rpart包

翟启东：1600022771

王海洋：1600022710

rpart包

主要包括rpart（）函数和Prune（）函数，其中rpart（）函数构造决策树prune（）函数负责对决策树进行剪枝。

rpart（）函数

函数表现形式

rpart（formula，date，weights，subset，na.action=na.rpart,methond,modle=FALSE,x=FALSE,y=ture，parms，control，cost,...）

rpart（）函数

参数说明

formula：回归方程的形式。

date:前面方程的变量数据框

Na.action：缺失去数据处理方法，默认删除应变量缺失的观测，保留自变量缺失的观测。

method：树的末端数据类型选择相应的变量分割方法: 连续性**method="anova"**,离散型**method="class"**,计数型**method="poisson"**,生存分析型**method="exp"**

parms用来设置三个参数:先验概率、损失矩阵、分类纯度的度量方法

rpart（）函数

示例

```
fit <- rpart(Kyphosis~Age + Number + Start, +
data=kyphosis, method="class", + parms = list(prior =
c(0.65,0.35), split = "information"))
```

- par(mfrow=c(1,3))
- plot(fit)
- text(fit,use.n=T,all=T,cex=0.9)

prune（）函数

函数表现形式

prune（tree，cp...）

参数说明

Tree:一个回归树的对象，常常是rpart（）的结果对象

Cp：复杂性参量，指定剪枝的采用的阈值

prune （） 函数

示例

```
> printcp(fit)

Classification tree:
rpart(formula = Kyphosis ~ Age + Number + Start, data = kyphosis,
    method = "class", parms = list(prior = c(0.65, 0.35), split = "information"),
    control = ct)

Variables actually used in tree construction:
[1] Age   Start

Root node error: 28.35/81 = 0.35

n= 81

        CP nsplit rel error xerror    xstd
1 0.30200      0   1.00000 1.0000 0.21559
2 0.20234      1   0.69800 1.3459 0.20307
3 0.10000      2   0.49567 1.2000 0.19959
```

prune（）函数

示例

```
➢ fit2 <- prune(fit, cp=0.01);
➢ > par(mfrow=c(1,3))
➢ > plot(fit2)
➢ > text(fit,use.n=T,all=T,cex=0.9)
```

实例操作

前期准备工作

```
install.packages("rpart")
install.packages("survival")

Library ("rpart")
Library ("survival")
```

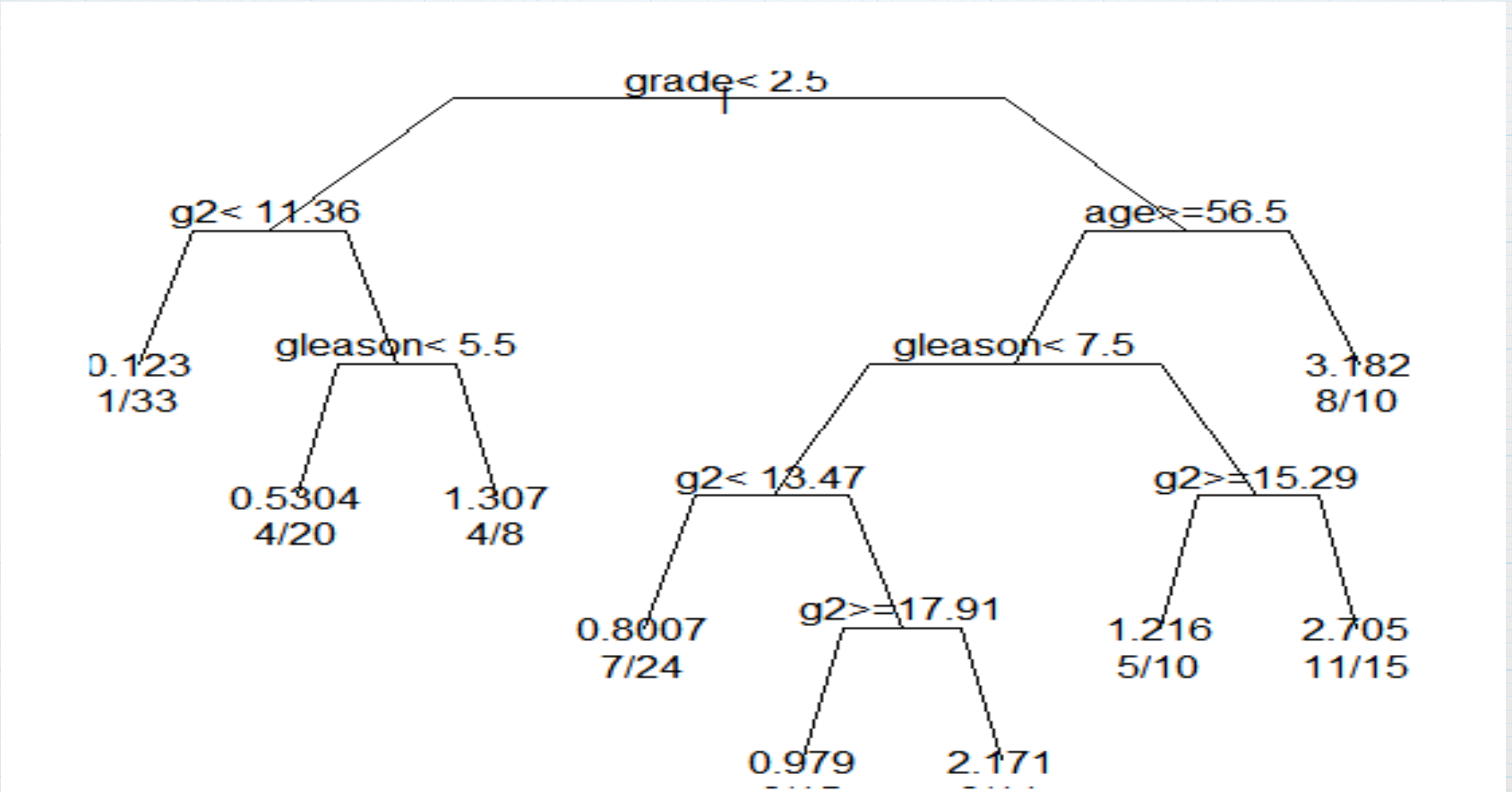
实例操作

数据查看

```
> library(rpart)
Warning message:
程辑包'rpart'是用R版本3.3.3 来建造的
> library(survival)
Warning message:
程辑包'survival'是用R版本3.3.3 来建造的
> stagec
    pgtime pgstat age eet    g2 grade gleason     ploidy
1      6.1      0  64   2 10.26     2       4    diploid
2      9.4      0  62   1    NA     3       8  aneuploid
3      5.2      1  59   2  9.99     3       7    diploid
4      3.2      1  62   2  3.57     2       4    diploid
5      1.9      1  64   2 22.56     4       8 tetraploid
6      4.8      0  69   1  6.14     3       7    diploid
7      5.8      0  75   2 13.69     2      NA tetraploid
8      7.3      0  71   2    NA     3       7  aneuploid
9      3.7      1  73   2 11.77     3       6    diploid
10    15.9      0  64   2 27.27     3       7 tetraploid
11     6.3      0  65   2 19.34     3       7 tetraploid
```

实例操作

利用rpart函数构建决策树



实例操作

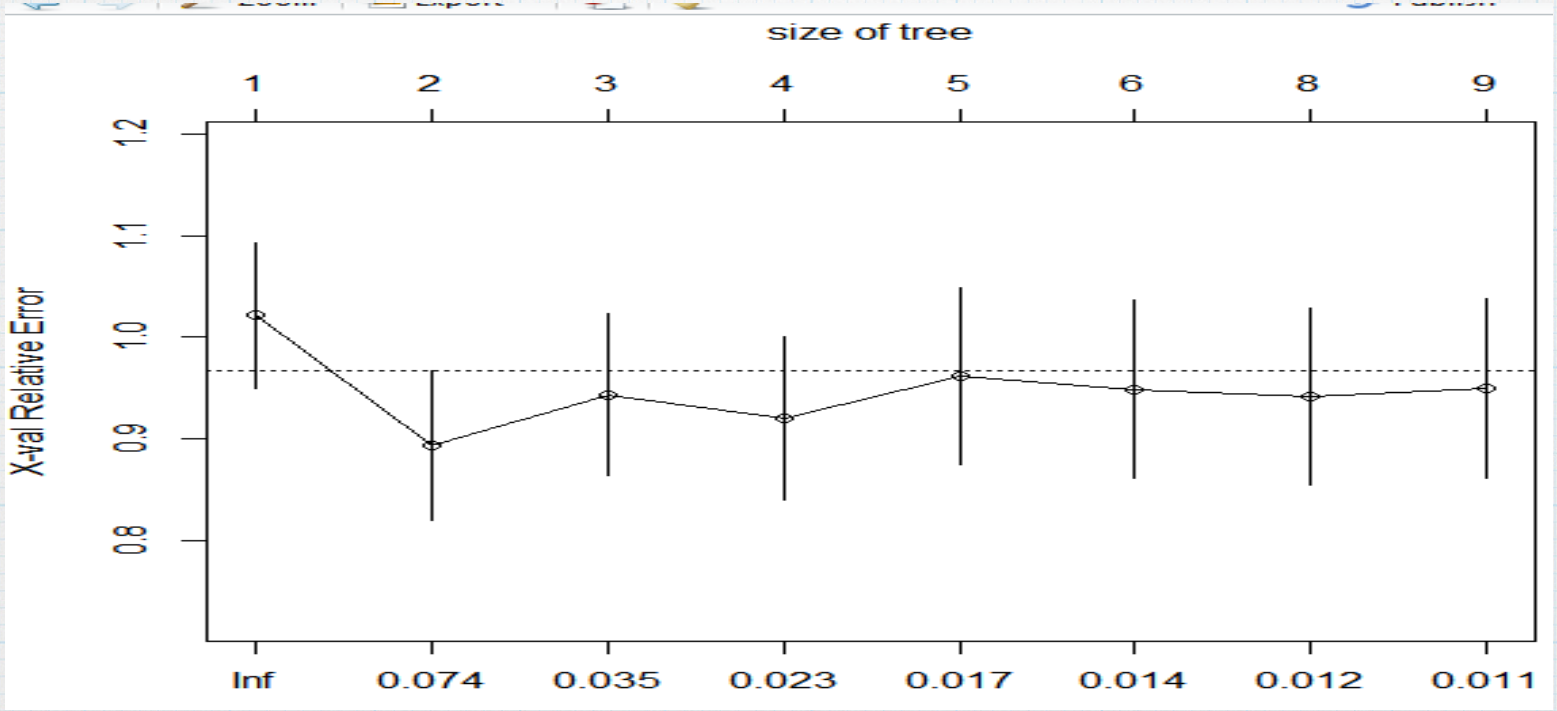
查看决策树的具体信息

```
> printcp(fit)

Survival regression tree:
rpart(formula = Surv(pgtime, pgstat) ~ age + eet + g2 + grade +
    gleason + ploidy, data = stagec, method = "exp")

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

Root node error: 192.11/146 = 1.3158

n= 146

        CP nsplit rel error  xerror     xstd
1 0.129460      0   1.00000 1.02160 0.071282
2 0.042056      1   0.87054 0.89378 0.073545
3 0.029200      2   0.82848 0.94335 0.079683
4 0.017989      3   0.79928 0.92023 0.080868
5 0.015406      4   0.78130 0.96194 0.087418
6 0.013354      5   0.76589 0.94893 0.087607
7 0.011506      7   0.73918 0.94205 0.086796
8 0.010000      8   0.72768 0.94945 0.088322
```

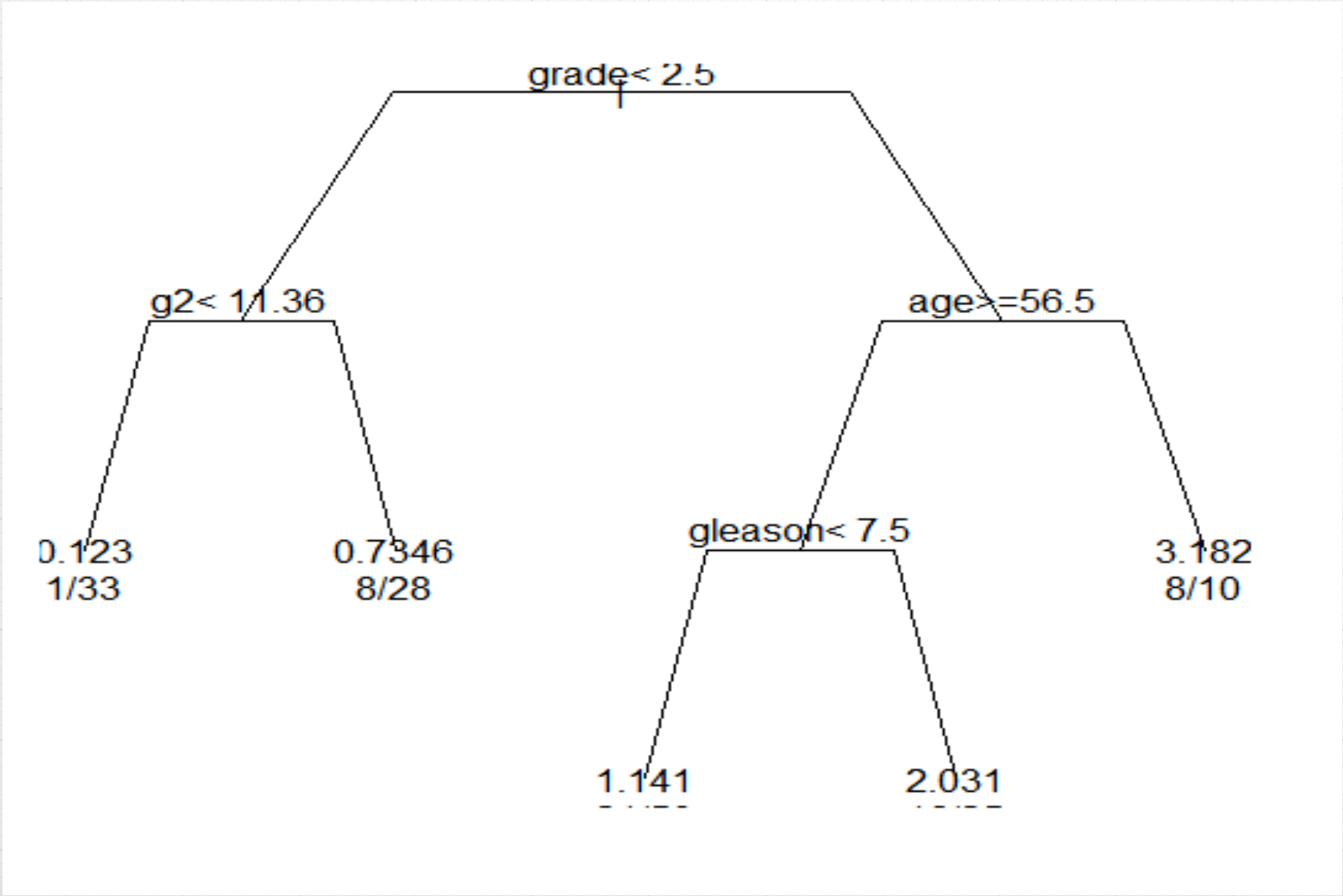
实例操作

查看决策树的具体信息

plotcp（fit）



实例操作


grade< 2.5
g2< 11.36
age>=56.5
0.123
1/33
0.7346
8/28
gleason< 7.5
3.182
8/10
1.141
2.031

作业

1.根据前文范例，利用rpart包中自带的数据集kyphosis，构建决策树。

2.通过prune函数对构造的决策树进行剪枝。

汇报完毕，谢谢大家