03_machine_learning

Isabel García Valdivia

4/25/2018

PART III. MACHINE LEARNING ANALYSIS

Using machine learning as a method to estimate f, in this case legal status, categorical variable 0/1 (in other words - F/T) for CPS 2016 data using CPS 2017 data.

Load Libraries/Packages

```
library(ggplot2)      #visuals
library(rpart)        #decision tree
library(rpart.plot)   #decision tree plots
```

Machine Learning: Decision trees

```
#omit the undoc variable and create a subset
# Choose method="anova" for a regression tree
# Instead of gini coefficient, select "information" inside the "parms" parameter
dt <- rpart(cps17_mlearning$undoc_log ~ ., data = subset(cps17_mlearning, select = -cps17_mlearning$undo
            method = "class",
            parms = list(split = 'information'))

# Here is the text-based display of the decision tree. Yikes!
print(dt)

# The plot is much easier to interpret.
rpart.plot(dt) #basic default plot

#Each node shows: the predictions as follows:
#- the predicted class (legal status or not),
#- the predicted probability of legal status (true),
#- the percentage of observations in the node.
binary_model <- rpart.plot(dt, box.palette=c("pink", "palegreen3"), #change the colors of the fill
           branch.lty=3, shadow.col="gray", nn=TRUE,
           fallen.leaves=F,
           tweak = 2) #increase size

#saving the decision tree diagram
png("binary_model.png", width = 1500, height = 1000, res = 200) #start saving the file to jpeg
binary_model <- rpart.plot(dt, box.palette=c("pink", "palegreen3"), #change the colors of the fill
           branch.lty=3, shadow.col="gray", nn=TRUE,
           fallen.leaves=F,
           tweak = 2)
dev.off() #end saving
```

```r
# Variable importance is also informative:
dt$variable.importance
```

Predict Undocumented Legal Status for CPS 2016.

```r
#Use dt to predict cps16_mlearning data.frame legal status
dt_pred <- predict(dt, newdata = cps16_mlearning)
predict(dt, newdata = cps16_mlearning[130,]) #check first observation
summary(dt_pred) #review the predictions
table(dt_pred) #check data
```