dt_hw.R

alexaubrey

Wed May 23 18:55:49 2018

```
library(rpart)
library(caret)
```

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

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

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

```
## Warning in as.POSIXlt.POSIXct(Sys.time()): unknown timezone 'zone/tz/2018c.
## 1.0/zoneinfo/America/Chicago'
```

```
data <- read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/glass/glass.
data",
                 col.names=c("ID", "RI", "Na", "Mg", "Al","Si","K",
                             "Ca","Ba","Fe", "glass"))

#Convert class num to label
data$glass_name <- "NULL"
data[data$glass == 1,]$glass_name <- "building_windows_float_processed"
data[data$glass == 2,]$glass_name <- "building_windows_non_float_processed"
data[data$glass == 3,]$glass_name <- "vehicle_windows_float_processed"

#none of class 4 in this dataset
#data[data$glass == 4,]$glass_name <- "vehicle_windows_non_float_processed"
data[data$glass == 5,]$glass_name <- "containers"
data[data$glass == 6,]$glass_name <- "tableware"
data[data$glass == 7,]$glass_name <- "headlamps"
data$glass_name <- as.factor(data$glass_name)

#Drop ID and the class number before we changed it
data$ID <- NULL
data$glass <- NULL

#Train/Test Split
n <- dim(data)[1]

t1 <- sample(1:n, n*.8)
t2 <- setdiff(1:n, t1)

train <- subset(data[t1,])
test <- subset(data[t2,], select =- glass_name)

### 1. BASIC ###
fitControl <- trainControl(method='cv', number=10)

Grid <- expand.grid(cp=seq(0,0.1, 0.005))
#Run the training, needed to remove NAs
trained_tree <- train(glass_name ~ ., data=train, method='rpart',
                      trControl=fitControl, metric='Accuracy', maximize=TRUE, tuneGrid=G
rid)
trained_tree
```

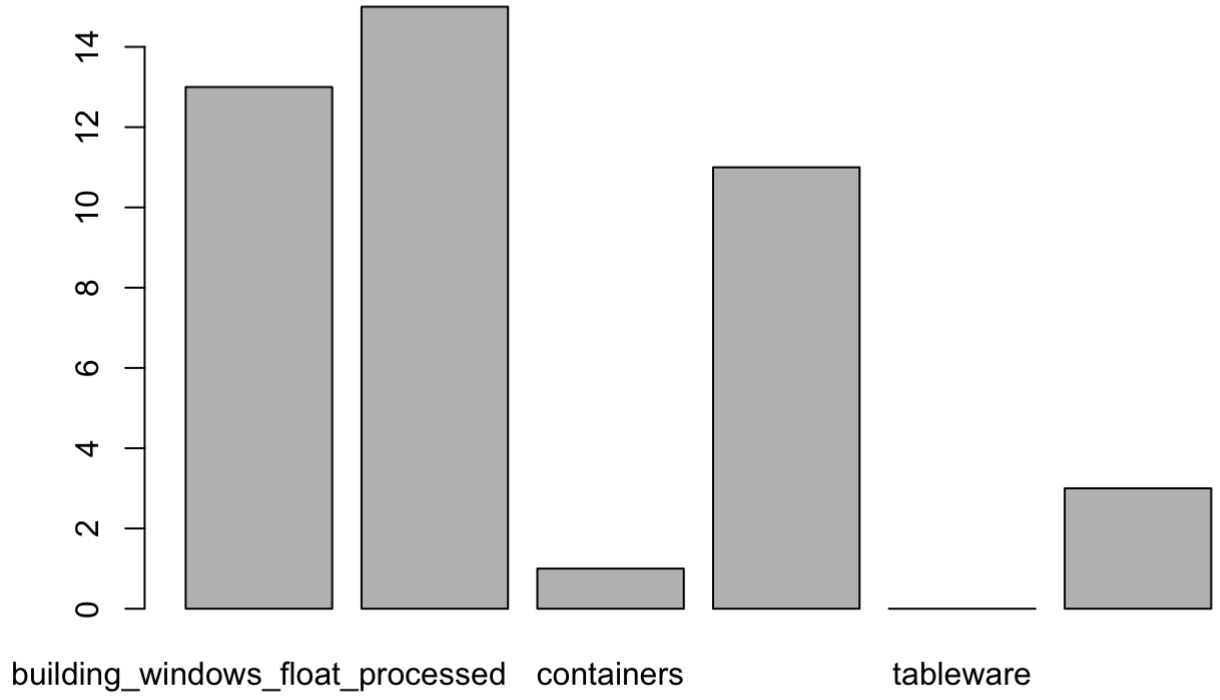
```
## CART
##
## 170 samples
##   9 predictor
##   6 classes: 'building_windows_float_processed', 'building_windows_non_float_processe
d', 'containers', 'headlamps', 'tableware', 'vehicle_windows_float_processed'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 153, 153, 154, 152, 155, 153, ...
## Resampling results across tuning parameters:
##
##   cp     Accuracy   Kappa
##   0.000  0.6703801  0.5450642
##   0.005  0.6703801  0.5450642
##   0.010  0.6703801  0.5450642
##   0.015  0.6703801  0.5450642
##   0.020  0.6703801  0.5450642
##   0.025  0.6693933  0.5404277
##   0.030  0.6635109  0.5315794
##   0.035  0.6635109  0.5288538
##   0.040  0.6579554  0.5212267
##   0.045  0.6697201  0.5348806
##   0.050  0.6697201  0.5348806
##   0.055  0.6586090  0.5151888
##   0.060  0.6586090  0.5151888
##   0.065  0.6356987  0.4746867
##   0.070  0.6356987  0.4746867
##   0.075  0.5882723  0.4038528
##   0.080  0.5882723  0.4038528
##   0.085  0.5816056  0.3895059
##   0.090  0.5816056  0.3895059
##   0.095  0.5757233  0.3805894
##   0.100  0.5757233  0.3805894
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.02.
```

```
pred <- predict(trained_tree, test, type="raw")
confusionMatrix(table(pred, data[t2,]$glass_name))
```

```
## Confusion Matrix and Statistics
##
##
## pred                                  building_windows_float_processed
##   building_windows_float_processed                                  11
##   building_windows_non_float_processed                               2
##   containers                                                         0
##   headlamps                                                          0
##   tableware                                                          0
##   vehicle_windows_float_processed                                    0
##
## pred                                  building_windows_non_float_processed
##   building_windows_float_processed                                     2
##   building_windows_non_float_processed                                11
##   containers                                                           0
##   headlamps                                                            0
##   tableware                                                            0
##   vehicle_windows_float_processed                                      0
##
## pred                                  containers headlamps tableware
##   building_windows_float_processed             0         0         0
##   building_windows_non_float_processed         0         0         0
##   containers                                   1         0         0
##   headlamps                                    1        10         0
##   tableware                                    0         0         0
##   vehicle_windows_float_processed              0         1         1
##
## pred                                  vehicle_windows_float_processed
##   building_windows_float_processed                                   0
##   building_windows_non_float_processed                               2
##   containers                                                         0
##   headlamps                                                          0
##   tableware                                                          0
##   vehicle_windows_float_processed                                    1
##
## Overall Statistics
##
##                Accuracy : 0.7907
##                  95% CI : (0.6396, 0.8996)
##     No Information Rate : 0.3023
##     P-Value [Acc > NIR] : 5.378e-11
##
##                   Kappa : 0.714
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: building_windows_float_processed
## Sensitivity                                           0.8462
## Specificity                                           0.9333
## Pos Pred Value                                        0.8462
## Neg Pred Value                                        0.9333
## Prevalence                                            0.3023
```

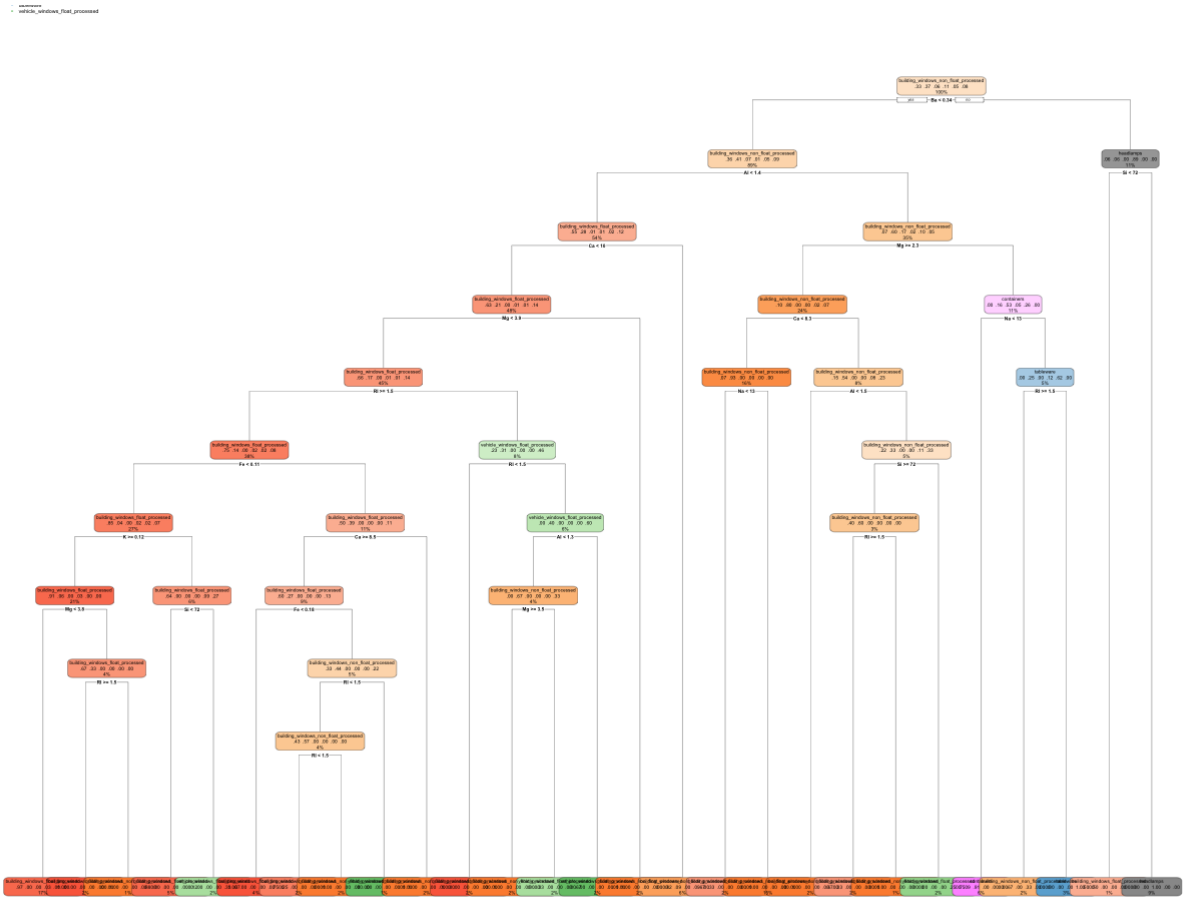
```
## Detection Rate                                        0.2558
## Detection Prevalence                                  0.3023
## Balanced Accuracy                                     0.8897
##                      Class: building_windows_non_float_processed
## Sensitivity                                               0.8462
## Specificity                                               0.8667
## Pos Pred Value                                            0.7333
## Neg Pred Value                                            0.9286
## Prevalence                                                0.3023
## Detection Rate                                            0.2558
## Detection Prevalence                                      0.3488
## Balanced Accuracy                                         0.8564
##                      Class: containers Class: headlamps Class: tableware
## Sensitivity                    0.50000           0.9091          0.00000
## Specificity                    1.00000           0.9688          1.00000
## Pos Pred Value                 1.00000           0.9091              NaN
## Neg Pred Value                 0.97619           0.9687          0.97674
## Prevalence                     0.04651           0.2558          0.02326
## Detection Rate                 0.02326           0.2326          0.00000
## Detection Prevalence           0.02326           0.2558          0.00000
## Balanced Accuracy              0.75000           0.9389          0.50000
##                      Class: vehicle_windows_float_processed
## Sensitivity                                         0.33333
## Specificity                                         0.95000
## Pos Pred Value                                      0.33333
## Neg Pred Value                                      0.95000
## Prevalence                                          0.06977
## Detection Rate                                      0.02326
## Detection Prevalence                                0.06977
## Balanced Accuracy                                   0.64167
```

```
plot(pred)
```




```
temp <- rpart.control(xval=10, minbucket = 2, minsplit = 4, cp = 0)
#Let's plot the tree with the rpart function, couldn't get fancy plot to work on mac
dfit <- rpart(glass_name ~ ., data=train, control=temp, cp=trained_tree)
rpart.plot::rpart.plot(dfit)
```

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```



```
### 2. Bagging functino ###
library(ipred)

### Find best nbagg checking from 1-70
acc <- NULL
for (i in 1:70) {
  baggedTree <- bagging(glass_name ~ ., data=train, nbagg=i)
  pred <- predict(baggedTree, test)

  #In case our test data has classes that aren't in the predicted
  u = union(pred, data[t2,]$glass_name)
  t = table(factor(pred,u), factor(data[t2,]$glass_name, u))
  c <- confusionMatrix(t)

  #c <- confusionMatrix(table(pred, data[t2,]$glass_name))
  acc[i] <- c$overall[1]
}

#What's the nbagg that gives us the max accuracy
nbagg <- which(max(acc) == acc, arr.ind = TRUE)[1]
baggedTree <- bagging(glass_name ~ ., data=train, nbagg=nbagg)
pred <- predict(baggedTree, test)
confusionMatrix(table(pred, data[t2,]$glass_name))
```

```
## Confusion Matrix and Statistics
##
##
## pred                                  building_windows_float_processed
##   building_windows_float_processed                                  12
##   building_windows_non_float_processed                               1
##   containers                                                         0
##   headlamps                                                          0
##   tableware                                                          0
##   vehicle_windows_float_processed                                    0
##
## pred                                  building_windows_non_float_processed
##   building_windows_float_processed                                     3
##   building_windows_non_float_processed                                10
##   containers                                                           0
##   headlamps                                                            0
##   tableware                                                            0
##   vehicle_windows_float_processed                                      0
##
## pred                                  containers headlamps tableware
##   building_windows_float_processed             0         0         0
##   building_windows_non_float_processed         0         1         0
##   containers                                   1         0         0
##   headlamps                                    1        10         0
##   tableware                                    0         0         1
##   vehicle_windows_float_processed              0         0         0
##
## pred                                  vehicle_windows_float_processed
##   building_windows_float_processed                                   1
##   building_windows_non_float_processed                               1
##   containers                                                         0
##   headlamps                                                          0
##   tableware                                                          0
##   vehicle_windows_float_processed                                    1
##
## Overall Statistics
##
##                Accuracy : 0.814
##                  95% CI : (0.666, 0.9161)
##     No Information Rate : 0.3023
##     P-Value [Acc > NIR] : 5.895e-12
##
##                   Kappa : 0.7442
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: building_windows_float_processed
## Sensitivity                                           0.9231
## Specificity                                           0.8667
## Pos Pred Value                                        0.7500
## Neg Pred Value                                        0.9630
## Prevalence                                            0.3023
```

```
## Detection Rate                                        0.2791
## Detection Prevalence                                  0.3721
## Balanced Accuracy                                     0.8949
##                      Class: building_windows_non_float_processed
## Sensitivity                                               0.7692
## Specificity                                               0.9000
## Pos Pred Value                                            0.7692
## Neg Pred Value                                            0.9000
## Prevalence                                                0.3023
## Detection Rate                                            0.2326
## Detection Prevalence                                      0.3023
## Balanced Accuracy                                         0.8346
##                      Class: containers Class: headlamps Class: tableware
## Sensitivity                    0.50000           0.9091          1.00000
## Specificity                    1.00000           0.9688          1.00000
## Pos Pred Value                 1.00000           0.9091          1.00000
## Neg Pred Value                 0.97619           0.9687          1.00000
## Prevalence                     0.04651           0.2558          0.02326
## Detection Rate                 0.02326           0.2326          0.02326
## Detection Prevalence           0.02326           0.2558          0.02326
## Balanced Accuracy              0.75000           0.9389          1.00000
##                      Class: vehicle_windows_float_processed
## Sensitivity                                         0.33333
## Specificity                                         1.00000
## Pos Pred Value                                      1.00000
## Neg Pred Value                                      0.95238
## Prevalence                                          0.06977
## Detection Rate                                      0.02326
## Detection Prevalence                                0.02326
## Balanced Accuracy                                   0.66667
```

```
### 3. RandomForest ###
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:ggplot2':
##
##     margin
```

```
control <- trainControl(method="repeatedcv", number=4, repeats=3)
metric <- 'Accuracy'
n <- round(sqrt(ncol(train)))
tuneGrid <- expand.grid(.mtry=seq(1,n,1))
rf_default <- train(glass_name ~ ., data=train, method="rf",
                    metric=metric, tuneGrid=tuneGrid, trControl=control)
print(rf_default)
```

```
## Random Forest
##
## 170 samples
##   9 predictor
##   6 classes: 'building_windows_float_processed', 'building_windows_non_float_processe
d', 'containers', 'headlamps', 'tableware', 'vehicle_windows_float_processed'
##
## No pre-processing
## Resampling: Cross-Validated (4 fold, repeated 3 times)
## Summary of sample sizes: 128, 127, 127, 128, 129, 127, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##   1     0.7333306  0.6179739
##   2     0.7706355  0.6755539
##   3     0.7569707  0.6571110
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.
```

```
pred <- predict(rf_default, test)
confusionMatrix(table(pred, data[t2,]$glass_name))
```

```
## Confusion Matrix and Statistics
##
##
## pred                                  building_windows_float_processed
##   building_windows_float_processed                                  12
##   building_windows_non_float_processed                               1
##   containers                                                         0
##   headlamps                                                          0
##   tableware                                                          0
##   vehicle_windows_float_processed                                    0
##
## pred                                  building_windows_non_float_processed
##   building_windows_float_processed                                     2
##   building_windows_non_float_processed                                11
##   containers                                                           0
##   headlamps                                                            0
##   tableware                                                            0
##   vehicle_windows_float_processed                                      0
##
## pred                                  containers headlamps tableware
##   building_windows_float_processed             0         0         0
##   building_windows_non_float_processed         1         1         0
##   containers                                   0         0         0
##   headlamps                                    1        10         0
##   tableware                                    0         0         1
##   vehicle_windows_float_processed              0         0         0
##
## pred                                  vehicle_windows_float_processed
##   building_windows_float_processed                                   1
##   building_windows_non_float_processed                               1
##   containers                                                         0
##   headlamps                                                          0
##   tableware                                                          0
##   vehicle_windows_float_processed                                    1
##
## Overall Statistics
##
##                Accuracy : 0.814
##                  95% CI : (0.666, 0.9161)
##     No Information Rate : 0.3023
##     P-Value [Acc > NIR] : 5.895e-12
##
##                   Kappa : 0.7421
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: building_windows_float_processed
## Sensitivity                                           0.9231
## Specificity                                           0.9000
## Pos Pred Value                                        0.8000
## Neg Pred Value                                        0.9643
## Prevalence                                            0.3023
```

```
## Detection Rate                                        0.2791
## Detection Prevalence                                  0.3488
## Balanced Accuracy                                     0.9115
##                      Class: building_windows_non_float_processed
## Sensitivity                                               0.8462
## Specificity                                               0.8667
## Pos Pred Value                                            0.7333
## Neg Pred Value                                            0.9286
## Prevalence                                                0.3023
## Detection Rate                                            0.2558
## Detection Prevalence                                      0.3488
## Balanced Accuracy                                         0.8564
##                      Class: containers Class: headlamps Class: tableware
## Sensitivity                    0.00000           0.9091          1.00000
## Specificity                    1.00000           0.9688          1.00000
## Pos Pred Value                     NaN           0.9091          1.00000
## Neg Pred Value                 0.95349           0.9687          1.00000
## Prevalence                     0.04651           0.2558          0.02326
## Detection Rate                 0.00000           0.2326          0.02326
## Detection Prevalence           0.00000           0.2558          0.02326
## Balanced Accuracy              0.50000           0.9389          1.00000
##                      Class: vehicle_windows_float_processed
## Sensitivity                                         0.33333
## Specificity                                         1.00000
## Pos Pred Value                                      1.00000
## Neg Pred Value                                      0.95238
## Prevalence                                          0.06977
## Detection Rate                                      0.02326
## Detection Prevalence                                0.02326
## Balanced Accuracy                                   0.66667
```