ID5059 Lab 01 - fitting, predicting and Kaggle

C. Donovan & T. Kelsey

Feb 2019

Table of Contents

Introductory lab

The objectives for this first lab a quite modest.

- We seek to fit some models to a fairly simple dataset
- Make predictions from these models
- Pass these over to the Kaggle platform
- Do some basic plotting and model evaluation

The intention is to get you used to these processes for your first project, with the Kaggle part being important for the group project - which has a competitive part administered on the Kaggle platform. Also, if you're new to R or Python, then you'll get to flex these a little.

*blah

Note, this document has been produced in an R notebook (basically R markdown with some extra features). I recommend that you explore these sorts of things as they can make your analysis life easier as well as collaboration. In short it merges the coding and analysis bits with report writing (or webpages, presenations etc).

The titanic dataset and Kaggle

- kaggle sign up
- download data
- read in and explore
- fit models (example)
- check how well you're doing
- make predictions
- upload to kaggle and check results

Kaggle titanic

Kaggle Python Tutorial on Machine Learning Kaggle R Tutorial on Machine Learning

https://www.datacamp.com/courses/free-introduction-to-r
https://www.datacamp.com/courses/intro-to-python-for-data-science

```
#= load in some useful packages

  library(ggplot2) # for pretty graphs
  library(tidyverse) # lots of useful data manip tools
```

To read in data, you need to know what format it is in. The file suffix is usually a clue, but may be wrong or ambiguous. Here our data is

```
## load train data
sink <- read.csv("L06-train.csv", header=TRUE)
attach(sink)



## load test data
check <- read.csv("L06-test.csv", header=TRUE)
attach(check)



##### Classification tree
library(rattle)
library(rpart.plot)
library(RColorBrewer)
library(rpart)

## women and children first!
wacf.train <- rpart(Survived ~ Sex + Age, method="class", data=sink)

## plot the tree
plot(wacf.train, uniform=TRUE,
   main="Women and Children First")
text(wacf.train, use.n=TRUE, all=TRUE, cex=.8)

## the rpart plot is terrible, so get something better
fancyRpartPlot(wacf.train)

## learn a tree for a larger subset of covariates
CART.train <- rpart(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare,
method="class", data=sink)

## view details
summary(CART.train)

## plot the tree
```

```r
fancyRpartPlot(CART.train)

# prune the tree
pfit<- prune(CART.train,
    cp= CART.train$cptable[which.min(CART.train$cptable[,"xerror"]),"CP"])

# plot the pruned tree
fancyRpartPlot(pfit)


# missing step - impute missing covariate values in the test data
check3 <- read.csv("test-imputed.csv", header=TRUE)
attach(check3)
preds <- predict(pfit,check3)
pid <- subset(check3, select=PassengerId)
out <- cbind(pid,round(preds[,2],digits=0))
write.csv(out,"ToKaggleCART.csv", row.names=FALSE)
detach(check3)
```