Alex\_Hyman\_HW5

Alex Hyman

10/29/2018

## Data Preparation

To conduct the decision tree analysis for the documents based on word frequency, the data needs to be split into a training set and a test set. The training and test sets will only include the federalist papers where the author is known, and the paper is written by only one author. The training set will be 80% of the documents, and the test set will be the remaining 20% of the documents. I will also check to make sure that the random sample has at least 3 documents for each author in the training set, and at least on document for each author in the test set. Additionally, the disputed papers will be indexed into their own dataset for later classification.

This process resulted in 57 documents being designated in the training set: 41 documents by Alexander Hamilton, 4 papers by John Jay, and 12 papers by James Madison. The test set contained 14 of the federalist papers: 10 by Hamilton, 1 by Jay, and 3 by Madison. This sampling meets the requirements specified. The disputed data frame contains the 11 federalist papers that have a disputed author with the word frequencies and the file name. The unused “dispt” and “HM” author levels were dropped for the training.

Note: Because decision trees do not use any distance measures, the word frequency does not need to be normalized.

library(rpart)  
library(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

library(tidyr)  
library(rpart.plot)  
library(matrixStats)

##   
## Attaching package: 'matrixStats'

## The following object is masked from 'package:dplyr':  
##   
## count

library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

set.seed(111)  
#Reading in the file  
df <- read.csv("fedPapers85.csv")  
#Removing the HM and disputed papers from the traiing set  
train\_df <- df %>% filter(!author %in% c("HM","dispt")) %>%  
 sample\_frac(0.8)  
  
#Dropping the undefined author levels  
train\_df$author <- droplevels(train\_df$author)  
  
#Get the remaining for test set  
test\_df <- df[!df$filename %in% train\_df$filename, ] %>%   
 filter(!author %in% c("dispt", "HM")) %>%  
 select(-filename)  
  
#Dropping the undefined author levels  
test\_df$author <- droplevels(test\_df$author)  
#get rid of the file name in the training set  
train\_df <- train\_df %>% select(-filename)  
  
#Creating a test set with just the disputed texts  
dispt <- df %>% filter(author == "dispt")  
  
print("Train set Author Distribution")

## [1] "Train set Author Distribution"

print(table(train\_df$author))

##   
## Hamilton Jay Madison   
## 41 4 12

print("Test set Author Distribution")

## [1] "Test set Author Distribution"

print(table(test\_df$author))

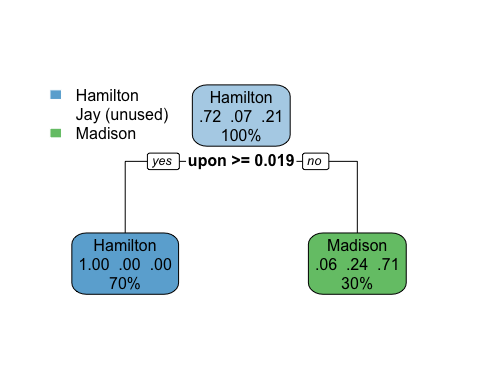
##   
## Hamilton Jay Madison   
## 10 1 3

## Build and Tune Decision Trees

Decision trees were then developed using the “rpart” and “caret” R packages. Two different decision trees were created; one with just the formula included, and another with some parameter adjustments. The decision trees will be evaluated by accuracy on the test set, and the most accurate decision tree will be used to determine which author wrote the disputed papers.

### Basic Decision Tree

The first decision tree was created with the default parameters, and used the word frequencies to predict the author. These settings determined that the root node was the only node required, and the decision was soley based on the frequency of the word “upon”. If the word “upon” made up 1.9% or more of the words in the paper, the author was determined ot be Hamilton. If the word “upon” was used less frequently, the model determined that the author was Madison. This model is very generic, and most likely underfit the data. One of the biggest flaws of this model is that there is no leaf node for papers by John Jay.



The test set was then used to evaluate the performance of the decision tree model. This model performed decently, with all of the papers by Hamilton being correctly classified, and all of the papers by Madison correctly predicted as being written by Madison. There was one misclassification in the sole paper by John Jay by the model. This was due to the model not having any Jay leaf nodes.

#Predict with decision tree 1  
dt1.pred <- predict(dt1, test\_df)  
#Table for the confusion matrix  
table(dt1.pred, test\_df$author)

##   
## dt1.pred Hamilton Jay Madison  
## Hamilton 10 0 0  
## Jay 0 0 0  
## Madison 0 1 3

### Decision Tree #2

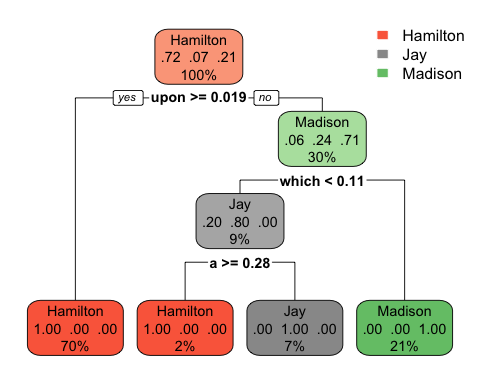
The biggest problem with the previous model was that there was not a leaf node for the papers by Jay. This was due to the default parameter that there has to be a minimum of 20 instances for a split to occur. The only author that had more than 20 papers was Hamilton, therefore the only decisions were made for papers by Hamilton. Madison was the default leaf node for the other decision tree because there were more papers by Madison than Jay in the training set. The newest version of the decision tree requires a minimum of 3 instances for a split to occur. This will now allow decisions to be made specific to papers by John Jay and James Madison.

This newest decision tree resulted in a tree with one root node, and two internal nodes. The frequent use of the word “upon” by Hamilton was still the first split made by the decision tree, and the lack of the use of the word “which” by Madison was the first internal node split. If the word “which” was used less than 1.1% of the time and upon was not used 1.9% of the time or more, then the classifer determined that the paper was written by Madison. If paper did not have frequent use if the word “upon”, but did have frequent usage of the word “which”, then the classifier looked at the usage of the word “a”. If the word “a” was used 2.8% of the time or more, then the classifier determined that the paper was written by Hamilton. If the word “a” was not used 2.8% of the time or more, then the classifier determined that the paper was written by John Jay.

set.seed(111)  
#changing the minimum split to be 3  
dt\_control <- rpart.control(minsplit = 3)  
#Predicting the test set  
dt\_prune <- train(author ~ ., data = train\_df, control = dt\_control,   
 method = "rpart", metric = "Accuracy")  
#confusion matrix of the decision tree  
table(train\_df$author, predict(dt\_prune$finalModel, train\_df, "class"))

##   
## Hamilton Jay Madison  
## Hamilton 41 0 0  
## Jay 0 4 0  
## Madison 0 0 12

#Plot of decision tree  
rpart.plot(dt\_prune$finalModel)



The new decision tree was evaluated on the test set, and it was 100% accuarate. Additionally, the decision tree was 100% accurate on the training set. This initially caused some worry that the model had been overfit, but the model evaluation determined that the decision tree was not overfit.

#Predictions of test set #2  
dt2.pred <- predict(dt\_prune$finalModel, test\_df, type = "class")  
#Confusion Matrix for decision tree 2  
table(dt2.pred, test\_df$author)

##   
## dt2.pred Hamilton Jay Madison  
## Hamilton 10 0 0  
## Jay 0 1 0  
## Madison 0 0 3

## Predict

Finally, a prediction was made for each of the disputed papers. All the the disputed papers were predicted to be written by James Madison, except for Federalist Paper 51. Federalist Paper 51 had used the word “upon” less frequently than 1.9%, but the word “a” had a high usage rate. Coincidentally, these results are the exact same as the clustering algorithm determined.

#Predicting the class of the disputed papers  
dispt.pred <- predict(dt\_prune$finalModel, dispt, type = "class")  
#adding the class to the disputed data frame  
dispt$prediction <- dispt.pred  
#Printing prediction and file name  
dispt[,c("prediction", "filename")]

## prediction filename  
## 1 Madison dispt\_fed\_49.txt  
## 2 Madison dispt\_fed\_50.txt  
## 3 Hamilton dispt\_fed\_51.txt  
## 4 Madison dispt\_fed\_52.txt  
## 5 Madison dispt\_fed\_53.txt  
## 6 Madison dispt\_fed\_54.txt  
## 7 Madison dispt\_fed\_55.txt  
## 8 Madison dispt\_fed\_56.txt  
## 9 Madison dispt\_fed\_57.txt  
## 10 Madison dispt\_fed\_62.txt  
## 11 Madison dispt\_fed\_63.txt

#Printing the number of predictions for each of the authors  
print(table(dispt.pred))

## dispt.pred  
## Hamilton Jay Madison   
## 1 0 10