Classification and Regression Trees in R

Oliver

# Sample Data

The following code reads the titanic data that we will use in our examples.

titanic = read.csv(  
 "http://facweb1.redlands.edu/fac/jim\_bentley/downloads/math111/titanic.csv") ### Note data are in Math111 not FYS04  
 titanic$AGE=factor(titanic$AGE,labels=c('Child','Adult'))  
 titanic$CLASS=factor(titanic$CLASS,labels=c('0','1','2','3'))  
 titanic$SEX=factor(titanic$SEX, labels=c('Female','Male'))  
 titanic$SURVIVED=factor(titanic$SURVIVED,labels=c('No','Yes'))

We can now check to see if the dataframe has been created by entering:

ls()

## [1] "titanic"

# Loading R Packages

## load a few packages  
 p\_load(xtable)  
 p\_load(rpart)  
 p\_load(rpart.plot)  
 #p\_load(rpartOrdinal)  
 p\_load(Hmisc)  
 p\_load(xtable)  
 p\_load(lattice)

# Fitting CART

The CARTs fitted here are analogous to the logistic models fitted in SAS and R.

## CLASS

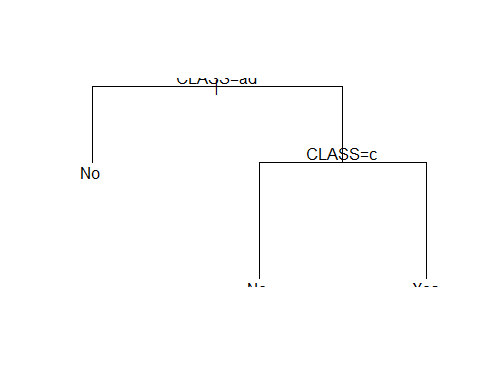
A classification tree to look at the predictive nature of class when looking at survival may be fitted using the **rpart** function.

titanic.rpart.class=rpart(SURVIVED~CLASS,data=titanic)  
 summary(titanic.rpart.class)

## Call:  
## rpart(formula = SURVIVED ~ CLASS, data = titanic)  
## n= 2201   
##   
## CP nsplit rel error xerror xstd  
## 1 0.05696203 0 1.0000000 1.0000000 0.03085662  
## 2 0.01000000 2 0.8860759 0.8860759 0.02982488  
##   
## Variable importance  
## CLASS   
## 100   
##   
## Node number 1: 2201 observations, complexity param=0.05696203  
## predicted class=No expected loss=0.323035 P(node) =1  
## class counts: 1490 711  
## probabilities: 0.677 0.323   
## left son=2 (1591 obs) right son=3 (610 obs)  
## Primary splits:  
## CLASS splits as LRRL, improve=69.6841, (0 missing)  
##   
## Node number 2: 1591 observations  
## predicted class=No expected loss=0.2451288 P(node) =0.7228532  
## class counts: 1201 390  
## probabilities: 0.755 0.245   
##   
## Node number 3: 610 observations, complexity param=0.05696203  
## predicted class=Yes expected loss=0.4737705 P(node) =0.2771468  
## class counts: 289 321  
## probabilities: 0.474 0.526   
## left son=6 (285 obs) right son=7 (325 obs)  
## Primary splits:  
## CLASS splits as -RL-, improve=13.46678, (0 missing)  
##   
## Node number 6: 285 observations  
## predicted class=No expected loss=0.4140351 P(node) =0.1294866  
## class counts: 167 118  
## probabilities: 0.586 0.414   
##   
## Node number 7: 325 observations  
## predicted class=Yes expected loss=0.3753846 P(node) =0.1476602  
## class counts: 122 203  
## probabilities: 0.375 0.625

A plot of the tree may be created using:

plot(titanic.rpart.class)  
 text(titanic.rpart.class)



## AGE and SEX

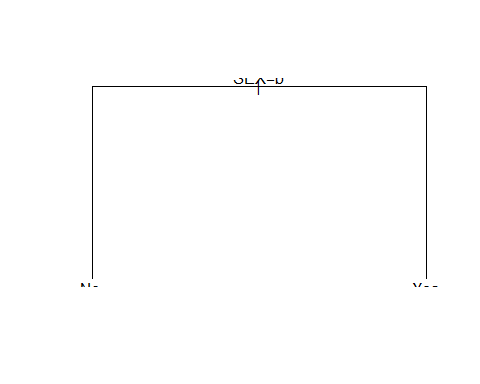
A classification tree to look at the predictive nature of age and sex when looking at survival may be fitted using the **rpart** function.

titanic.rpart.agesex=rpart(SURVIVED~AGE+SEX,data=titanic)  
 summary(titanic.rpart.agesex)

## Call:  
## rpart(formula = SURVIVED ~ AGE + SEX, data = titanic)  
## n= 2201   
##   
## CP nsplit rel error xerror xstd  
## 1 0.3066104 0 1.0000000 1.0000000 0.03085662  
## 2 0.0100000 1 0.6933896 0.6933896 0.02750982  
##   
## Variable importance  
## SEX   
## 100   
##   
## Node number 1: 2201 observations, complexity param=0.3066104  
## predicted class=No expected loss=0.323035 P(node) =1  
## class counts: 1490 711  
## probabilities: 0.677 0.323   
## left son=2 (1731 obs) right son=3 (470 obs)  
## Primary splits:  
## SEX splits as RL, improve=199.821600, (0 missing)  
## AGE splits as RL, improve= 9.165241, (0 missing)  
##   
## Node number 2: 1731 observations  
## predicted class=No expected loss=0.2120162 P(node) =0.7864607  
## class counts: 1364 367  
## probabilities: 0.788 0.212   
##   
## Node number 3: 470 observations  
## predicted class=Yes expected loss=0.2680851 P(node) =0.2135393  
## class counts: 126 344  
## probabilities: 0.268 0.732

A plot of the tree may be created using:

plot(titanic.rpart.agesex)  
 text(titanic.rpart.agesex)



## CLASS, AGE and SEX

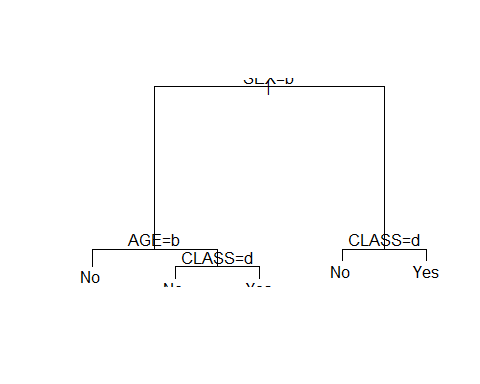
A classification tree to look at the predictive nature of class, age and sex when looking at survival may be fitted using the **rpart** function.

titanic.rpart.classagesex=rpart(SURVIVED~CLASS+AGE+SEX,data=titanic)  
 summary(titanic.rpart.classagesex)

## Call:  
## rpart(formula = SURVIVED ~ CLASS + AGE + SEX, data = titanic)  
## n= 2201   
##   
## CP nsplit rel error xerror xstd  
## 1 0.30661041 0 1.0000000 1.0000000 0.03085662  
## 2 0.02250352 1 0.6933896 0.6933896 0.02750982  
## 3 0.01125176 2 0.6708861 0.6849508 0.02738989  
## 4 0.01000000 4 0.6483826 0.6736990 0.02722731  
##   
## Variable importance  
## SEX CLASS AGE   
## 73 23 4   
##   
## Node number 1: 2201 observations, complexity param=0.3066104  
## predicted class=No expected loss=0.323035 P(node) =1  
## class counts: 1490 711  
## probabilities: 0.677 0.323   
## left son=2 (1731 obs) right son=3 (470 obs)  
## Primary splits:  
## SEX splits as RL, improve=199.821600, (0 missing)  
## CLASS splits as LRRL, improve= 69.684100, (0 missing)  
## AGE splits as RL, improve= 9.165241, (0 missing)  
##   
## Node number 2: 1731 observations, complexity param=0.01125176  
## predicted class=No expected loss=0.2120162 P(node) =0.7864607  
## class counts: 1364 367  
## probabilities: 0.788 0.212   
## left son=4 (1667 obs) right son=5 (64 obs)  
## Primary splits:  
## AGE splits as RL, improve=7.726764, (0 missing)  
## CLASS splits as LRLL, improve=7.046106, (0 missing)  
##   
## Node number 3: 470 observations, complexity param=0.02250352  
## predicted class=Yes expected loss=0.2680851 P(node) =0.2135393  
## class counts: 126 344  
## probabilities: 0.268 0.732   
## left son=6 (196 obs) right son=7 (274 obs)  
## Primary splits:  
## CLASS splits as RRRL, improve=50.015320, (0 missing)  
## AGE splits as LR, improve= 1.197586, (0 missing)  
## Surrogate splits:  
## AGE splits as LR, agree=0.619, adj=0.087, (0 split)  
##   
## Node number 4: 1667 observations  
## predicted class=No expected loss=0.2027594 P(node) =0.757383  
## class counts: 1329 338  
## probabilities: 0.797 0.203   
##   
## Node number 5: 64 observations, complexity param=0.01125176  
## predicted class=No expected loss=0.453125 P(node) =0.02907769  
## class counts: 35 29  
## probabilities: 0.547 0.453   
## left son=10 (48 obs) right son=11 (16 obs)  
## Primary splits:  
## CLASS splits as -RRL, improve=12.76042, (0 missing)  
##   
## Node number 6: 196 observations  
## predicted class=No expected loss=0.4591837 P(node) =0.08905043  
## class counts: 106 90  
## probabilities: 0.541 0.459   
##   
## Node number 7: 274 observations  
## predicted class=Yes expected loss=0.0729927 P(node) =0.1244889  
## class counts: 20 254  
## probabilities: 0.073 0.927   
##   
## Node number 10: 48 observations  
## predicted class=No expected loss=0.2708333 P(node) =0.02180827  
## class counts: 35 13  
## probabilities: 0.729 0.271   
##   
## Node number 11: 16 observations  
## predicted class=Yes expected loss=0 P(node) =0.007269423  
## class counts: 0 16  
## probabilities: 0.000 1.000

A plot of the tree may be created using:

plot(titanic.rpart.classagesex)  
 text(titanic.rpart.classagesex)



# Additional Functions

The documentation for the function **rpart** shows how to prune classification trees. There are also a number of sites on the web that show how to interpret **rpart** output.