0.5    to the left,  improve=0.28422950, (0 missing)
      RBI     < 82.5   to the left,  improve=0.17440740, (0 missing)
      Runs    < 83.5   to the left,  improve=0.15926140, (0 missing)
      HR      < 16.5   to the left,  improve=0.15236050, (0 missing)
      Doubles < 31.5   to the left,  improve=0.07377436, (0 missing)
  Surrogate splits:
      Arb   < 0.5    to the right, agree=0.799, adj=0.559, (0 split)
      HR    < 14.5   to the left,  agree=0.624, adj=0.176, (0 split)
      Err   < 8.5    to the right, agree=0.624, adj=0.176, (0 split)
      Walks < 30.5   to the left,  agree=0.604, adj=0.132, (0 split)
      SB    < 11.5   to the right, agree=0.604, adj=0.132, (0 split)

Node number 4: 135 observations,    complexity param=0.01992146
  mean=370.9556, MSE=139364.4
  left son=8 (108 obs) right son=9 (27 obs)
  Primary splits:
      Arb     < 0.5    to the left,  improve=0.5470509, (0 missing)
      Doubles < 7.5    to the left,  improve=0.1443754, (0 missing)
      Hits    < 101.5  to the left,  improve=0.1420688, (0 missing)
      RBI     < 11.5   to the left,  improve=0.1277459, (0 missing)
      Runs    < 21.5   to the left,  improve=0.1268071, (0 missing)
  Surrogate splits:
      Hits      < 101.5  to the left,  agree=0.822, adj=0.111, (0 split)
      Prior_Arb < 0.5    to the left,  agree=0.822, adj=0.111, (0 split)
      Doubles   < 23.5   to the left,  agree=0.807, adj=0.037, (0 split)

Node number 5: 53 observations,    complexity param=0.02166442
  mean=1205.755, MSE=741717.2
  left son=10 (23 obs) right son=11 (30 obs)
  Primary splits:
      Runs  < 28.5   to the left,  improve=0.2847245, (0 missing)
      Hits  < 62.5   to the left,  improve=0.2106064, (0 missing)
      HR    < 7.5    to the left,  improve=0.2092991, (0 missing)
      RBI   < 25.5   to the left,  improve=0.1954400, (0 missing)
      Walks < 35.5   to the left,  improve=0.1438194, (0 missing)
  Surrogate splits:
      Hits    < 57.5   to the left,  agree=0.811, adj=0.565, (0 split)
      HR      < 3.5    to the left,  agree=0.811, adj=0.565, (0 split)
```

```
      RBI     < 30     to the left,  agree=0.811, adj=0.565, (0 split)
      SO      < 41     to the left,  agree=0.774, adj=0.478, (0 split)
      Doubles < 6.5    to the left,  agree=0.755, adj=0.435, (0 split)

Node number 6: 68 observations,    complexity param=0.08888501
  mean=1295.25, MSE=1317909
  left son=12 (30 obs) right son=13 (38 obs)
  Primary splits:
      Arb  < 0.5    to the left,  improve=0.5124201, (0 missing)
      Runs < 86.5   to the left,  improve=0.2205733, (0 missing)
      RBI  < 82.5   to the left,  improve=0.1676543, (0 missing)
      Hits < 174    to the left,  improve=0.1625680, (0 missing)
      HR   < 24     to the left,  improve=0.1563424, (0 missing)
  Surrogate splits:
      SO      < 87.5   to the right, agree=0.676, adj=0.267, (0 split)
      BatAvg  < 0.2575 to the left,  agree=0.618, adj=0.133, (0 split)
      Hits    < 123.5  to the left,  agree=0.618, adj=0.133, (0 split)
      Doubles < 15.5   to the left,  agree=0.618, adj=0.133, (0 split)
      OBP     < 0.359  to the right, agree=0.603, adj=0.100, (0 split)

Node number 7: 81 observations,    complexity param=0.04766668
  mean=2699.914, MSE=1161236
  left son=14 (33 obs) right son=15 (48 obs)
  Primary splits:
      RBI     < 64.5   to the left,  improve=0.2618190, (0 missing)
      Runs    < 76.5   to the left,  improve=0.1873776, (0 missing)
      HR      < 17.5   to the left,  improve=0.1538262, (0 missing)
      Doubles < 31     to the left,  improve=0.1229190, (0 missing)
      Hits    < 162    to the left,  improve=0.1194826, (0 missing)
  Surrogate splits:
      HR      < 12.5   to the left,  agree=0.852, adj=0.636, (0 split)
      SO      < 41     to the left,  agree=0.728, adj=0.333, (0 split)
      Hits    < 129.5  to the left,  agree=0.716, adj=0.303, (0 split)
      Doubles < 16.5   to the left,  agree=0.716, adj=0.303, (0 split)
      SB      < 10.5   to the right, agree=0.691, adj=0.242, (0 split)

Node number 8: 108 observations
  mean=232.8981, MSE=25438.92

Node number 9: 27 observations
  mean=923.1852, MSE=213869.1

Node number 10: 23 observations
  mean=680.913, MSE=133118.1

Node number 11: 30 observations
  mean=1608.133, MSE=835216.2

Node number 12: 30 observations
  mean=370.3667, MSE=22682.43

Node number 13: 38 observations,    complexity param=0.03075163
  mean=2025.421, MSE=1131983
  left son=26 (27 obs) right son=27 (11 obs)
```

```
  Primary splits:
      RBI     < 81.5   to the left,  improve=0.3693487, (0 missing)
      HR      < 24     to the left,  improve=0.3339221, (0 missing)
      BatAvg  < 0.287  to the left,  improve=0.2554011, (0 missing)
      Doubles < 34.5   to the left,  improve=0.2105206, (0 missing)
      Walks   < 55.5   to the left,  improve=0.1941625, (0 missing)
  Surrogate splits:
      HR      < 21.5   to the left,  agree=0.921, adj=0.727, (0 split)
      BatAvg  < 0.298  to the left,  agree=0.816, adj=0.364, (0 split)
      Doubles < 34.5   to the left,  agree=0.816, adj=0.364, (0 split)
      OBP     < 0.365  to the left,  agree=0.789, adj=0.273, (0 split)
      Runs    < 89.5   to the left,  agree=0.789, adj=0.273, (0 split)

Node number 14: 33 observations,    complexity param=0.01520262
  mean=2034.909, MSE=804609.9
  left son=28 (26 obs) right son=29 (7 obs)
  Primary splits:
      SB       < 27.5   to the left,  improve=0.2958087, (0 missing)
      Runs     < 83.5   to the left,  improve=0.2882175, (0 missing)
      Prior_FA < 0.5    to the right, improve=0.2079851, (0 missing)
      Err      < 10.5   to the right, improve=0.1738314, (0 missing)
      SO       < 82.5   to the right, improve=0.1187657, (0 missing)
  Surrogate splits:
      Runs  < 79     to the left,  agree=0.939, adj=0.714, (0 split)
      Hits  < 162    to the left,  agree=0.879, adj=0.429, (0 split)
      Walks < 92.5   to the left,  agree=0.848, adj=0.286, (0 split)

Node number 15: 48 observations,    complexity param=0.02095516
  mean=3157.104, MSE=893360.1
  left son=30 (34 obs) right son=31 (14 obs)
  Primary splits:
      RBI    < 94.5   to the left,  improve=0.25247320, (0 missing)
      SO     < 120.5  to the left,  improve=0.15891980, (0 missing)
      BatAvg < 0.3005 to the left,  improve=0.10163010, (0 missing)
      Runs   < 77     to the left,  improve=0.09583728, (0 missing)
      Hits   < 149.5  to the left,  improve=0.09090962, (0 missing)
  Surrogate splits:
      HR      < 28.5   to the left,  agree=0.854, adj=0.500, (0 split)
      Runs    < 96.5   to the left,  agree=0.792, adj=0.286, (0 split)
      SO      < 118.5  to the left,  agree=0.792, adj=0.286, (0 split)
      Doubles < 33.5   to the left,  agree=0.771, adj=0.214, (0 split)
      Hits    < 145    to the left,  agree=0.750, adj=0.143, (0 split)

Node number 26: 27 observations
  mean=1612.704, MSE=389480.5

Node number 27: 11 observations
  mean=3038.455, MSE=1510157

Node number 28: 26 observations
  mean=1781.769, MSE=635429.7

Node number 29: 7 observations
  mean=2975.143, MSE=310943.3
```

```
Node number 30: 34 observations
  mean=2852.353, MSE=450014.5

Node number 31: 14 observations
  mean=3897.214, MSE=1196744
```

```
par(mfrow=c(1,1),mar=rep(5,4))
plot(fit, uniform=TRUE,
    main="Regression Tree for Baseball Salary Data")
text(fit, use.n=TRUE, all=TRUE, cex=.8)
```




```
cp_vals <- fit$cptable
printcp(fit)
```

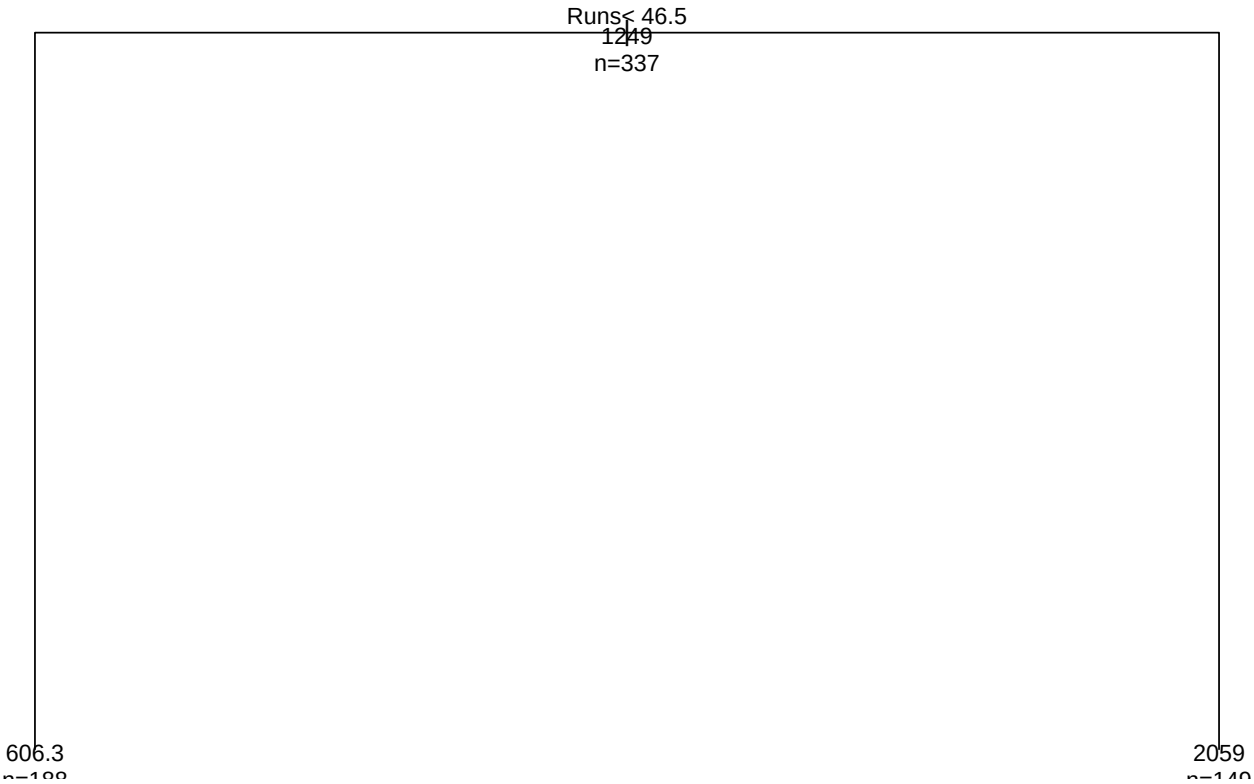
```
##
## Regression tree:
## rpart(formula = Salary ~ ., data = bball, method = "anova")
##
## Variables actually used in tree construction:
## [1] Arb  FA   RBI  Runs SB
##
## Root node error: 516644690/337 = 1533070
##
## n= 337
##
##          CP nsplit rel error  xerror     xstd
## 1  0.339462      0   1.00000 1.01141 0.090026
## 2  0.141176      1   0.66054 0.84809 0.079946
## 3  0.088885      2   0.51936 0.67477 0.068038
## 4  0.051336      3   0.43048 0.52623 0.058663
## 5  0.047667      4   0.37914 0.50852 0.058413
## 6  0.030752      5   0.33147 0.43711 0.052995
## 7  0.021664      6   0.30072 0.41882 0.052736
## 8  0.020955      7   0.27906 0.42296 0.054340
## 9  0.019921      8   0.25810 0.41665 0.053834
```

```
## 10 0.015203      9   0.23818 0.36985 0.049611
## 11 0.010000     10   0.22298 0.35769 0.046516
```

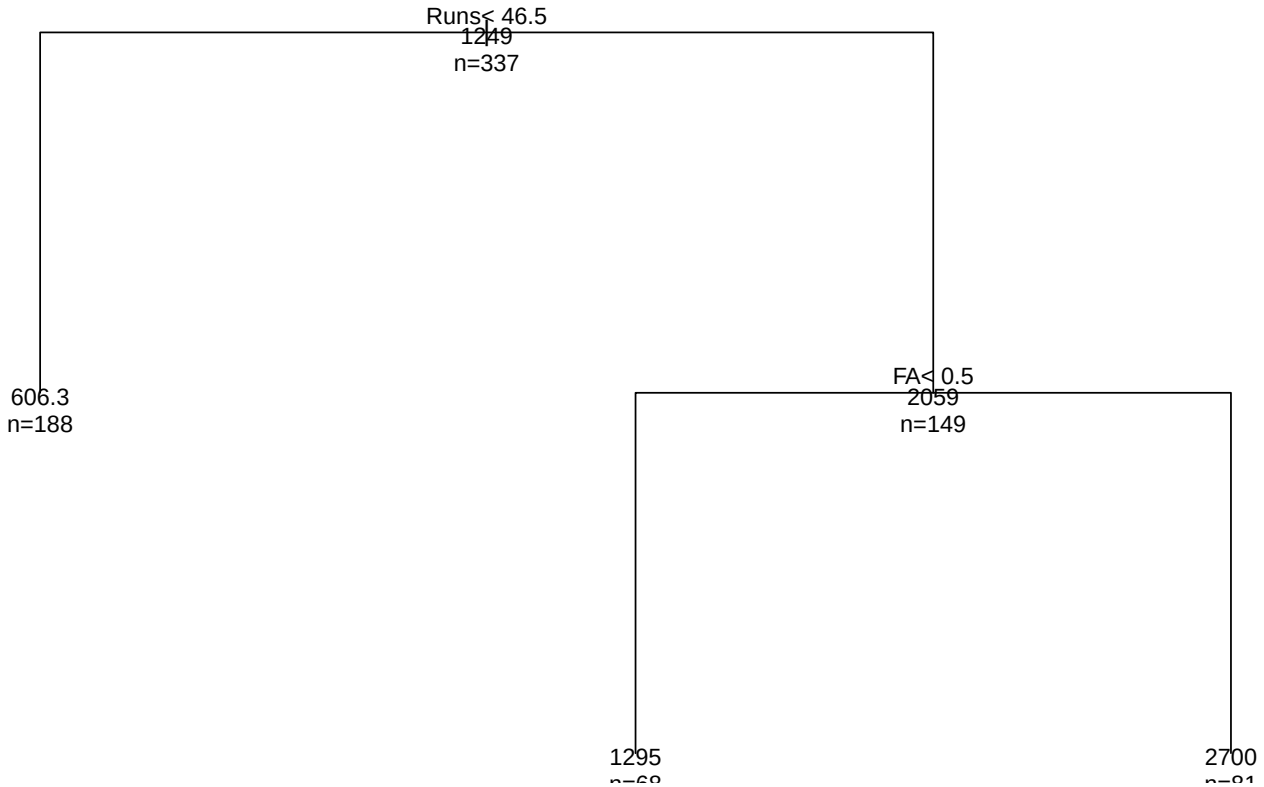
- 'CP' is the value of the complexity parameter divided by the resubstitution estimate $C(T_{root}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ where $\hat{y}_i$ is estimated from the root tree denoted by T_{root}.

- The 'rel error' is $C(T)/C(T_{root})$ where T varies by row.
- The 'xerror' column is the CV error divided by $C(T_{root})$.

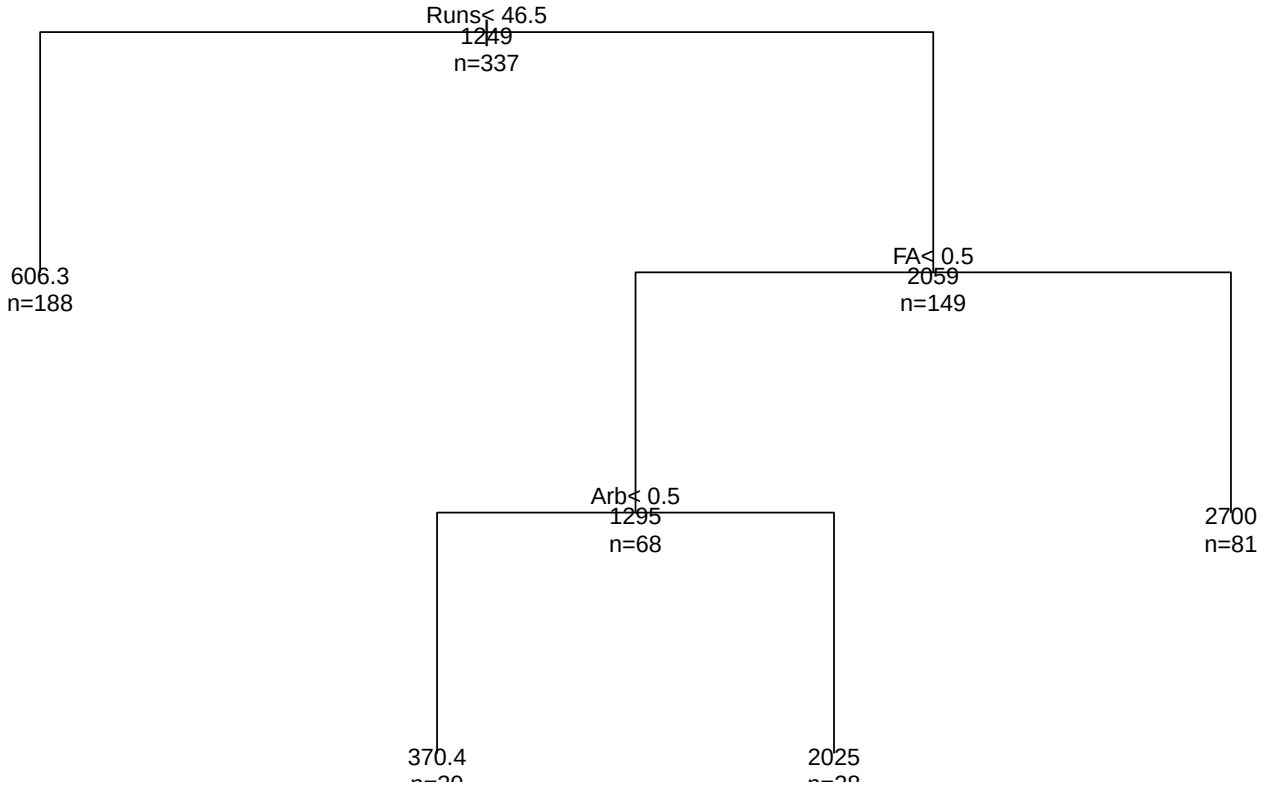
```
par(mfrow=c(1,1),mar=rep(6,4))
plot(prune(fit,cp=cp_vals[2,1]), uniform=TRUE)
text(prune(fit,cp=cp_vals[2,1]), use.n=TRUE, all=TRUE, cex=.8)
```



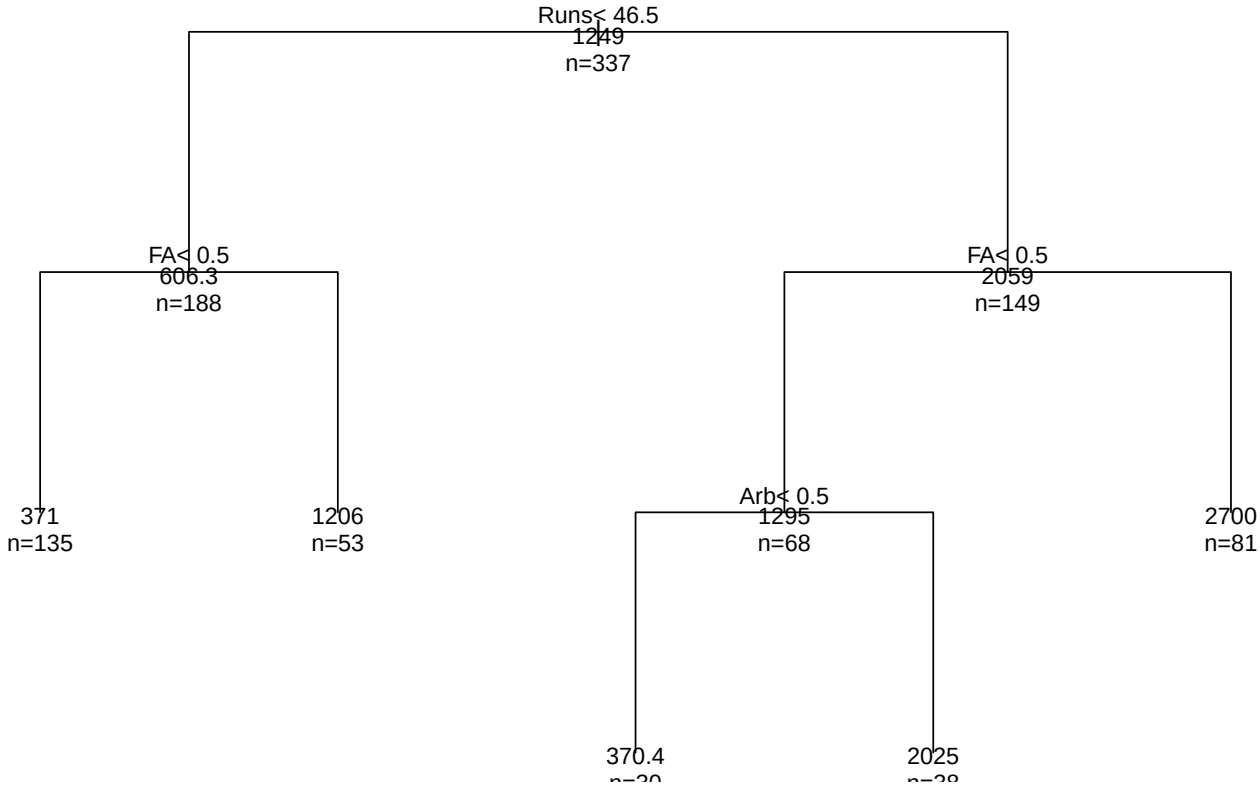

```
plot(prune(fit,cp=cp_vals[3,1]), uniform=TRUE)
text(prune(fit,cp=cp_vals[3,1]), use.n=TRUE, all=TRUE, cex=.8)
```




```
plot(prune(fit,cp=cp_vals[4,1]), uniform=TRUE)
text(prune(fit,cp=cp_vals[4,1]), use.n=TRUE, all=TRUE, cex=.8)
```



```
plot(prune(fit,cp=cp_vals[5,1]), uniform=TRUE)
text(prune(fit,cp=cp_vals[5,1]), use.n=TRUE, all=TRUE, cex=.8)
```



Classification Trees

In this example we're going to look at the Cleveland heart study:

```
Cle_heart <- read.csv("Cle_heart.csv")
attach(Cle_heart)
head(Cle_heart[,1:12])
```

age	gender	cp	trestbps	chol	fbs	restecg	thatach	exang	oldpeak	slope	ca
63	male	angina	145	233	TRUE	hyp	150	fal	2.3	down	0
67	male	asympt	160	286	fal	hyp	108	TRUE	1.5	flat	3
67	male	asympt	120	229	fal	hyp	129	TRUE	2.6	flat	2
37	male	notang	130	250	fal	norm	187	fal	3.5	down	0
41	fem	abnang	130	204	fal	hyp	172	fal	1.4	up	0
56	male	abnang	120	236	fal	norm	178	fal	0.8	up	0

```
head(Cle_heart[,13:(dim(Cle_heart)[2])])
```

thal	diag	Col15
fix	buff	H
norm	sick	S2
rev	sick	S1
norm	buff	H

thal	diag	Col15
norm	buff	H
norm	buff	H

```
sort(unique(age))
```

```
##  [1] 29 34 35 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58
## [26] 59 60 61 62 63 64 65 66 67 68 69 70 71 74 76 77
```

```
tbl <- table(age>65,diag)
tbl
```

/diag	buff	sick
FALSE	143	120
TRUE	17	16

```
prop.table(tbl)
```

/diag	buff	sick
FALSE	0.4831081	0.4054054
TRUE	0.0574324	0.0540541

```
table(cp)
```

abnang	angina	asympt	notang
49	23	141	83

```
table(thal)
```

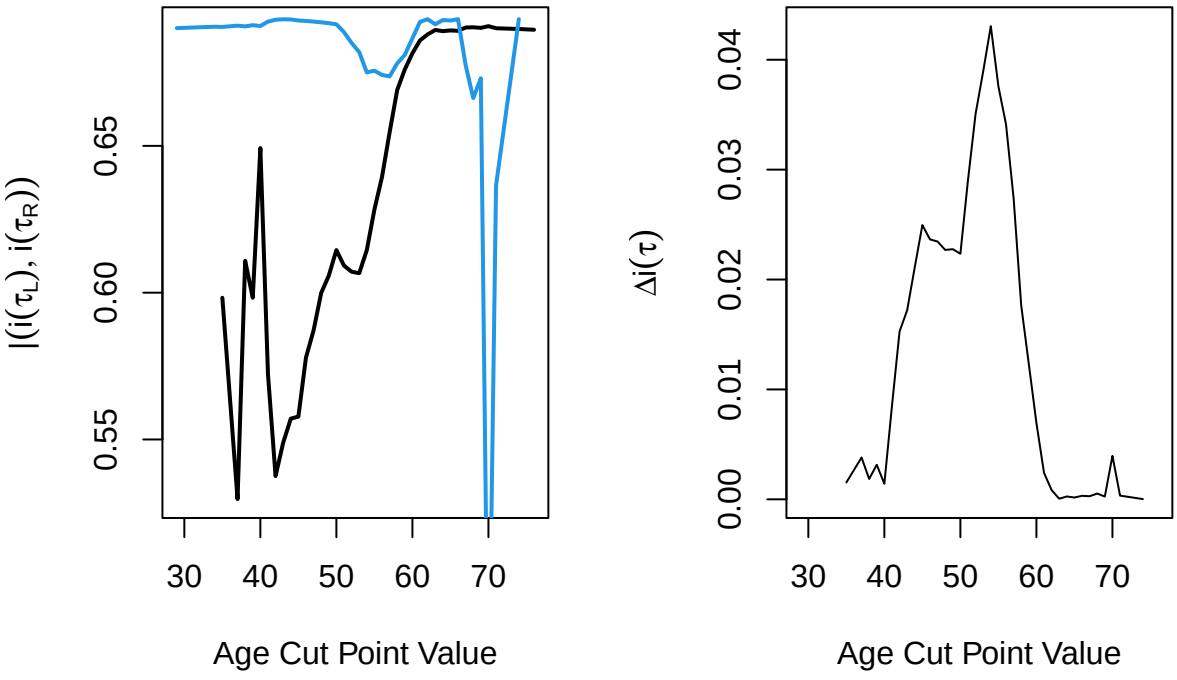
fix	norm	rev
18	163	115

First, we'll write a function to estimate the impurity:

```
impur_func <- function(cov,i,cond){
  tbl <- table(cov>i,cond)
  n <- sum(tbl)
  n1p <- sum(tbl[1,])
  p_tau_l <- tbl[1,1]/n1p
  q_tau_l <- 1-p_tau_l
  i_tau_l <- -p_tau_l*log(p_tau_l) - q_tau_l*log(q_tau_l)
  n2p <- sum(tbl[2,])
  p_tau_r <- tbl[2,1]/n2p
  q_tau_r <- 1-p_tau_r
  i_tau_r <- -p_tau_r*log(p_tau_r) - q_tau_r*log(q_tau_r)
  p <- sum(tbl[,1])/n
  q <- 1-p
```

```
  i_tau <- -p*log(p) - q*log(q)
  p_l <- n1p/n
  p_r <- n2p/n
  delta_i <- i_tau - p_l*i_tau_l - p_r*i_tau_r
  return(list(delta_i=delta_i,i_tau_r=i_tau_r,i_tau_l=i_tau_l))
  }
```

```
impur <- NULL
vals <- sort(unique(age))
vals <- vals[-which.max(vals)]
for(i in vals){
  t_impur <- impur_func(age,i,diag)
  impur <- rbind(impur,c(t_impur$delta_i,t_impur$i_tau_l,t_impur$i_tau_r))
}
par(mfrow=c(1,2))
plot(vals,impur[,2],xlab="Age Cut Point Value",
     ylab=expression(i(tau[L])|i(tau[R])),type="l",lwd=2)
lines(vals,impur[,3],lwd=2,col=4)
plot(vals,impur[,1],type="l",xlab="Age Cut Point Value",
     ylab=expression(Delta * i(tau)))
```



Now we'll estimate with `rpart`:

```
library(rpart)
Cle_heart2 <- Cle_heart[,-15]
fit <- rpart(diag~.,data=Cle_heart2,method="class")
summary(fit)
```

```
## Call:
## rpart(formula = diag ~ ., data = Cle_heart2, method = "class")
##   n= 296
##
##           CP nsplit rel error    xerror       xstd
```

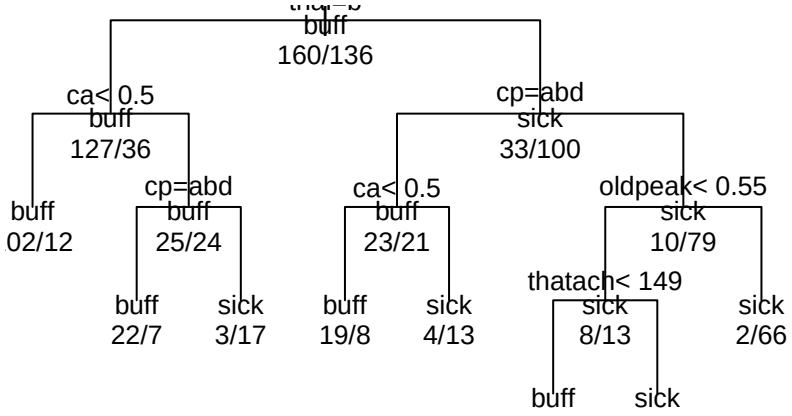
```
## 1 0.49264706      0 1.0000000 1.0000000 0.06304413
## 2 0.05147059      1 0.5073529 0.5735294 0.05573022
## 3 0.04044118      3 0.4044118 0.4852941 0.05265639
## 4 0.01102941      5 0.3235294 0.3970588 0.04885621
## 5 0.01000000      7 0.3014706 0.4044118 0.04920453
##
## Variable importance
##     thal       cp  thatach  oldpeak       ca    exang   gender trestbps
##       25       19       14       11        9        9        6        2
##    slope      age
##        2        2
##
## Node number 1: 296 observations,    complexity param=0.4926471
##   predicted class=buff  expected loss=0.4594595  P(node) =1
##     class counts:   160   136
##    probabilities: 0.541 0.459
##   left son=2 (163 obs) right son=3 (133 obs)
##   Primary splits:
##       thal    splits as  RLR,       improve=41.30481, (0 missing)
##       cp      splits as  LLRL,      improve=37.51762, (0 missing)
##       ca      < 0.5   to the left,  improve=35.03455, (0 missing)
##       thatach < 147.5 to the right, improve=26.72095, (0 missing)
##       exang   splits as  LR,        improve=26.56741, (0 missing)
##   Surrogate splits:
##       thatach < 150.5 to the right, agree=0.679, adj=0.286, (0 split)
##       cp      splits as  LLRL,      agree=0.676, adj=0.278, (0 split)
##       exang   splits as  LR,        agree=0.669, adj=0.263, (0 split)
##       gender  splits as  LR,        agree=0.662, adj=0.248, (0 split)
##       oldpeak < 1.55  to the left,  agree=0.662, adj=0.248, (0 split)
##
## Node number 2: 163 observations,    complexity param=0.05147059
##   predicted class=buff  expected loss=0.2208589  P(node) =0.5506757
##     class counts:   127    36
##    probabilities: 0.779 0.221
##   left son=4 (114 obs) right son=5 (49 obs)
##   Primary splits:
##       ca      < 0.5   to the left,  improve=10.134680, (0 missing)
##       cp      splits as  LRRL,      improve= 8.181990, (0 missing)
##       oldpeak < 2.1   to the left,  improve= 6.279698, (0 missing)
##       thatach < 119.5 to the right, improve= 6.279698, (0 missing)
##       age     < 54.5  to the left,  improve= 5.870762, (0 missing)
##   Surrogate splits:
##       age     < 64.5  to the left,  agree=0.736, adj=0.122, (0 split)
##       thatach < 134   to the right, agree=0.724, adj=0.082, (0 split)
##       cp      splits as  LRLL,      agree=0.706, adj=0.020, (0 split)
##       oldpeak < 1.7   to the left,  agree=0.706, adj=0.020, (0 split)
##
## Node number 3: 133 observations,    complexity param=0.04044118
##   predicted class=sick  expected loss=0.2481203  P(node) =0.4493243
##     class counts:    33   100
##    probabilities: 0.248 0.752
##   left son=6 (44 obs) right son=7 (89 obs)
##   Primary splits:
##       cp      splits as  LLRL,      improve=9.916706, (0 missing)
```

```
##       ca      < 0.5   to the left,  improve=9.308898, (0 missing)
##       oldpeak < 0.7   to the left,  improve=7.335688, (0 missing)
##       thatach < 144.5 to the right, improve=6.169350, (0 missing)
##       exang   splits as  LR,        improve=5.288013, (0 missing)
##   Surrogate splits:
##       thatach  < 172   to the right, agree=0.722, adj=0.159, (0 split)
##       exang    splits as  LR,        agree=0.692, adj=0.068, (0 split)
##       age      < 66.5  to the right, agree=0.684, adj=0.045, (0 split)
##       trestbps < 106.5 to the left,  agree=0.684, adj=0.045, (0 split)
##
## Node number 4: 114 observations
##   predicted class=buff  expected loss=0.1052632  P(node) =0.3851351
##     class counts:   102    12
##    probabilities: 0.895 0.105
##
## Node number 5: 49 observations,    complexity param=0.05147059
##   predicted class=buff  expected loss=0.4897959  P(node) =0.1655405
##     class counts:    25    24
##    probabilities: 0.510 0.490
##   left son=10 (29 obs) right son=11 (20 obs)
##   Primary splits:
##       cp      splits as  LLRL,      improve=8.769106, (0 missing)
##       gender  splits as  LR,        improve=5.503264, (0 missing)
##       slope   splits as  RRL,       improve=4.576003, (0 missing)
##       thatach < 119.5 to the right, improve=4.251701, (0 missing)
##       exang   splits as  LR,        improve=3.432653, (0 missing)
##   Surrogate splits:
##       thatach  < 125.5 to the right, agree=0.755, adj=0.4, (0 split)
##       exang    splits as  LR,        agree=0.755, adj=0.4, (0 split)
##       trestbps < 115   to the right, agree=0.714, adj=0.3, (0 split)
##       oldpeak  < 0.85  to the left,  agree=0.714, adj=0.3, (0 split)
##       slope    splits as  RRL,       agree=0.714, adj=0.3, (0 split)
##
## Node number 6: 44 observations,    complexity param=0.04044118
##   predicted class=buff  expected loss=0.4772727  P(node) =0.1486486
##     class counts:    23    21
##    probabilities: 0.523 0.477
##   left son=12 (27 obs) right son=13 (17 obs)
##   Primary splits:
##       ca      < 0.5   to the left,  improve=4.577639, (0 missing)
##       thatach < 143.5 to the right, improve=4.183712, (0 missing)
##       slope   splits as  LRL,       improve=3.963092, (0 missing)
##       chol    < 207.5 to the left,  improve=2.426768, (0 missing)
##       oldpeak < 1.95  to the left,  improve=2.043434, (0 missing)
##   Surrogate splits:
##       cp      splits as  LL-R,      agree=0.705, adj=0.235, (0 split)
##       thatach < 125.5 to the right, agree=0.705, adj=0.235, (0 split)
##       oldpeak < 1.95  to the left,  agree=0.682, adj=0.176, (0 split)
##       age     < 67.5  to the left,  agree=0.659, adj=0.118, (0 split)
##       chol    < 190.5 to the right, agree=0.636, adj=0.059, (0 split)
##
## Node number 7: 89 observations,    complexity param=0.01102941
##   predicted class=sick  expected loss=0.1123596  P(node) =0.3006757
##     class counts:    10    79
```

```
##    probabilities: 0.112 0.888
##   left son=14 (21 obs) right son=15 (68 obs)
##   Primary splits:
##       oldpeak < 0.55  to the left,  improve=3.965694, (0 missing)
##       ca      < 0.5   to the left,  improve=1.893160, (0 missing)
##       chol    < 236.5 to the left,  improve=1.652505, (0 missing)
##       slope   splits as  RRL,       improve=1.132809, (0 missing)
##       restecg splits as  RRL,       improve=1.116074, (0 missing)
##   Surrogate splits:
##       thatach  < 146.5 to the right, agree=0.831, adj=0.286, (0 split)
##       slope    splits as  RRL,       agree=0.798, adj=0.143, (0 split)
##       trestbps < 109   to the left,  agree=0.787, adj=0.095, (0 split)
##
## Node number 10: 29 observations
##   predicted class=buff  expected loss=0.2413793  P(node) =0.09797297
##     class counts:    22     7
##    probabilities: 0.759 0.241
##
## Node number 11: 20 observations
##   predicted class=sick  expected loss=0.15  P(node) =0.06756757
##     class counts:     3    17
##    probabilities: 0.150 0.850
##
## Node number 12: 27 observations
##   predicted class=buff  expected loss=0.2962963  P(node) =0.09121622
##     class counts:    19     8
##    probabilities: 0.704 0.296
##
## Node number 13: 17 observations
##   predicted class=sick  expected loss=0.2352941  P(node) =0.05743243
##     class counts:     4    13
##    probabilities: 0.235 0.765
##
## Node number 14: 21 observations,    complexity param=0.01102941
##   predicted class=sick  expected loss=0.3809524  P(node) =0.07094595
##     class counts:     8    13
##    probabilities: 0.381 0.619
##   left son=28 (7 obs) right son=29 (14 obs)
##   Primary splits:
##       thatach < 149   to the left,  improve=2.333333, (0 missing)
##       chol    < 237.5 to the left,  improve=2.293651, (0 missing)
##       ca      < 0.5   to the left,  improve=1.250216, (0 missing)
##       age     < 50    to the right, improve=1.190476, (0 missing)
##       restecg splits as  -RL,       improve=1.190476, (0 missing)
##   Surrogate splits:
##       ca       < 2.5   to the right, agree=0.762, adj=0.286, (0 split)
##       age      < 62.5  to the right, agree=0.714, adj=0.143, (0 split)
##       gender   splits as  LR,        agree=0.714, adj=0.143, (0 split)
##       trestbps < 137.5 to the right, agree=0.714, adj=0.143, (0 split)
##       oldpeak  < 0.05  to the right, agree=0.714, adj=0.143, (0 split)
##
## Node number 15: 68 observations
##   predicted class=sick  expected loss=0.02941176  P(node) =0.2297297
##     class counts:     2    66
```

```
##    probabilities: 0.029 0.971
##
## Node number 28: 7 observations
##   predicted class=buff  expected loss=0.2857143  P(node) =0.02364865
##     class counts:     5     2
##    probabilities: 0.714 0.286
##
## Node number 29: 14 observations
##   predicted class=sick  expected loss=0.2142857  P(node) =0.0472973
##     class counts:     3    11
##    probabilities: 0.214 0.786
```

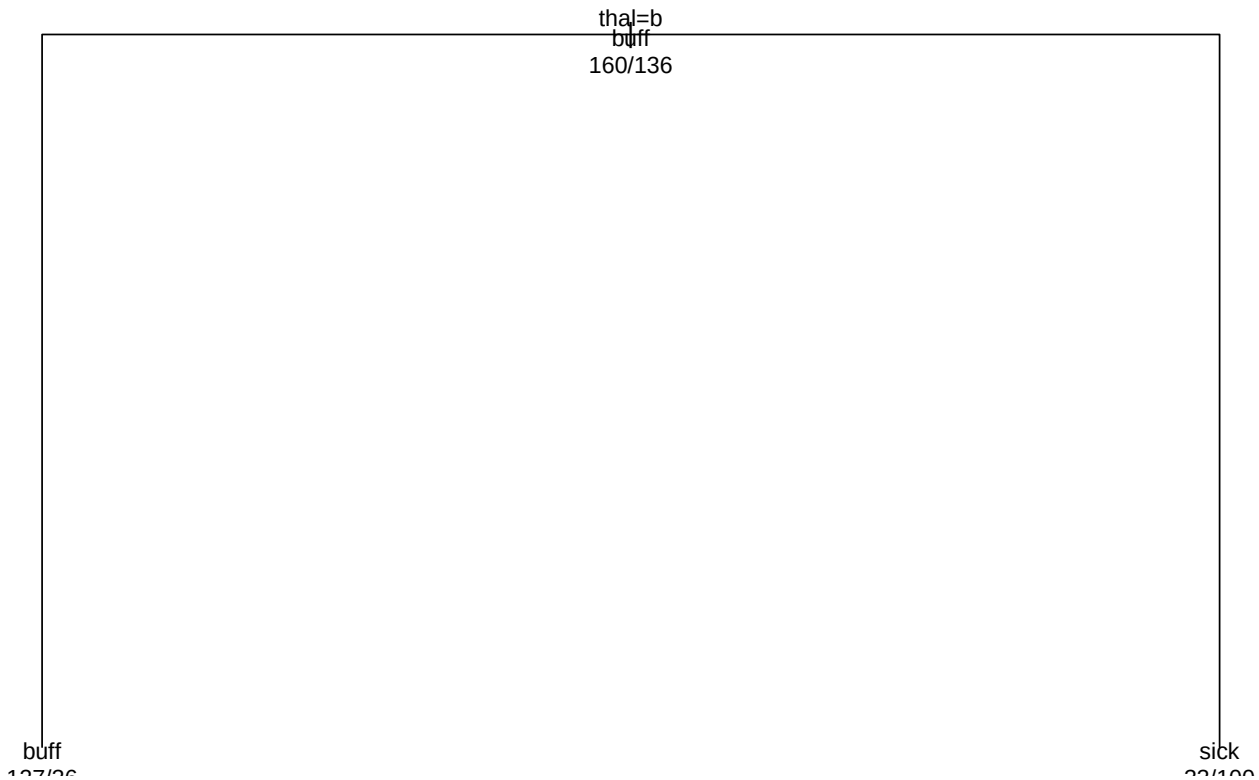
```
par(mfrow=c(1,1),mar=rep(6,4))
plot(fit, uniform=TRUE,
    main="Classification Tree for Heart Disease")
text(fit, use.n=TRUE, all=TRUE, cex=.8)
```



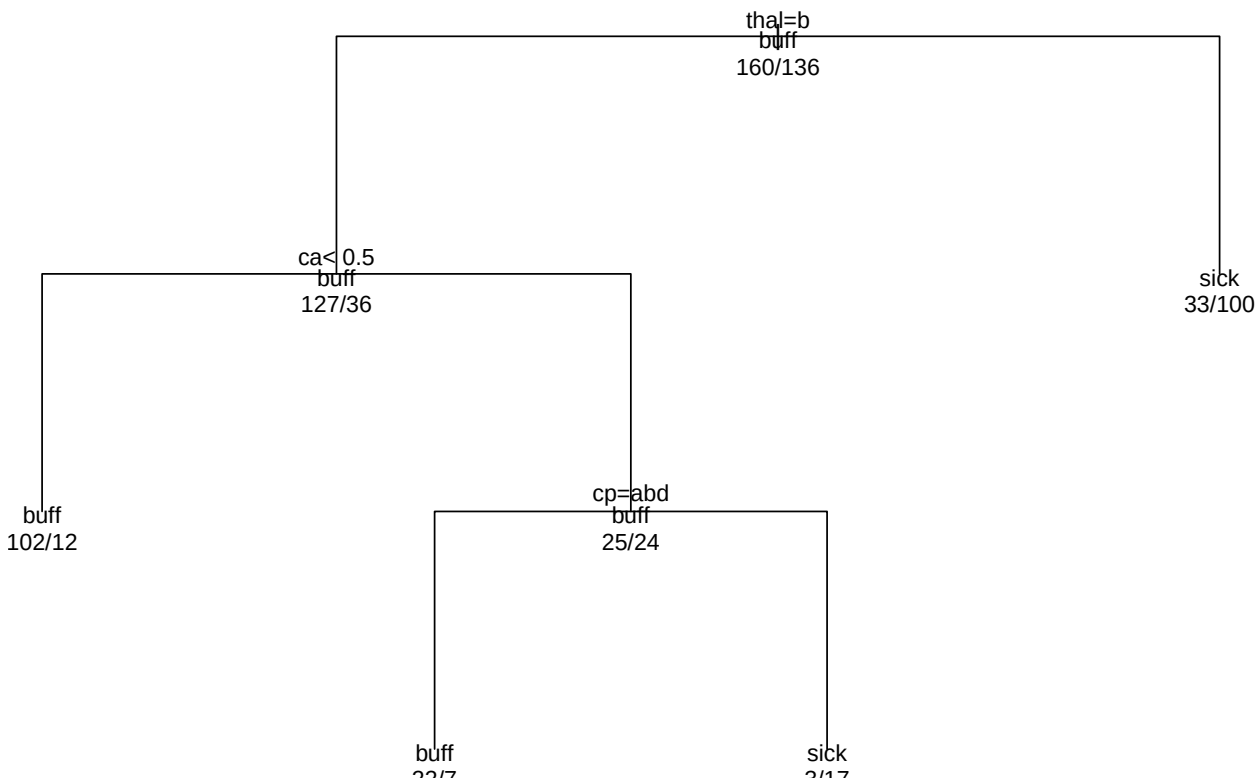

```
cp_vals <- fit$cptable
printcp(fit)
```

```
##
## Classification tree:
## rpart(formula = diag ~ ., data = Cle_heart2, method = "class")
##
## Variables actually used in tree construction:
## [1] ca      cp      oldpeak thal    thatach
##
## Root node error: 136/296 = 0.45946
##
## n= 296
##
##         CP nsplit rel error  xerror     xstd
## 1 0.492647      0   1.00000 1.00000 0.063044
## 2 0.051471      1   0.50735 0.57353 0.055730
## 3 0.040441      3   0.40441 0.48529 0.052656
## 4 0.011029      5   0.32353 0.39706 0.048856
## 5 0.010000      7   0.30147 0.40441 0.049205
```

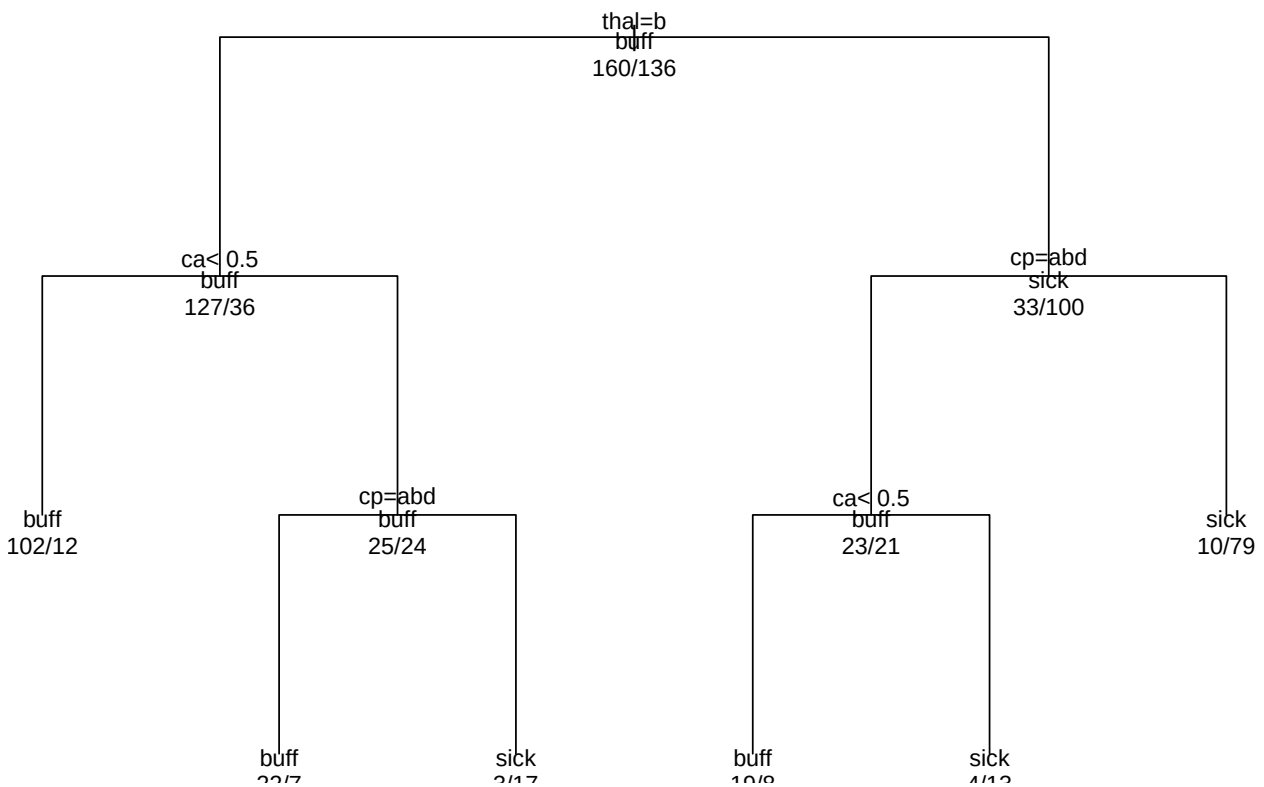
```
par(mfrow=c(1,1),mar=rep(6,4))
plot(prune(fit,cp=cp_vals[2,1]), uniform=TRUE,
    main="Classification Tree for Heart Disease")
text(prune(fit,cp=cp_vals[2,1]), use.n=TRUE, all=TRUE, cex=.8)
```




```
plot(prune(fit,cp=cp_vals[3,1]), uniform=TRUE,
    main="Classification Tree for Heart Disease")
text(prune(fit,cp=cp_vals[3,1]), use.n=TRUE, all=TRUE, cex=.8)
```




```
plot(prune(fit,cp=cp_vals[4,1]), uniform=TRUE,
    main="Classification Tree for Heart Disease")
text(prune(fit,cp=cp_vals[4,1]), use.n=TRUE, all=TRUE, cex=.8)
```




```
plot(prune(fit,cp=cp_vals[5,1]), uniform=TRUE,
    main="Classification Tree for Heart Disease")
text(prune(fit,cp=cp_vals[5,1]), use.n=TRUE, all=TRUE, cex=.8)
```

Classification Tree for Heart Disease