2 Classification of Titanic Survials

Tobias Machnitzki and Finn Burgemeister

08 Dezember 2017

```
library(knitr)
library(markdown)
library(rpart)
library(rpart.plot)
library(plyr)
library(ggplot2)
```

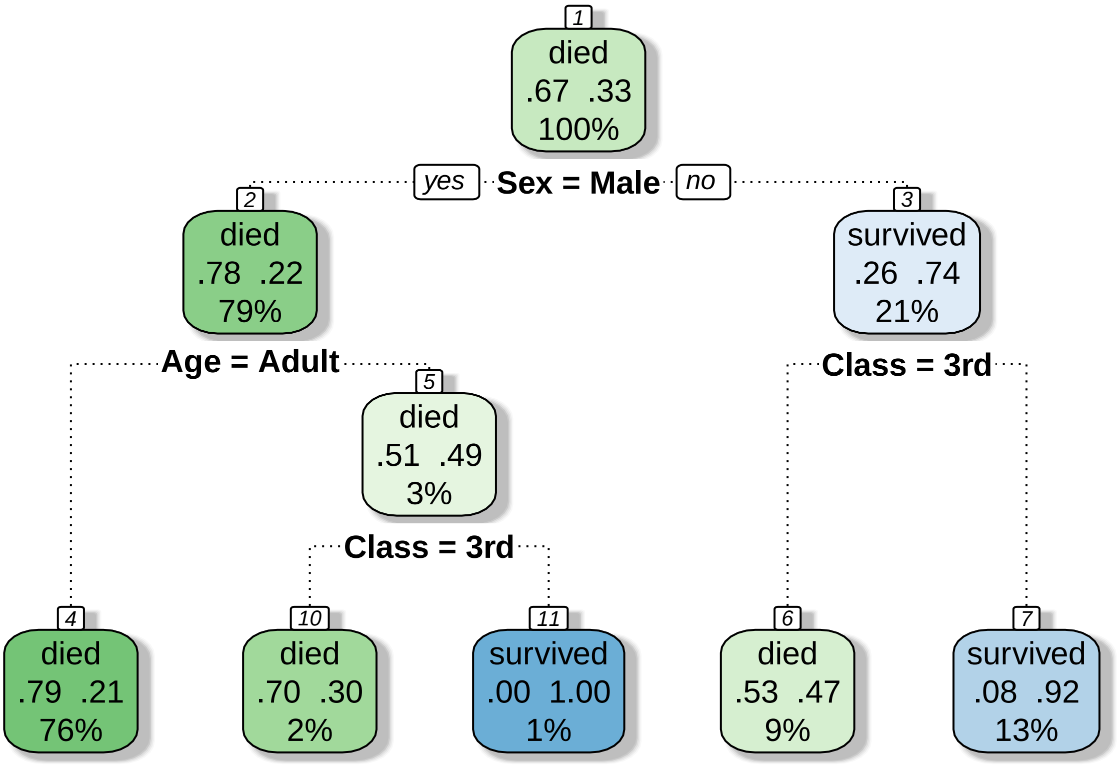
0 Import Data

```
# Import data
dTitanic = read.csv("/home/finn/Schreibtisch/Studium/17_WiSe/BigDataAnalytics/WORK/Blatt7/titanic.csv",
                    header=T,
                    sep=",",
                    quote="\"")

colnames(dTitanic)[1] = "Num"

dTitanic$Survived = gsub("No", "died", dTitanic$Survived)
dTitanic$Survived = gsub("Yes", "survived", dTitanic$Survived)

# Create one training dataset
#mask = sample(2, nrow(dTitanic), repl=T, prob=c(0.9,0.1))
#validation = dTitanic[mask==1, ]
#training = dTitanic[mask==2, ]
```

1 Create a decision tree for all passengers and try to deduct useful rules for determining the survival or demise of a passenger

```
set.seed(123)

# create 10 folds
folds = split(dTitanic, cut(sample(1:nrow(dTitanic)),10))
errs = rep(NA, length(folds))
errFP = rep(NA, length(folds))
errFN = rep(NA, length(folds))

for (i in 1:length(folds)) {
 validation = ldply(folds[i], data.frame)
 training = ldply(folds[-i], data.frame)

 m = rpart(Survived ~ Age+Class+Sex,
          data=training,
          method="class")

 p = predict(m,
             newdata = validation,
             type = "class")

 confmat = table(validation$Survived, p)

 # Compute the error rate of the predictions
 errs[i] = 1-sum(diag(confmat))/sum(confmat)
 errFN[i] = sum(confmat[2,1])/sum(confmat)
```

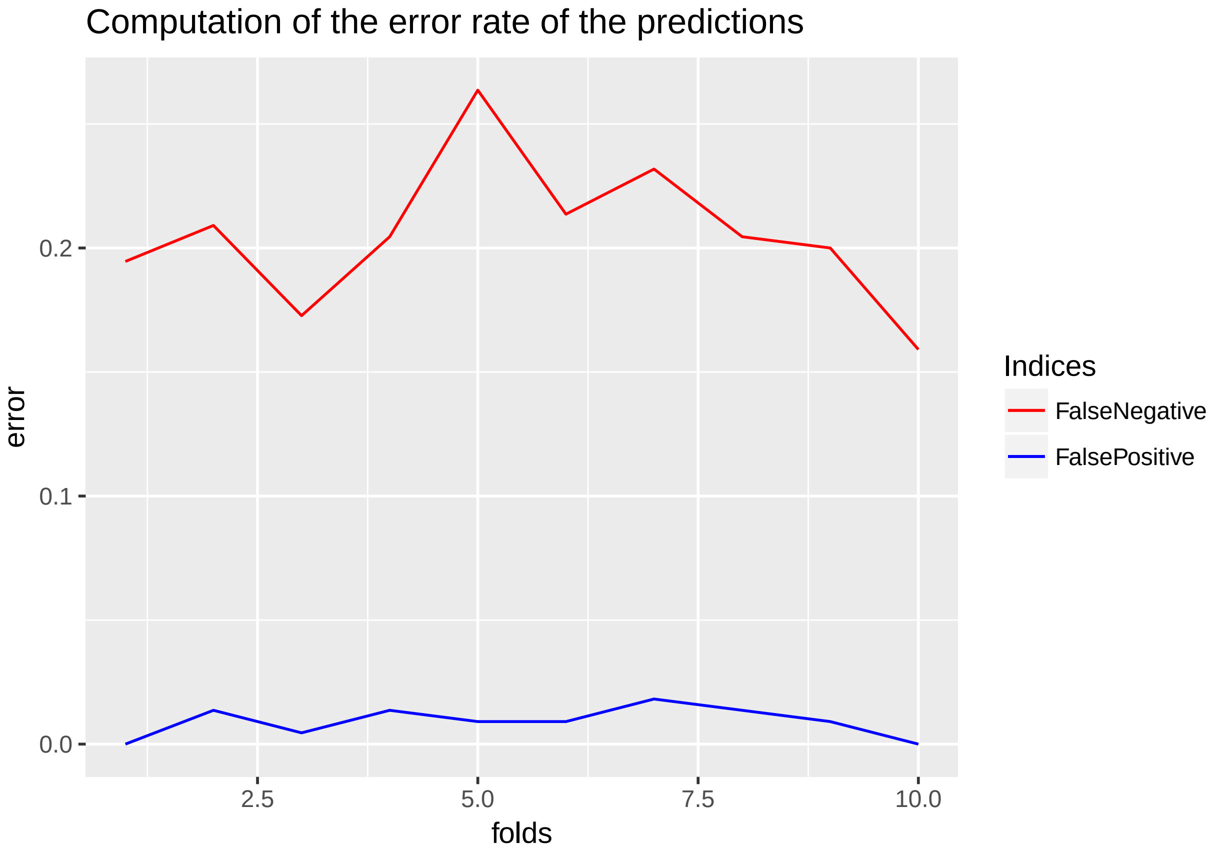
```
  errFP[i] = sum(confmat[1,2])/sum(confmat)
}

rpart.plot(m,
           extra=104, box.palette="GnBu",
           branch.lty=3, shadow.col="gray", nn=TRUE)
```



```
# evaluate the accuracy of your decision tree
print(sprintf("average error using k-fold cross-validation: %.3f percent", 100*mean(errs)))
```

```
## [1] "average error using k-fold cross-validation: 21.446 percent"
```

```
# Create data
dataFN=data.frame(folds=seq(1,10), errorFalseNegative=errFN)
dataFP=data.frame(folds=seq(1,10), errorFalsePositive=errFP)

plot = ggplot() +
  geom_line(data=dataFN, aes(x=folds, y=errorFalseNegative, colour="FalseNegative")) +
  geom_line(data=dataFP, aes(x=folds, y=errorFalsePositive, colour="FalsePositive")) +
  scale_color_manual(values = c(FalseNegative="red", FalsePositive="blue")) +
  labs(color="Indices") +
  xlab("folds") +
  ylab("error") +
  ggtitle("Computation of the error rate of the predictions")

plot
```



We have 10 subsets of the data for validation and training. In the decision tree we predict the class of survival with the predictors of the other classes age, sex and class/ticket of the training dataset.

The following description is only for one possible decision tree. (Node 1:) The root node shows 67 % of the passengers died, while 33 % survived used 100 % of the dataset. (Node 2:) If the passenger was male 78 % of the male passengers died. 79 % of the passengers were male. (Node 4:) If the male passenger was an adult (76 % of the passengers) they died with a probability of 79 %. (Node 5:) If the male passenger was a child (3 %) they died with a probability of 51 %. If the male child had a ticket for the 3rd class he died with a chance of 70 %, if not he survived. (Node 3:) If the passenger was female 73 % survived of the female passengers (21 %). If the female had a ticket for the third class, she died with a probability of 53 %. Otherwise they survived with a chance of 92 % (13 % of all passengers).

Based on this decision tree a useful rule for survival would be: be female without a ticket for the third class or be a male child in the class 1 or 2. A useful rule for demise is: be an adult male passenger.

For the errors: True positive (TP): observation is true, predicted as true (AA) predicted died and died False positive (FP): observation is false, prediction is true (BA, CA) predicted survived and died True negative (TN): observation is false, predicted as false (BB, CC) predicted survived and survived False negative (FN): observation is true, prediction is false (else) predicted died and survived