Classification and Regression Trees (CART) update

ACM

October 9th, 2023

Growing CART Further

```
library(printr)
```

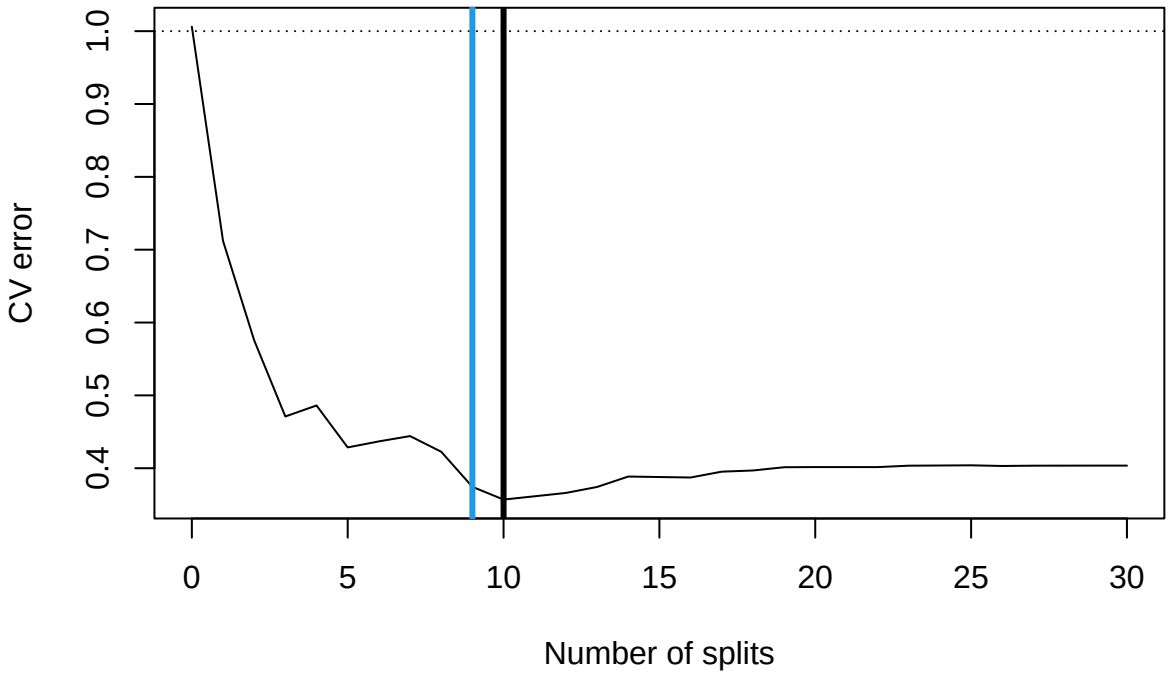
```
Registered S3 method overwritten by 'printr':
  method                from
  knit_print.data.frame rmarkdown
```

```
library(rpart)
library(rpart.plot)
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

```
set.seed(456546)
fit <- rpart(Salary ~ ., data = bball, method = "anova", control = list(minbucket = 5,
    cp = 2e-07, xval = 100))
```

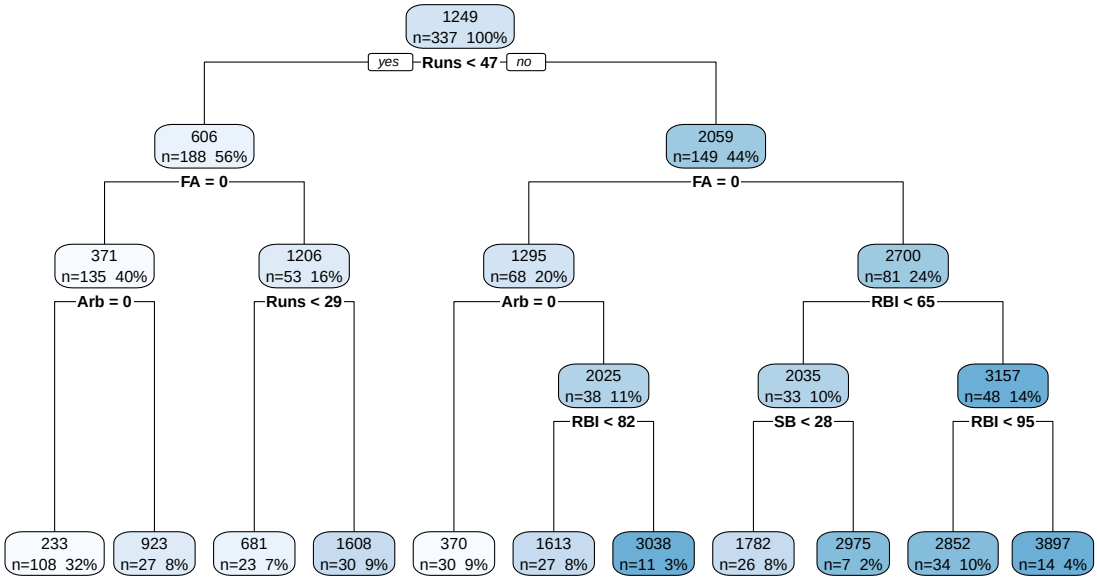
```
Fit_table <- as.data.frame(fit$cptable)
plot(Fit_table$nsplit,Fit_table$xerror,
     xlab = "Number of splits",ylab="CV error",
     type = "l", xlim=c(0,min(c(max(Fit_table$nsplit),100))))
abline(h=1,lty=3)
Best_mod <- which.min(Fit_table$xerror)
Best_mod_1sd <- which.max(Fit_table$xerror <
                            c(Fit_table$xerror+Fit_table$xstd) [which.min(Fit_table$xerror)]
                          )
abline(v = Fit_table$nsplit[Best_mod] ,lwd=3)
abline(v = Fit_table$nsplit[Best_mod_1sd], lwd = 3, col = 4)
```




```
print(Fit_table[Best_mod,])
```

```
##             CP nsplit rel error    xerror       xstd
## 11 0.009079524     10  0.222979 0.3569545 0.05090667
```

```
cp <- Fit_table$CP[Best_mod]
fit.best <- prune(fit,cp=cp)
rpart.plot(fit.best,extra = 101)
```



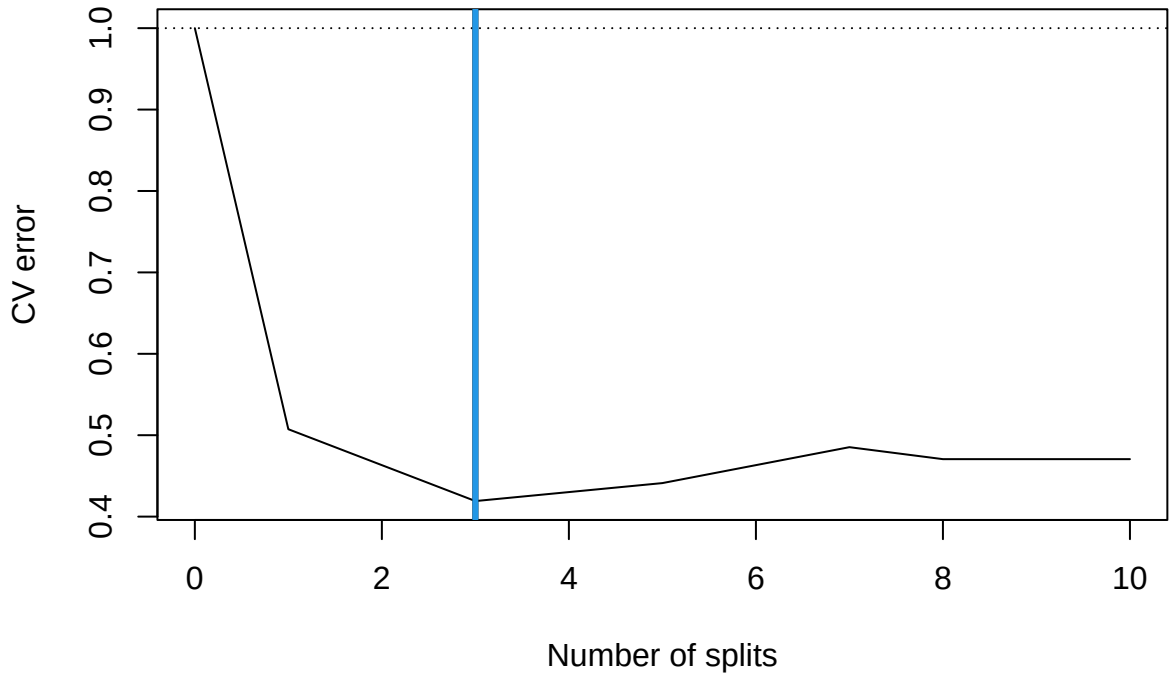
In this example we're going to look at the Cleveland heart study:

```
Cle_heart <- read.csv("Cle_heart.csv")
Cle_heart2 <- Cle_heart[,-15]
fit <- rpart(diag~.,data=Cle_heart2,method="class",
             control = list(minbucket = 5, cp = 1e-10, xval = 100))
```

```r
Fit_table <- as.data.frame(fit$cptable)
Fit_table
```

CP	nsplit	rel error	xerror	xstd
0.4926471	0	1.0000000	1.0000000	0.0630441
0.0514706	1	0.5073529	0.5073529	0.0534876
0.0404412	3	0.4044118	0.4191176	0.0498829
0.0110294	5	0.3235294	0.4411765	0.0508565
0.0073529	7	0.3014706	0.4852941	0.0526564
0.0036765	8	0.2941176	0.4705882	0.0520774
0.0000000	10	0.2867647	0.4705882	0.0520774

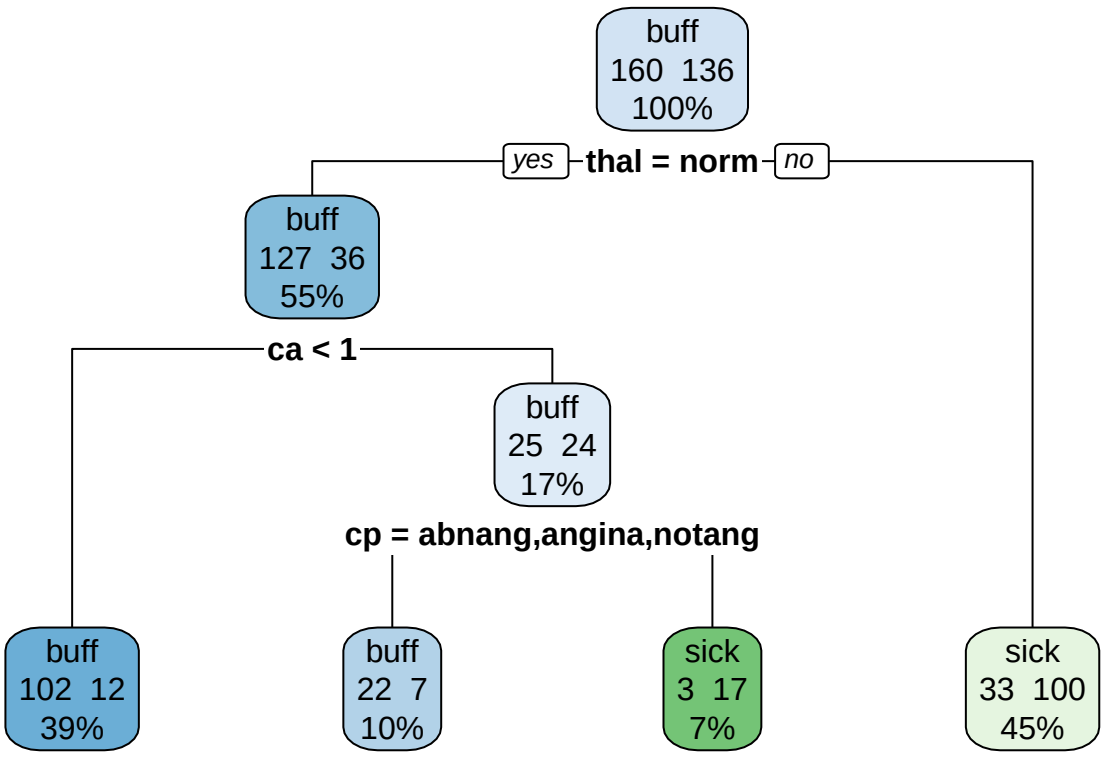
```r
plot(Fit_table$nsplit,Fit_table$xerror,
     xlab = "Number of splits",ylab="CV error",
     type = "l", xlim=c(0,min(c(max(Fit_table$nsplit),100))))
abline(h=1,lty=3)
Best_mod <- which.min(Fit_table$xerror)
Best_mod_1sd <- which.max(Fit_table$xerror <
                            c(Fit_table$xerror+Fit_table$xstd)[which.min(Fit_table$xerror)]
                          )
abline(v = Fit_table$nsplit[Best_mod] ,lwd=3)
abline(v = Fit_table$nsplit[Best_mod_1sd], lwd = 3, col = 4)
```




```r
print(Fit_table[Best_mod,])
```

```
##           CP nsplit rel error    xerror       xstd
## 3 0.04044118      3 0.4044118 0.4191176 0.04988289
```

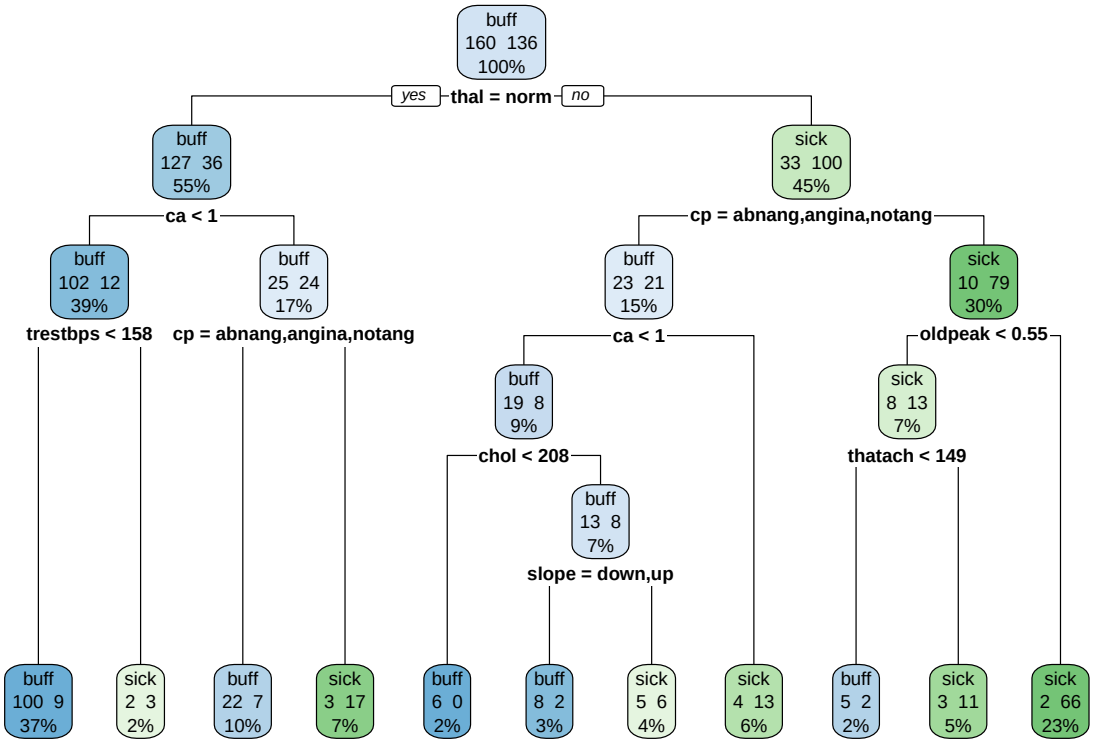
```r
cp <- Fit_table$CP[Best_mod]
fit.best <- prune(fit,cp=cp)
rpart.plot(fit.best,extra = 101)
```




```
print(Fit_table[7,])
```

```
##      CP nsplit rel error    xerror       xstd
## 7 1e-10     10 0.2867647 0.4705882 0.05207739
```

```
cp <- Fit_table$CP[7]
fit.best <- prune(fit,cp=cp)
rpart.plot(fit.best,extra = 101)
```


buff
160 136
100%
yes
thal = norm
no
buff
127 36
55%
sick
33 100
45%
ca < 1
cp = abnang,angina,notang
buff
102 12
39%
buff
25 24
17%
buff
23 21
15%
sick
10 79
30%
trestbps < 158
cp = abnang,angina,notang
ca < 1
oldpeak < 0.55
buff
19 8
9%
sick
8 13
7%
chol < 208
thatach < 149
buff
13 8
7%
slope = down,up
buff
100 9
37%
sick
2 3
2%
buff
22 7
10%
sick
3 17
7%
buff
6 0
2%
buff
8 2
3%
sick
5 6
4%
sick
4 13
6%
buff
5 2
2%
sick
3 11
5%
sick
2 66
23%