Imputing excused missed tests

Evan Green

2/14/2017

```
library(igraph)
library(dplyr)
library(MASS)
library(randomForest)
library(RColorBrewer)
nice_colors <- brewer.pal(5,"Set1")
setwd("/Users/evangreen/Desktop/Senior\ Thesis/Grade\ Data")

grades <- read.csv("FullGrades.csv", as.is=T, strip.white = TRUE)
genders <- read.csv("CodeGender.csv", as.is = T, strip.white=TRUE)
grades <- merge(genders,grades,by.x = "code",by.y = "student")

mylist <- list()
my_files <- paste("CollabPS", 1:7, ".csv", sep = "")
names_vec <- paste("CollabPS", 1:7, sep = "")
raw_collabs <- lapply(my_files, read.csv, as.is = T,strip.white = TRUE)
names(raw_collabs) <- names_vec
```

```
#for this purpose we dont want anyone who has droppped
grades <- grades[!grades$did_drop_after4,]

grades$hw_mean <- rowMeans(grades[,grepl("hw",colnames(grades))])

#these missed tests were excused
held_out <- grades$test1 == 0

lm1 <- lm(test1 ~ . - code - hw6, subset = !held_out, data = grades)
summary(lm1)
lm2 <- stepAIC(lm1)
summary(lm2)

#I am seeing overfitting since hw6 should not have a negative coefficent
lm3 <- lm(test1 ~ hw1 + hw3 + hw5 +test2, subset = !held_out, data = grades)
summary(lm3)

lm3 <- lm(test1 ~  hw3 +test2, subset = !held_out, data = grades)
summary(lm3)

plot(lm3)
#test mean squared error
mean((predict(lm3) - grades$test1[!held_out]) ** 2)

hist(grades$test1[!held_out], breaks = 20)
hist(predict(lm3))


rf1 <- randomForest(test1 ~  hw3 +test2, subset = !held_out, data = grades)
```

```r
mean((predict(rf1) - grades$test1[!held_out]) ** 2)
plot(grades$test1[!held_out], grades$test1[!held_out] -predict(rf1) )

rf2 <- randomForest(test1 ~ . -code - Sex, subset = !held_out, data = grades)
mean((predict(rf2) - grades$test1[!held_out]) ** 2)
plot(grades$test1[!held_out], grades$test1[!held_out] -predict(rf2) )



#lm3 is the result that is the best

grades$test1[held_out] <- predict(lm3, newdata = grades[held_out,])
```

```r
#these missed tests were excused
held_out <- grades$test2 == 0

lm1 <- lm(test2 ~ . -code, subset = !held_out, data = grades)
summary(lm1)
lm2 <- stepAIC(lm1)
summary(lm2)

plot(lm2)


mean((predict(lm2) - grades$test2[!held_out]) ** 2)


rf1 <- randomForest(test2 ~  hw3 +hw1 +hw7, subset = !held_out, data = grades)
mean((predict(rf1) - grades$test2[!held_out]) ** 2)
plot(grades$test1[!held_out], grades$test2[!held_out] -predict(rf1) )

rf2 <- randomForest(test2 ~ . -code - Sex, subset = !held_out, data = grades)
mean((predict(rf2) - grades$test1[!held_out]) ** 2)
plot(grades$test1[!held_out], grades$test1[!held_out] -predict(rf2) )


predict(lm2, newdata = grades[held_out,])

grades$test2[held_out] <- predict(lm2, newdata = grades[held_out,])
```

```r
#don't overwrite
#Also should make it possible to tell which students are imputed, they aren't integers
setwd("/Users/evangreen/Desktop/Senior\ Thesis/Grade\ Data")
write.csv(grades,file = "Grades_Imputed.csv",row.names = F)
```