Homework-5

Subbu Kandhaswamy

11/6/2019

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

# Homework Instructions - Last week we used Clustering and in this we are using the decision tree algorithm to solve the disputed essay problem.

## 1. Data Preparation

We’ll start by loading the necesary packages to work with data partitioning, model design, and graphics outputs.

require(caret)

## Loading required package: caret

## Loading required package: lattice

## Loading required package: ggplot2

require(dplyr)

## Loading required package: dplyr

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

require(ggplot2)  
require(rpart)

## Loading required package: rpart

require(rpart.plot)

## Loading required package: rpart.plot

require(e1071)

## Loading required package: e1071

Now loading the data. Since this is the same dataset as last week’s homework, we will skip summary, structure, and head.

dat <- read.csv("C:/Users/Administrator/Documents/MSDatascience/IST707/Week5/fedPapers85.csv")

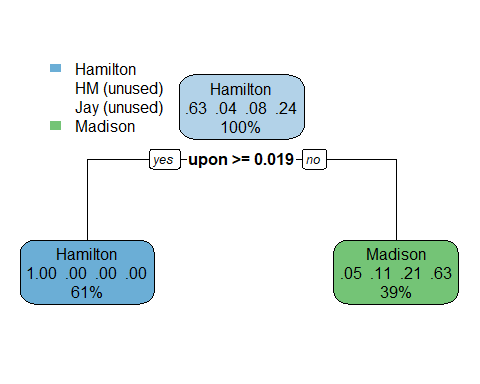
Then separate the dataset into 3 groups: Training, Testing & Verification set

Now extract the disputed papers from the set as we want to know if the model can predict to who these belong to. The training set will consist of 2/3 of the dataset after removing verification set; and the test set will be the remaining 1/3.

fed <- dat %>% filter(author != "dispt")  
fed$author <- as.factor(as.character(fed$author))  
  
# Randomly select 2/3 of the dataset - these will be the training set.  
split <- sample(nrow(fed), nrow(fed) \* 2/3)  
train <- fed[split, ]  
test <- fed[-split, ]  
  
# The verification set,  
ver <- dat %>% filter(author == "dispt")

## 2. Build and train the model.

# We'll create a first tree using rpart. and we exclude filename variable  
# considering the type (factor) and doesn't add much to the decision tree.  
feds.tree <- rpart(author ~ . - filename, data = train, method = "class", control = rpart.control(cp = 0))  
  
# Plotting to see how the model works vefore pruning  
rpart.plot(feds.tree)

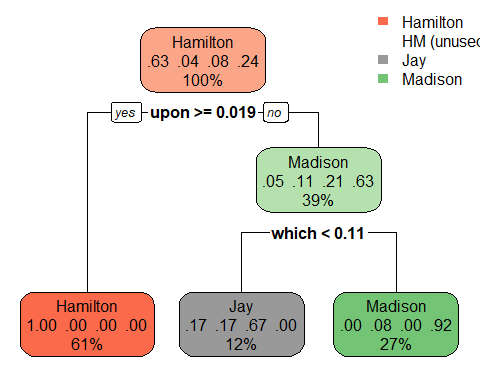


# And also look under the hood.  
summary(feds.tree)

## Call:  
## rpart(formula = author ~ . - filename, data = train, method = "class",   
## control = rpart.control(cp = 0))  
## n= 49   
##   
## CP nsplit rel error xerror xstd  
## 1 0.6111111 0 1.0000000 1.0000000 0.1874764  
## 2 0.0000000 1 0.3888889 0.4444444 0.1437363  
##   
## Variable importance  
## upon there on by and to   
## 25 19 16 15 13 12   
##   
## Node number 1: 49 observations, complexity param=0.6111111  
## predicted class=Hamilton expected loss=0.3673469 P(node) =1  
## class counts: 31 2 4 12  
## probabilities: 0.633 0.041 0.082 0.245   
## left son=2 (30 obs) right son=3 (19 obs)  
## Primary splits:  
## upon < 0.019 to the right, improve=15.725030, (0 missing)  
## there < 0.0115 to the right, improve=10.832480, (0 missing)  
## on < 0.081 to the left, improve= 9.768089, (0 missing)  
## by < 0.1385 to the left, improve= 7.345963, (0 missing)  
## to < 0.5315 to the right, improve= 7.228107, (0 missing)  
## Surrogate splits:  
## there < 0.0115 to the right, agree=0.898, adj=0.737, (0 split)  
## on < 0.0745 to the left, agree=0.857, adj=0.632, (0 split)  
## by < 0.1385 to the left, agree=0.837, adj=0.579, (0 split)  
## and < 0.4245 to the left, agree=0.816, adj=0.526, (0 split)  
## to < 0.5 to the right, agree=0.796, adj=0.474, (0 split)  
##   
## Node number 2: 30 observations  
## predicted class=Hamilton expected loss=0 P(node) =0.6122449  
## class counts: 30 0 0 0  
## probabilities: 1.000 0.000 0.000 0.000   
##   
## Node number 3: 19 observations  
## predicted class=Madison expected loss=0.3684211 P(node) =0.3877551  
## class counts: 1 2 4 12  
## probabilities: 0.053 0.105 0.211 0.632

Notice the model has a starting node as Hamilton - 71 percent of the texts in the training set belong to Hamilton. From there, the model asks whether the word ‘upon’ is used above 1.9 percent. If the ‘upon’ appears more frequently, then the model asigns the paper to Hamilton. But, if ‘upon’ is used less than 1.9 percent, then the model immediately accredit the text to Madison. Also notice that at this point in the model - no papers are assigned to Jay or to the dual Hamilton-Madison papers. The model will have to be tuned to fix this. Furthermore, when checking the model summary, we find that at two splits, the relative error of the model is 0.429. This is not desirable as we want a more accurate model.

# We'll set a minimum split of 10 papers in a bucket, and a max depth of 4  
# leaf nodes, and check how the model works from there.  
feds.tree2 <- rpart(author ~ . - filename, data = train, method = "class", control = rpart.control(cp = 0,   
 minsplit = 10, maxdepth = 4))  
  
rpart.plot(feds.tree2)

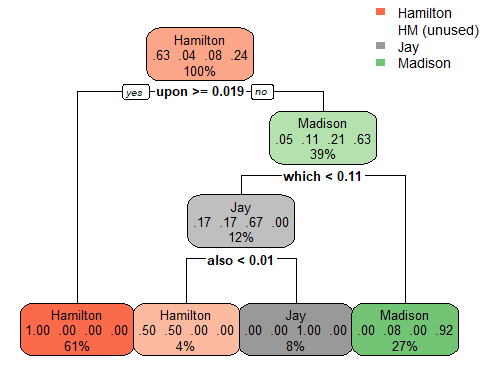


summary(feds.tree2)

## Call:  
## rpart(formula = author ~ . - filename, data = train, method = "class",   
## control = rpart.control(cp = 0, minsplit = 10, maxdepth = 4))  
## n= 49   
##   
## CP nsplit rel error xerror xstd  
## 1 0.6111111 0 1.0000000 1.0000000 0.1874764  
## 2 0.2222222 1 0.3888889 0.4444444 0.1437363  
## 3 0.0000000 2 0.1666667 0.3888889 0.1360828  
##   
## Variable importance  
## upon and there on by to which no when an been   
## 18 14 13 11 10 9 6 5 5 4 4   
##   
## Node number 1: 49 observations, complexity param=0.6111111  
## predicted class=Hamilton expected loss=0.3673469 P(node) =1  
## class counts: 31 2 4 12  
## probabilities: 0.633 0.041 0.082 0.245   
## left son=2 (30 obs) right son=3 (19 obs)  
## Primary splits:  
## upon < 0.019 to the right, improve=15.725030, (0 missing)  
## there < 0.0115 to the right, improve=10.832480, (0 missing)  
## on < 0.081 to the left, improve= 9.768089, (0 missing)  
## by < 0.1385 to the left, improve= 7.345963, (0 missing)  
## to < 0.5315 to the right, improve= 7.228107, (0 missing)  
## Surrogate splits:  
## there < 0.0115 to the right, agree=0.898, adj=0.737, (0 split)  
## on < 0.0745 to the left, agree=0.857, adj=0.632, (0 split)  
## by < 0.1385 to the left, agree=0.837, adj=0.579, (0 split)  
## and < 0.4245 to the left, agree=0.816, adj=0.526, (0 split)  
## to < 0.5 to the right, agree=0.796, adj=0.474, (0 split)  
##   
## Node number 2: 30 observations  
## predicted class=Hamilton expected loss=0 P(node) =0.6122449  
## class counts: 30 0 0 0  
## probabilities: 1.000 0.000 0.000 0.000   
##   
## Node number 3: 19 observations, complexity param=0.2222222  
## predicted class=Madison expected loss=0.3684211 P(node) =0.3877551  
## class counts: 1 2 4 12  
## probabilities: 0.053 0.105 0.211 0.632   
## left son=6 (6 obs) right son=7 (13 obs)  
## Primary splits:  
## which < 0.113 to the left, improve=5.469636, (0 missing)  
## and < 0.6235 to the right, improve=5.249123, (0 missing)  
## of < 0.7465 to the left, improve=5.249123, (0 missing)  
## the < 1.098 to the left, improve=5.249123, (0 missing)  
## no < 0.0175 to the left, improve=3.944361, (0 missing)  
## Surrogate splits:  
## no < 0.0175 to the left, agree=0.947, adj=0.833, (0 split)  
## when < 0.0145 to the right, agree=0.947, adj=0.833, (0 split)  
## an < 0.043 to the left, agree=0.895, adj=0.667, (0 split)  
## and < 0.6235 to the right, agree=0.895, adj=0.667, (0 split)  
## been < 0.0275 to the left, agree=0.895, adj=0.667, (0 split)  
##   
## Node number 6: 6 observations  
## predicted class=Jay expected loss=0.3333333 P(node) =0.122449  
## class counts: 1 1 4 0  
## probabilities: 0.167 0.167 0.667 0.000   
##   
## Node number 7: 13 observations  
## predicted class=Madison expected loss=0.07692308 P(node) =0.2653061  
## class counts: 0 1 0 12  
## probabilities: 0.000 0.077 0.000 0.923

We can see that this tree is much more accurate, producing a relative error of 0.143 at two splits. We also see that the tree further elaborates beyond the Madison leaf node, going so far as to identify the Jay papers. Returning to the Madison leaf node, if the author uses ‘been’ less than 5.3 percent of the time, then it is a Jay paper. Anything more is still attributed to Madison. Unfortunately, the model is still failing to determine the Hamilton-Madison papers, and we see that it’s struggling to correctly identify the Jay papers, as only 67 percent of the papers attributed to Jay are in fact his.

# Let's reduce the minimum split to five, while also increasing the max  
# depth to five.  
feds.tree3 <- rpart(author ~ . - filename, data = train, method = "class", control = rpart.control(cp = 0,   
 minsplit = 5, maxdepth = 5))  
  
rpart.plot(feds.tree3, cex = 0.8)



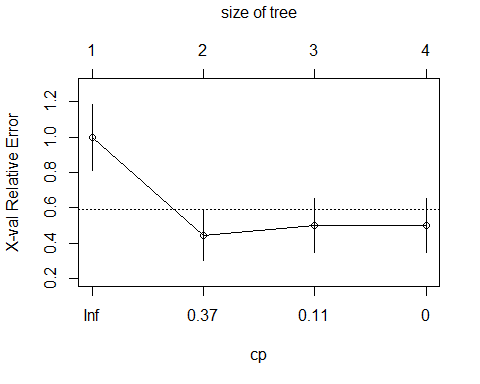
summary(feds.tree3)

## Call:  
## rpart(formula = author ~ . - filename, data = train, method = "class",   
## control = rpart.control(cp = 0, minsplit = 5, maxdepth = 5))  
## n= 49   
##   
## CP nsplit rel error xerror xstd  
## 1 0.61111111 0 1.0000000 1.0000000 0.1874764  
## 2 0.22222222 1 0.3888889 0.4444444 0.1437363  
## 3 0.05555556 2 0.1666667 0.5000000 0.1505847  
## 4 0.00000000 3 0.1111111 0.5000000 0.1505847  
##   
## Variable importance  
## upon and there on by to which no when an been also   
## 16 14 12 10 9 7 5 5 5 4 4 2   
## but if. it not   
## 2 2 2 2   
##   
## Node number 1: 49 observations, complexity param=0.6111111  
## predicted class=Hamilton expected loss=0.3673469 P(node) =1  
## class counts: 31 2 4 12  
## probabilities: 0.633 0.041 0.082 0.245   
## left son=2 (30 obs) right son=3 (19 obs)  
## Primary splits:  
## upon < 0.019 to the right, improve=15.725030, (0 missing)  
## there < 0.0115 to the right, improve=10.832480, (0 missing)  
## on < 0.081 to the left, improve= 9.768089, (0 missing)  
## by < 0.1385 to the left, improve= 7.345963, (0 missing)  
## to < 0.5315 to the right, improve= 7.228107, (0 missing)  
## Surrogate splits:  
## there < 0.0115 to the right, agree=0.898, adj=0.737, (0 split)  
## on < 0.0745 to the left, agree=0.857, adj=0.632, (0 split)  
## by < 0.1385 to the left, agree=0.837, adj=0.579, (0 split)  
## and < 0.4245 to the left, agree=0.816, adj=0.526, (0 split)  
## to < 0.5 to the right, agree=0.796, adj=0.474, (0 split)  
##   
## Node number 2: 30 observations  
## predicted class=Hamilton expected loss=0 P(node) =0.6122449  
## class counts: 30 0 0 0  
## probabilities: 1.000 0.000 0.000 0.000   
##   
## Node number 3: 19 observations, complexity param=0.2222222  
## predicted class=Madison expected loss=0.3684211 P(node) =0.3877551  
## class counts: 1 2 4 12  
## probabilities: 0.053 0.105 0.211 0.632   
## left son=6 (6 obs) right son=7 (13 obs)  
## Primary splits:  
## which < 0.113 to the left, improve=5.469636, (0 missing)  
## and < 0.6235 to the right, improve=5.249123, (0 missing)  
## of < 0.7465 to the left, improve=5.249123, (0 missing)  
## the < 1.098 to the left, improve=5.249123, (0 missing)  
## no < 0.0175 to the left, improve=3.944361, (0 missing)  
## Surrogate splits:  
## no < 0.0175 to the left, agree=0.947, adj=0.833, (0 split)  
## when < 0.0145 to the right, agree=0.947, adj=0.833, (0 split)  
## an < 0.043 to the left, agree=0.895, adj=0.667, (0 split)  
## and < 0.6235 to the right, agree=0.895, adj=0.667, (0 split)  
## been < 0.0275 to the left, agree=0.895, adj=0.667, (0 split)  
##   
## Node number 6: 6 observations, complexity param=0.05555556  
## predicted class=Jay expected loss=0.3333333 P(node) =0.122449  
## class counts: 1 1 4 0  
## probabilities: 0.167 0.167 0.667 0.000   
## left son=12 (2 obs) right son=13 (4 obs)  
## Primary splits:  
## also < 0.01 to the left, improve=2, (0 missing)  
## and < 0.5945 to the left, improve=2, (0 missing)  
## but < 0.041 to the left, improve=2, (0 missing)  
## if. < 0.013 to the left, improve=2, (0 missing)  
## it < 0.153 to the left, improve=2, (0 missing)  
## Surrogate splits:  
## and < 0.5945 to the left, agree=1, adj=1, (0 split)  
## but < 0.041 to the left, agree=1, adj=1, (0 split)  
## if. < 0.013 to the left, agree=1, adj=1, (0 split)  
## it < 0.153 to the left, agree=1, adj=1, (0 split)  
## not < 0.0695 to the left, agree=1, adj=1, (0 split)  
##   
## Node number 7: 13 observations  
## predicted class=Madison expected loss=0.07692308 P(node) =0.2653061  
## class counts: 0 1 0 12  
## probabilities: 0.000 0.077 0.000 0.923   
##   
## Node number 12: 2 observations  
## predicted class=Hamilton expected loss=0.5 P(node) =0.04081633  
## class counts: 1 1 0 0  
## probabilities: 0.500 0.500 0.000 0.000   
##   
## Node number 13: 4 observations  
## predicted class=Jay expected loss=0 P(node) =0.08163265  
## class counts: 0 0 4 0  
## probabilities: 0.000 0.000 1.000 0.000

This model gives us a relative error of 0.071, which is amamzing. Most of the papers are being assigned correctly to Hamilton, Jay, and Madison. However, at the first Jay node, if the word ‘also’ appears less than one percent of the time, the model assigns the text to Hamilton with a 50 percent probability of getting it right. We can assume that both Hamilton’s and the Hamilton Madison papers are being classified here.

Let’s check the cp plot and see how the model works on the test set.

plotcp(feds.tree3)



# Now we'll use this model to predict the test set.  
test\_pred <- data.frame(predict(feds.tree3, newdata = test))  
  
# Some data transformation needs to be done to check the confusion matrix.  
results <- test\_pred %>% mutate(results = ifelse(Madison == 1, "Madison", ifelse(Hamilton ==   
 1, "Hamilton", ifelse(Jay == 1, "Jay", "HM"))))  
# Lets introduce some bias here to explicitly statethat if none of the rules  
# apply to the results, then the result should be HM This will cause some  
# texts which are 50-50 to be classified as HM.  
  
row.names(test) <- NULL  
testResult <- test %>% bind\_cols(results)  
testResult$results <- as.factor(testResult$results)  
  
confusionMatrix(testResult$result, testResult$author)

## Warning in levels(reference) != levels(data): longer object length is not a  
## multiple of shorter object length

## Warning in confusionMatrix.default(testResult$result, testResult$author):  
## Levels are not in the same order for reference and data. Refactoring data  
## to match.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Hamilton HM Jay Madison  
## Hamilton 20 0 0 0  
## HM 0 1 0 3  
## Jay 0 0 1 0  
## Madison 0 0 0 0  
##   
## Overall Statistics  
##   
## Accuracy : 0.88   
## 95% CI : (0.6878, 0.9745)  
## No Information Rate : 0.8   
## P-Value [Acc > NIR] : 0.234   
##   
## Kappa : 0.6591   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: Hamilton Class: HM Class: Jay Class: Madison  
## Sensitivity 1.0 1.0000 1.00 0.00  
## Specificity 1.0 0.8750 1.00 1.00  
## Pos Pred Value 1.0 0.2500 1.00 NaN  
## Neg Pred Value 1.0 1.0000 1.00 0.88  
## Prevalence 0.8 0.0400 0.04 0.12  
## Detection Rate 0.8 0.0400 0.04 0.00  
## Detection Prevalence 0.8 0.1600 0.04 0.00  
## Balanced Accuracy 1.0 0.9375 1.00 0.50

Reviewing the confusion matrix, we could confirm that that the model correctly assigns the Hamilt, Jay , and Madison texts to each author. Also confirms the issue lies with the HM papers (HM or disputed papers).

## 3. Prediction

But, we have one more test to actually see how the model fares: can the model predict who the disputed texts belong to?

ver\_pred <- predict(feds.tree3, newdata = ver)  
  
# We'll skip the confusion matrix in this case and just check the results.  
ver\_pred

## Hamilton HM Jay Madison  
## 1 0.0 0.07692308 0 0.9230769  
## 2 0.0 0.07692308 0 0.9230769  
## 3 0.5 0.50000000 0 0.0000000  
## 4 0.0 0.07692308 0 0.9230769  
## 5 0.0 0.07692308 0 0.9230769  
## 6 0.0 0.07692308 0 0.9230769  
## 7 0.0 0.07692308 0 0.9230769  
## 8 0.0 0.07692308 0 0.9230769  
## 9 0.0 0.07692308 0 0.9230769  
## 10 0.0 0.07692308 0 0.9230769  
## 11 0.0 0.07692308 0 0.9230769

So Out of the 11 disputed papers, the model classifies 3 as authored by Madison, one by Jay, and the remaining as a 50/50 chance being authored either by Hamilton or Hamilton & Madison. In comparison, last week using clustering technique was less effective, where there was less evidence to determine the authors where the average method showed that some of the papers were more likely to be authored by Hamilton, some by Madison, and a couple that couldn’t be determined since the language used by the authors were closely similar.