Classification

Anisha Mittal

```r
library("lemon")  # Pretty printing of data frames
kint_print.data.frame <- lemon_print
library(mlbench) # Access the data
library(rpart) # For fitting classification trees
library(nnet) # For fitting multinomial logistic regression
```

```r
data("Satellite")

# Re-order class labels alphabetically and remove spacing
Satellite$classes <- gsub(" ", "_", Satellite$classes)
Satellite$classes <- factor(as.character(Satellite$classes))

# Rename classes column to y
colnames(Satellite)[37] <- "y"

# To have the same initial split
set.seed(22021)
N <- nrow(Satellite)
keep <- sample(1:N, 5000)
test <- setdiff(1:N, keep)
# Training & validation data
dat <- Satellite[keep,]
N_train <- nrow(dat)
# Testing data
dat_test <- Satellite[test,]


# Function to compute classification accuracy
class_acc <- function(y, yhat) {
  tab <- table(y, yhat)
  return( sum(diag(tab))/sum(tab))
  }

# just to identify the classifiers
classifiers <- c("class_tree", "Mlog_reg")

K <- 5 # set number of folds
R <- 350 # set number of replicates --- NOTE : could be slow
out <- matrix(NA,R,4)
colnames(out) <- c(classifiers, "best_fit", "test_fold_acc")
out <- as.data.frame(out)
```

```r
for ( r in 1:R ) {
  folds <- rep( 1:K, ceiling(N/K))
  folds <- sample(folds) # random permute
  folds <- folds[1:N_train] # ensure we got N_train data points
  for ( k in 1:K ) {
    train_fold <- which(folds != k)
    validation <- setdiff(1:N_train, train_fold)
    # fit classifiers on the training data
    # ------------------------------ classification tree
    fit_ct <- rpart(y ~ ., data = Satellite, subset = train_fold)

    # ------------------------------ logistic regression
    fit_Mlog <- multinom(y ~ ., data = Satellite, subset = train_fold)

    # Predict classification of the test data observations in the dropped fold

    # Classification Tree
    pred_ct <- predict(fit_ct, type = "class", newdata = dat[validation,])
    tab_ct <- table(dat$y[validation], pred_ct)
    out[r,1] <- class_acc(pred_ct, dat$y[validation])

    # Multinomial logistic Regression
    pred_Mlog <- predict(fit_Mlog, type = "class", newdata = dat[validation,])
    tab_Mlog <- table(dat$y[validation], pred_Mlog)
    out[r,2] <- class_acc(pred_Mlog, dat$y[validation])

    # Accuracy of each Classifier
    acc <- c(class_tree = out[r,1], Mlog = out[r,2] )

    # Find the best fit and fold accuracy on test data
    best <- names(which.max(acc))
    switch(best,
       class_tree = {
         predTestCt <- predict(fit_ct, type = "class", newdata = dat_test)
         tabTestCt <- table(dat$y[test], predTestCt)
         accBest <- sum(diag(tabTestCt))/sum(tabTestCt)
       },
       Mlog = {
         predTestLog <- predict(fit_Mlog, type = "class", newdata = dat_test)
         tabTestLog <- table(dat$y[test], predTestLog)
         accBest <- sum(diag(tabTestLog))/sum(tabTestLog)
         }
       )
    out[r,3]<- best
    out[r,4]<- accBest
    }

  print(r) # print iteration number
}
```

```r
# Check first 25 entries of out data frame
head(out,25)
```

class_tree	Mlog_reg	best_fit	test_fold_acc
0.8158698	0.8321465	Mlog	0.1989199
0.8026183	0.8106747	Mlog	0.1863186
0.7982018	0.8301698	Mlog	0.1962196
0.8202899	0.8038647	class_tree	0.1962196
0.8090452	0.8271357	Mlog	0.1953195
0.8152493	0.8435973	Mlog	0.1980198
0.8035892	0.7946162	class_tree	0.1881188
0.8093812	0.8053892	class_tree	0.1989199
0.8394816	0.8185444	class_tree	0.1917192
0.8324821	0.8130746	class_tree	0.1935194
0.7982283	0.7992126	Mlog	0.1989199
0.8066802	0.8522267	Mlog	0.1944194
0.8474576	0.8414756	class_tree	0.1935194
0.8138138	0.8248248	Mlog	0.1971197
0.8242972	0.8012048	class_tree	0.1935194
0.8089552	0.8079602	class_tree	0.1944194
0.8019608	0.8441176	Mlog	0.1971197
0.7898990	0.8313131	Mlog	0.1944194
0.8081918	0.8231768	Mlog	0.2025203
0.7983789	0.8145897	Mlog	0.1944194
0.7976767	0.8180058	Mlog	0.2007201
0.8179980	0.8109201	class_tree	0.1962196
0.8160804	0.8231156	Mlog	0.1980198
0.8000000	0.7721393	class_tree	0.1917192
0.7881526	0.8192771	Mlog	0.1962196

```r
## SUMMARY
# check out the error rate summary statistics
table(out[,3])
```

```
##
## class_tree       Mlog
##        104        246
```

```r
tapply(out[,4], out[,3], summary)
```

```
## $class_tree
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1854  0.1899  0.1926  0.1927  0.1953  0.1998
##
## $Mlog
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1863  0.1935  0.1962  0.1961  0.1989  0.2070
```

Logistic is better.

1.

When looked at the results for each iteration we see the accuracy for both the models pretty similar. Although, The model classification tree has higher count of best classifier than the multinomial logistic regression. Therefore, classification tree is a better model fit for this data.

2.

Looking at the summary statistics of the fold accuracy received from test data, we see that both the classification tree and multinomial logistic regression classifiers again have pretty similar summary statistics. Even though classification tree has higher 1st Qu, 3rd Qu, Median, Max, and Mean accuracy value than Multinomial logistic regression. But still both the models have equal minimum accuracy value.