Project 3

Keran Li, Mingming Liu, Zhongxing Xue, Yuhan Zha, Junkai Zhang

March 19, 2018

Step 0: Prepare packages

```
# Load the libraries
packages.used <- c("EBImage","gbm","caret","nnet","xgboost","xgboost",
                   "haven","tidyverse","mlr","plyr","ggplot2","randomForest",
                   "adabag","neuralnet","grDevices","kernlab","tree")

#if(!require("EBImage")){
#  source("https://bioconductor.org/biocLite.R")
#  biocLite("EBImage")
#}

# Check packages
packages.needed=setdiff(packages.used,
                        intersect(installed.packages()[,1],
                                  packages.used))
# install needed packages
if(length(packages.needed)>0){
  install.packages(packages.needed, dependencies = TRUE)
}

library("EBImage")
```

```
## Warning: package 'EBImage' was built under R version 3.4.3
```

```
library("gbm")
```

```
## Loading required package: survival
```

```
## Loading required package: lattice
```

```
## Loading required package: splines
```

```
## Loading required package: parallel
```

```
## Loaded gbm 2.1.3
```

```
library("kernlab")
library("tree")
```

```
## Warning: package 'tree' was built under R version 3.4.4
```

```
library("caret")
```

```
## Warning: package 'caret' was built under R version 3.4.3
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:kernlab':
##
##     alpha
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:survival':
##
##     cluster
```

```
library("nnet")
library("xgboost")
```

```
## Warning: package 'xgboost' was built under R version 3.4.3
```

```
library("haven")
```

```
## Warning: package 'haven' was built under R version 3.4.3
```

```
library("tidyverse")
```

```
## ── Attaching packages ─────────────────────────────────────── tidyve
rse 1.2.1 ──
```

```
## ✔ tibble  1.4.2     ✔ purrr   0.2.4
## ✔ tidyr   0.8.0     ✔ dplyr   0.7.4
## ✔ readr   1.1.1     ✔ stringr 1.3.0
## ✔ tibble  1.4.2     ✔ forcats 0.3.0
```

```
## Warning: package 'tibble' was built under R version 3.4.3
```

```
## Warning: package 'tidyr' was built under R version 3.4.3
```

```
## Warning: package 'stringr' was built under R version 3.4.3
```

```
## Warning: package 'forcats' was built under R version 3.4.3
```

```
## ── Conflicts ────────────────────────────────────── tidyverse_co
nflicts() ──
## ✖ ggplot2::alpha()   masks kernlab::alpha()
## ✖ dplyr::combine()   masks EBImage::combine()
## ✖ purrr::cross()     masks kernlab::cross()
## ✖ dplyr::filter()    masks stats::filter()
## ✖ dplyr::lag()       masks stats::lag()
## ✖ purrr::lift()      masks caret::lift()
## ✖ dplyr::slice()     masks xgboost::slice()
## ✖ purrr::transpose() masks EBImage::transpose()
```

```
library("mlr")
```

```
## Loading required package: ParamHelpers
```

```
##
## Attaching package: 'mlr'
```

```
## The following object is masked from 'package:caret':
##
##     train
```

```
library("plyr")
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr
:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following object is masked from 'package:purrr':
##
##     compact
```

```
library("ggplot2")
library("randomForest")
```

```
## Warning: package 'randomForest' was built under R version 3.4.4
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
## The following object is masked from 'package:EBImage':
##
##     combine
```

```
library("adabag")
```

```
## Warning: package 'adabag' was built under R version 3.4.3
```

```
## Loading required package: rpart
```

```
## Warning: package 'rpart' was built under R version 3.4.3
```

```
## Loading required package: foreach
```

```
## Warning: package 'foreach' was built under R version 3.4.3
```

```
##
## Attaching package: 'foreach'
```

```
## The following objects are masked from 'package:purrr':
##
##     accumulate, when
```

```
## Loading required package: doParallel
```

```
## Loading required package: iterators
```

```
## Warning: package 'iterators' was built under R version 3.4.3
```

```
library("neuralnet")
```

```
##
## Attaching package: 'neuralnet'
```

```
## The following object is masked from 'package:dplyr':
##
##     compute
```

```
library("grDevices")
```

Step 0: specify directories.

Set the working directory to the image folder. Specify the training and the testing set. For data without an independent test/validation set, you need to create your own testing data by random subsampling. In order to obain reproducible results, set.seed() whenever randomization is used.

```
#setwd("~/Desktop/Spring2018-Project3-Group3/")
```

```
experiment_dir <- "../data/"
img_train_dir <- paste(experiment_dir, "train/", sep="")
img_test_dir <- paste(experiment_dir, "test/", sep="")
```

Step 1: Set up controls for evaluation experiments.

In this chunk, we have a set of controls for the evaluation experiments.

(T/F) process features for training set (T/F) run evaluation on an independent test set (T/F) process features for test set

```
run.feature.train = TRUE
run.test = TRUE
run.feature.test = TRUE
run.feature.RGB <- FALSE
```

Set up controls indicating which model to run

```
run.xgboost = FALSE
run.adaboost = FALSE
run.tree = FALSE
run.gbm = TRUE
run.lg = FALSE
run.nnt = FALSE
run.rf = FALSE
run.svm = FALSE
```

Step 2: Extract features

```
source("../lib/SelectFeature.R")

# tm_feature_train <- NA
# if(run.feature.train){
#   tm_feature_train <- system.time(dat_train <- SelectFeature(SIFTname, SubGroup))
# }
# cat("Time for constructing training features=", tm_feature_train[1], "s \n")

#tm_feature_test <- NA
#if(run.feature.test){
#  tm_feature_test <- system.time(dat_test <-SelectFeature())
#}
# Running time
#cat("Time for constructing testing features=", tm_feature_test[1], "s \n")

#save(dat_train, file="./output/feature_train.RData")
#save(dat_test, file="./output/feature_test.RData")
if (run.feature.RGB)
{
  TotalImage <- 2100
  lottery <- sample(1:5, TotalImage, replace = TRUE)
  FeatureData <- SelectFeature(N = TotalImage)
  class_train <- c(rep(1,1000), rep(2, 1000), rep(3, 1000))
  class_train <- class_train[1:TotalImage]
  dat_train <- cbind(class_train[which(lottery != 1)], FeatureData[which(lottery != 1
),])
  colnames(dat_train)[1] <- "class"
  dat_test <-  FeatureData[which(lottery == 1),]
}
if (!run.feature.RGB)
{
  TotalImage <- 3000
  lottery <- sample(1:5, TotalImage, replace = TRUE)
  class_train <- c(rep(1,1000), rep(2, 1000), rep(3, 1000))
  dat_train <- read.csv("../lib/SIFT_test.csv")
  dat_train <- dat_train[,-1]
  dat_test <- dat_train[which(lottery == 1),]
  dat_train <- cbind(class_train[which(lottery != 1)], dat_train[which(lottery != 1),
])
  colnames(dat_train)[1] <- "class"
}
```

Step 3: Import training and test data of best features.

```
#train_all_data <- rbind(tr_data, te_data)
#train_all_class <- rbind(tr_class, te_class)

# Training data
#dat_train <- cbind(train_all_class[,-1], train_all_data[,-1])


# testdata <- cbind(te_class[,-1], te_data[,-1])
# colnames(test)[1] <- "class"
# dim(dat_train)
```

Step 4: Train a classification model with training images

Call the train model and test model from library.

`train.R` and `test.R` should be wrappers for all your model training steps and your classification/prediction steps. + `train.R` + Input: a path that points to the training set features. + Input: an R object of training sample labels. + Output: an RData file that contains trained classifiers in the forms of R objects: models/settings/links to external trained configurations. + `test.R` + Input: a path that points to the test set features. + Input: an R object that contains a trained classifier. + Output: an R object of class label predictions on the test set. If there are multiple classifiers under evaluation, there should be multiple sets of label predictions.

```
source("../lib/train.R")
source("../lib/test.R")
```

- Train the model with the entire training set using the selected model.

```
if(run.xgboost){
  dat_train_xgb <- as.matrix(dat_train)
  xgboost_train <- train_xgboost(dat_train_xgb)
  xgbPred <- xgboost_test(xgboost_train$fit, as.matrix(dat_test))
  cat("Accuracy", mean(xgbPred == class_train[which(lottery == 1)]))
  cat("Time for testing model=", xgboost_train$time, "s \n")
}
if(run.adaboost){
  adaboost_train <- train_adaboosting(dat_train)
  cat("Time for testing model=", adaboost_train$time, "s \n")
}
if(run.tree){
  tree_train <- train_tree(dat_train)
  cat("Time for testing model=", tree_train$time, "s \n")
}
if(run.gbm){
  gbm_train <-  train_GBM(dat_train)
  gbmPred <- gbm_test(gbm_train$fit, dat_test)
  cat("Accuracy", mean(gbmPred == class_train[which(lottery == 1)]))
  cat("Time for testing model=", gbm_train$time, "s \n")
}
```

```
## Using test method...
## Accuracy 0.6504854Time for testing model= 1.290544 s
```

```
if(run.lg){
  lg_train <- train_logistic(dat_train)
  cat("Time for testing model=", lg_train$time, "s \n")
}
if(run.nnt){
  nnt_train <- train_nn(dat_train)
  cat("Time for testing model=", nnt_train$time, "s \n")
}
if(run.rf){
  rf_train <- train_rf(dat_train)
  cat("Time for testing model=", rf_train$time, "s \n")
}
if(run.svm){
  svm_train <- train_svm(dat_train)
  cat("Time for testing model=", svm_train$time, "s \n")
}
```

Step 5: Make prediction and summarize Running Time

Feed the final training model with the completely holdout testing data. Prediction performance matters, so does the running times for constructing features and for training the model, especially when the computation resource is limited.

```
tm_xgboost_test=NA
tm_adaboost_test=NA
tm_tree_test=NA
tm_gbm_test=NA
tm_lg_test=NA
tm_nnt_test=NA
tm_rf_test=NA
tm_svm_test=NA

if(run.xgboost){
  load(file="../lib/XgBoost/XgBoost.RData")
  tm_xgboost_test <- system.time(xgboost.pred <- predict(model, as.matrix(dat_test))
+ 1)
  save(xgboost.pred, file="../output/xgboost.pred.RData")
  cat("Accuracy", mean(xgboost.pred == class_train[which(lottery == 1)]))
  cat("Time for testing model=", tm_xgboost_test[1], "s \n")
}
if(run.adaboost){
  load(file="../lib/AdaBoost/adaboost.RData")
  tm_adaboost_test <- system.time(ada.pred <- predict(adaall, dat_test))
  save(ada.pred , file="../output/ada.pred .RData")
  cat("Time for testing model=", tm_adaboost_test[1], "s \n")
}
if(run.tree){
  load(file="../lib/ClassficationTree/tree.RData")
  tm_tree_test <- system.time(tree.pred <- predict(tree, dat_test))
  save(tree.pred, file="../output/tree.pred.RData")
  cat("Time for testing model=", tm_tree_test[1], "s \n")
}
if(run.gbm){
  load(file="../lib/GBM/GBM.RData")
  tm_gbm_test <- system.time(gbm.pred <- predict(gbm_train$fit, dat_test))
  save(gbm.pred, file="../output/gbm.pred.RData")
  #cat("Accuracy", mean(gbm.pred == class_train[which(lottery == 1)]))
  cat("Time for testing model=", tm_gbm_test[1], "s \n")
}
```

```
## Using 19 trees...
## Time for testing model= 0.109 s
```

```
if(run.lg){
  load(file="../lib/Logistic/logistic.RData")
  tm_lg_test <- system.time(lg.pred <- predict(fit, dat_test))
  save(lg.pred, file="../output/lg.pred.RData")
  cat("Time for testing model=", tm_lg_test[1], "s \n")
}
if(run.nnt){
  load(file="../lib/NeuronNetwork/NeuronNetwork.RData")
  tm_nnt_test <- system.time(nnt.pred <- predict(pr.nn, dat_test))
  save(nnt.pred, file="../output/nnt.pred.RData")
  cat("Time for testing model=", tm_nnt_test[1], "s \n")
}
if(run.rf){
  load(file="../lib/RandomForest/randomforest.RData")
  tm_rf_test <- system.time(rf.pred <- predict(rf_all, dat_test, type="response"))
  save(rf.pred , file="../output/rf.pred .RData")
  cat("Time for testing model=", tm_rf_test[1], "s \n")
}
if(run.svm){
  load(file="../lib/SVM/svm.RData")
  tm_svm_test <- system.time(svm.pred <- predict(rad_svm_fit, dat_test))
  save(svm.pred , file="../output/svm.pred.RData")
  cat("Time for testing model=", tm_svm_test[1], "s \n")
}
```