Subject: GBM variable importance
From: Mike Wurm <mjwurm@gmail.com>
Date: 3/23/15, 12:12
To: Fred Boehm <frederick.boehm@gmail.com>

Fred,
Here's the gbm variable importance list and the code. For HAC2004 and doc2004, use v2 because this excludes the 2011 covariates. I updated the 2004-2011 models to exclude subjects that already had an event before 2004.

Mike

```
require(caret)
require(car)
require(gbm)
require(tree)
require(rpart)
require(glmnet)
require(party)
require(randomForest)

## append missing value indicators to covariates "x"
dat <- read.csv("/mnt/DATA/Main/Spring 2015/998/Proj2/docs/WLS2.csv")
x1 <- dat[, !colnames(dat) %in% c("HA2011", "HA2004", "HAC2011", "HAC2004", "doc2011",
"doc2004")]
miss_pct <- sapply(colnames(x1), function(nam) mean(is.na(dat[,nam])))
x2 <- data.frame((is.na(x1[,miss_pct>.1])*1))
names(x2) <- paste0(names(x1[miss_pct>.1]), "_missid")
x <- data.frame(x1,x2)  # appended with a missing value indicator for each variable
id_11 <- grepl("2011", names(x))
x04 <- x[,!id_11]  # excluding 2011 covariates
id1_11 <- grepl("2011", names(x1))
x104 <- x1[,!id1_11]  # no missing value indicators, excluding 2011 covariates

## create cross validation folds for response variables
## Note gbm requires integer response in {0,1}
h11 <- Recode(dat$HAC2011, "'Yes'=1; 'No'=0", as.factor=F)
h04 <- Recode(dat$HAC2004, "'Yes'=1; 'No'=0", as.factor=F)
d11 <- Recode(dat$doc2011, "'Yes'=1; 'No'=0", as.factor=F)
d04 <- Recode(dat$doc2004, "'Yes'=1; 'No'=0", as.factor=F)
h0411 <- rep(0, length(h11))
h0411[is.na(h04+h11) | h04==1] <- NA
h0411[h04==0 & h11==1] <- 1
d0411 <- rep(0, length(d11))
d0411[is.na(d04+d11) | d04==1] <- NA
d0411[d04==0 & d11==1] <- 1

set.seed(9881829)
folds_h11_temp <- createFolds(which(!is.na(h11)), k = 5, list=T, returnTrain=F)
folds_h11 <- lapply(folds_h11_temp, function(x) which(!is.na(h11))[x])
folds_h04_temp <- createFolds(which(!is.na(h04)), k = 5, list=T, returnTrain=F)
folds_h04 <- lapply(folds_h04_temp, function(x) which(!is.na(h04))[x])
```

```
folds_h0411_temp <- createFolds(which(!is.na(h0411)), k = 5, list=T, returnTrain=F)
folds_h0411 <- lapply(folds_h0411_temp, function(x) which(!is.na(h0411))[x])
folds_d11_temp <- createFolds(which(!is.na(d11)), k = 5, list=T, returnTrain=F)
folds_d11 <- lapply(folds_d11_temp, function(x) which(!is.na(d11))[x])
folds_d04_temp <- createFolds(which(!is.na(d04)), k = 5, list=T, returnTrain=F)
folds_d04 <- lapply(folds_d04_temp, function(x) which(!is.na(d04))[x])
folds_d0411_temp <- createFolds(which(!is.na(d0411)), k = 5, list=T, returnTrain=F)
folds_d0411 <- lapply(folds_d0411_temp, function(x) which(!is.na(d0411))[x])
rm(folds_h11_temp, folds_h04_temp, folds_h0411_temp, folds_d11_temp, folds_d04_temp,
folds_d0411_temp)

gbmInit <- function(y, x, folds, tree_inc=1000) {
  id <- Reduce(union, folds)
  xfit <- x[id,]
  yfit <- y[id]
  ntree <- tree_inc
  gbm_fit <- gbm.fit(xfit, yfit, distribution="adaboost", n.tree=tree_inc)
  while (gbm.perf(gbm_fit, method="OOB")==ntree) {
    ntree <- ntree + tree_inc
    gbm_fit <- gbm.more(gbm_fit, n.new.trees=tree_inc)
  }
  gbm_fit
}
# set.seed(3382624)
# gbm_h11 <- gbmInit(h11, x, folds_h11, 1000)
# gbm_h04 <- gbmInit(h04, x, folds_h04, 1000)
# gbm_d11 <- gbmInit(d11, x, folds_d11, 1000)
# gbm_d04 <- gbmInit(d04, x, folds_d04, 1000)
# set.seed(858874)
# gbm_h0411 <- gbmInit(h0411, x04, folds_h0411, 1000)
# gbm_d0411 <- gbmInit(d0411, x04, folds_d0411, 1000)
# set.seed(1909459)
# gbm_h04v2 <- gbmInit(h04, x04, folds_h04, 1000)
# gbm_d04v2 <- gbmInit(d04, x04, folds_d04, 1000)
# save(gbm_h11, gbm_h04, gbm_d11, gbm_d04, gbm_h0411, gbm_d0411, gbm_h04v2, gbm_d04v2,
#      file="/mnt/DATA/Main/Spring 2015/998/Proj2/models/gbm-initial.rda")
load("/mnt/DATA/Main/Spring 2015/998/Proj2/models/gbm-initial.rda")

best_h11 <- gbm.perf(gbm_h11, method="OOB")
best_h04 <- gbm.perf(gbm_h04, method="OOB")
best_d11 <- gbm.perf(gbm_d11, method="OOB")
best_d04 <- gbm.perf(gbm_d04, method="OOB")
best_h0411 <- gbm.perf(gbm_h0411, method="OOB")
best_d0411 <- gbm.perf(gbm_d0411, method="OOB")
best_h04v2 <- gbm.perf(gbm_h04v2, method="OOB")
best_d04v2 <- gbm.perf(gbm_d04v2, method="OOB")

summary(gbm_h11, n.trees=best_h11)[summary(gbm_h11, n.trees=best_h11)$rel.inf>0,]
summary(gbm_h04, n.trees=best_h04)[summary(gbm_h04, n.trees=best_h04)$rel.inf>0,]
summary(gbm_d11, n.trees=best_d11)[summary(gbm_d11, n.trees=best_d11)$rel.inf>0,]
summary(gbm_d04, n.trees=best_d04)[summary(gbm_d04, n.trees=best_d04)$rel.inf>0,]
summary(gbm_h0411, n.trees=best_h0411)[summary(gbm_h0411, n.trees=best_h0411)$rel.inf>0,]
summary(gbm_d0411, n.trees=best_d0411)[summary(gbm_d0411, n.trees=best_d0411)$rel.inf>0,]
summary(gbm_h04v2, n.trees=best_h04v2)[summary(gbm_h04v2, n.trees=best_h04v2)$rel.inf>0,]
summary(gbm_d04v2, n.trees=best_d04v2)[summary(gbm_d04v2, n.trees=best_d04v2)$rel.inf>0,]
```

Attachments:

gbm-inital.xlsx	26.4 KB