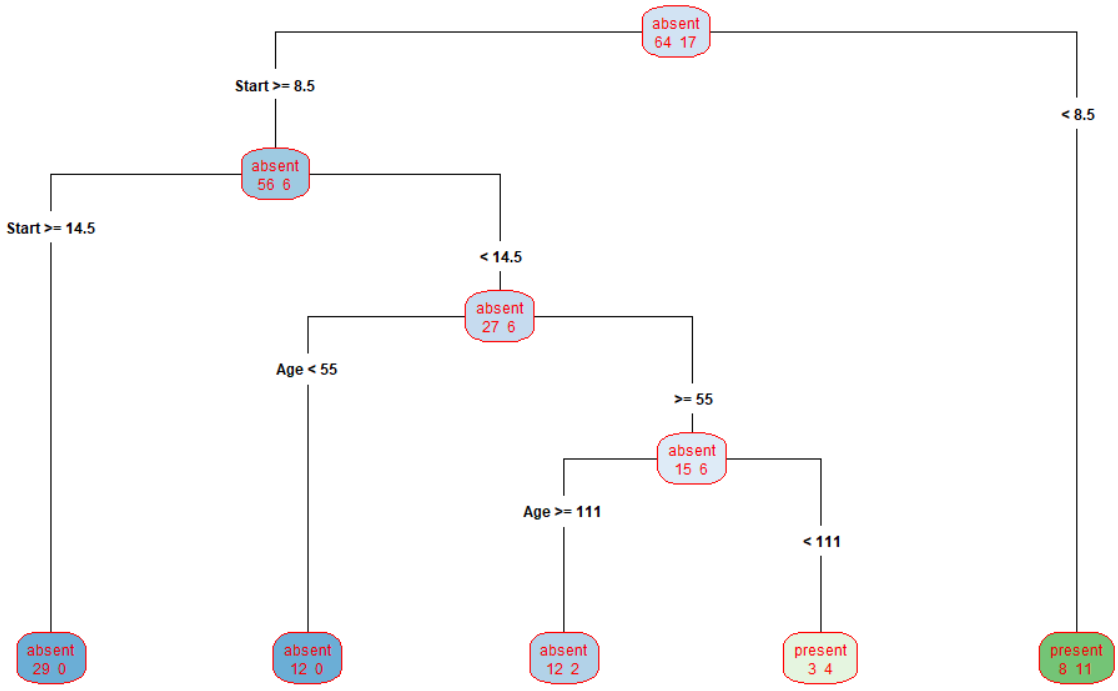
26 BE 7082 + 26 PH 7028 + 20 BME 7082
Introduction to Data Science
Autumn 2020

MB Rao

Home Work 4: Due date: September 24, 2020 Maximum points: 30

A preamble: In Lecture 4, we have seen how to present a regression tree as a polygonal graph if there are only two predictors showing up in the tree. I want you develop a polygonal tree in the environment of Classification Trees.

1. Build a classification tree for the data 'kyphosis' from the 'rpart' package. 5 points



I used the following lines of code to generate the classification tree:
data(kyphosis)
SF<-rpart(Kyphosis ~., data=kyphosis)
rpart.plot(SF, type=4, extra=1, digits=3, roundint=FALSE, col="red")

2. Describe the tree verbally. 5 points

The root node contains a total of 81 subject data out of which 64 have Kyphosis absent and 17 have Kyphosis present. The root node is split based on Start value (The number of first or topmost vertebra operated on). For case that have start >= 9, they split to the left and Kyphosis is found to be absent in 56 and present in 6 such cases. That is a total of 62 cases. Those cases with Start <9 go towards a terminal node on the right and a total of 8 out of 19 cases are found to have kyphosis absent. The node is classified as terminal as the total number of cases for this node is less than 20 which is the default minimum split size of the node in R. The node on the left is split into two subgroups based on start >= 15 towards left and Start <15 towards right. The sub-split on the left is pure as it contains only 29 absent cases and no present cases and hence the left sub-node node is classified as terminal. The sub-node on the right is classified to have 26 absent and 6 present cases and is further split. This next level of split is based on the age of the case subjects (Children) in months. For children with age >= 55 months, the classification tree splits into a pure node with 12 absent cases to the left. For children <55 months the node is split to the right with 15 absent and 6 present cases. This node is further divided based on Age of the children. For children with age >= 111 months the node is sub-divided into a terminal node to the left with 12 kyphosis absent and 2 kyphosis present cases. On the right for children aged <111 but more than 55 the node is divided to another terminal node with 3 absent cases and 4 present cases. Finally, using print(SF) I obtained the following description of the classification tree:

```
> print(SF)
n= 81
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 81 17 absent (0.79012346 0.20987654)
   2) Start>=8.5 62  6 absent (0.90322581 0.09677419)
     4) Start>=14.5 29  0 absent (1.00000000 0.00000000) *
     5) Start< 14.5 33  6 absent (0.81818182 0.18181818)
      10) Age< 55 12  0 absent (1.00000000 0.00000000) *
      11) Age>=55 21  6 absent (0.71428571 0.28571429)
        22) Age>=98 16  2 absent (0.87500000 0.12500000) *
        23) Age< 98 5  1 present (0.20000000 0.80000000) *
   3) Start< 8.5 19  8 present (0.42105263 0.57894737)
     6) Age< 93 10  4 absent (0.60000000 0.40000000)
      12) Number< 4.5 4  0 absent (1.00000000 0.00000000) *
      13) Number>=4.5 6  2 present (0.33333333 0.66666667) *
     7) Age>=93 9  2 present (0.22222222 0.77777778) *
```

3. Develop a polygonal tree. Make it informative. 20 points

I used the following lines of code to generate the polygonal tree:

```
plot(kyphosis$Start, kyphosis$Age, type=n)
polygon(c(1, 9, 9, 1), c(1, 1, 206, 206), col="mistyrose")
polygon(c(15,18, 18, 15), c(1, 1, 206, 206), col="green")
polygon(c(9, 15, 15, 9), c(1, 1, 55, 55), col="green")
polygon(c(9, 15, 15, 9), c(111, 111, 206, 206), col="green")
polygon(c(9, 15, 15, 9), c(55,55,111,111), col="mistyrose")
axis(side=2, at=c(1,55, 111,206), labels=c(1,55,  111, 206), col="magenta")
axis(side=1, at=c(1,9,15, 18), labels=c(1, 9,15,18), col="magenta")
title(ylab="Age (in months)", xlab="Start (number of first or topmost vertebra
operated on)", main="Regression Tree - Polygonal Graph", sub="Kyphosis
Data") text(4.5,103, "Present", col="magenta")
text(16.5, 103, "Absent", col="magenta")
text(12,27.5, "Absent", col="magenta")
text(12, 83, "Present", col="magenta")
text(12, 158.5, "Absent", col="magenta")
```