DMC@ISU: Iowa State University Data Mining Cup Team 2015

Initial Exploration

Spring 2015, Iowa State

Due Date: April 25 2015

I am using the following packages:

```
library(ggplot2)
library(lubridate)
library(dplyr)
library(reshape2)
library(sqldf)
```

and my working directory is set to `dmc2015/ian/features/feature_files/R/`.

0.1 Reading the Data

This file updates the feature matrix `featureMatrix_v0`. I am adding features Pete created:

```
# read in the training set
featureMatrix <- readRDS("~/dmc2015/data/featureMatrix/featMat_v0.0.rds")

# split the list feature matrix
trn <- featureMatrix$train
cls <- featureMatrix$class

# get the raw training and test set too for
# reference
trn.raw <- read.delim("~/dmc2015/data/raw_data/DMC_2015_orders_train.txt",
    stringsAsFactors = FALSE, sep = "|", quote = "")
cls.raw <- read.delim("~/dmc2015/data/raw_data/DMC_2015_orders_class.txt",
    stringsAsFactors = FALSE, sep = "|", quote = "")
```

0.2 Reading the Features

Pete stored his features in files that can be read into R using the following:

```
# Training and test features of coupon counts
nCoupTrain <- read.csv("../nCoupTrain.csv")
nCoupClass <- read.csv("../nCoupClass.csv")

# NAs should be 0
for (i in 1:nrow(nCoupTrain)) {
    for (j in 1:ncol(nCoupTrain)) {
        if (is.na(nCoupTrain[i, j]))
            nCoupTrain[i, j] <- 0
    }
}
```

```
# NAs should be 0
for (i in 1:nrow(nCoupClass)) {
    for (j in 1:ncol(nCoupClass)) {
        if (is.na(nCoupClass[i, j]))
            nCoupClass[i, j] <- 0
    }
}
```

Add the batch features:

```
# It's so easy with dplyr
trn <- trn %>% left_join(nCoupTrain, by = "orderID")
cls <- cls %>% left_join(nCoupClass, by = "orderID")
```

61 Features!

0.3 Writing the Feature Matrix

We can save the features as CSV files and R objects (using `saveRDS` and writing the training and test sets as lists):

```
write.csv(trn, file = "~/dmc2015/data/featureMatrix/train_ver1.0.csv",
    row.names = FALSE, na = "", quote = FALSE)
write.csv(cls, file = "~/dmc2015/data/featureMatrix/class_ver1.0.csv",
    row.names = FALSE, na = "", quote = FALSE)

featMat <- list(train = trn, class = cls)
saveRDS(featMat, file = "~/dmc2015/data/featureMatrix/featMat_v1.0.rds")
```