2020-0701 IST 707 Data Analytics

Homework #5

Sathish Kumar Rajendiran

2020-08-14

Table of Contents

[Task Summary 2](#_Toc48434486)

[Dataset Preparation 3](#_Toc48434487)

[Decision Trees (DT) 6](#_Toc48434488)

[Prediction 10](#_Toc48434489)

[Other Analysis 12](#_Toc48434490)

[Summary 14](#_Toc48434491)

[Conclusion 14](#_Toc48434492)

## **Task Summary**

This dataset consists of data from 85 federalist essays appeared in New York newspapers during 1787 - 1788 published under the pen name “Publius”. These are still considered as one of the most important sources for interpreting and understanding the original intent of the Constitution. There were 3 authors Hamilton, Madison and Jay had written most of these essays. Hamilton and Madison co-authored 3, Author Jay had written 5, Madison has 15 and Hamilton wrote 51 of them. However, remaining 11 are considered as disputed and claims were made to prove ownership of those essays, however none remain conclusive.

In this dataset, the features are a set of “function words” and it’s based on the percentage of word occurrence in an essay. Let’s try to solve this mystery by using decision tree algorithm. First, build a DT model default setting, and then tune parameters to see if better model can be generated. In addition, compare these models using evaluation measures. Lastly, apply it to disputed essays to find out who wrote those, Hamilton or Madison? Task includes (but not limited to),

1. Dataset Preparation
   1. Data extraction
   2. Prepare Training and Test Data set
2. Build Decision Tree model
   1. Build DT model with default setting
   2. Tune DT model – Grow, Plot and Summarize
3. Prediction
   1. Run prediction on test data and disputed papers
4. Other Analysis
   1. Apply Random Forest
5. Summary
   1. Conclude on who wrote the disputed essays, Hamilton or Madison. Does the DT model reach the same conclusion as the clustering algorithms did?

## **Dataset Preparation**

In order to build decision tree models, first step is to understand the data and prepare training and test data sets. It involves the following steps,

* 1. Data extraction - import necessary packages for data extraction.

A screenshot of a social media post

Description automatically generated

fedPapers85.csv – has 85 rows and 72 columns. 85 rows represent all the essays by their respective author categories.

A screenshot of a cell phone

Description automatically generated A screen shot of a computer

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Preview and datatype of the dataset below,

A picture containing window, large, standing, flying

Description automatically generated

A close up of text on a white background

Description automatically generated

* 1. Prepare Training and Test data set – In this case, lets ignore the filename column and all the disputed rows to have a cleaner dataset with actual authors mapped. From this, training and test data set can be prepared by taking random split at 70:30 ratio. Try to split the data sets, such that all categories have values.

A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Having only 85 rows, the probability of splitting the train and test data set to include all combinations were minimal and 70:30 works better among other combinations.

## **Decision Trees (DT)**

Next step is to build and tune DT models

1. Build DT model with default setting

rpart library – from R Studio is used for building the decision trees. Initial case, involving the default parameters for the classification model on the training dataset. This process involves, building the tree, plot and summarize it with cross validation results.

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generated

From above, Decision tree with default settings,

* Function word “upon” is considered key word to split the tree into Hamilton vs Madison, upon > 0.015= resulting in root node error as 0.29412.
* Out of 51 samples from training set 35/36 mapped correctly to Hamilton’s writing
* 10 papers were mapped as Madison’s. However, 6 are not assigned correctly. Clearly, suggesting need for improvement in this model.
* CP split with relative error is 0.4

1. Tune Decision tree model with control measures (cp, minsplit and maxdepth)

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

From above, DT with additional parameters (cp=0, minsplit=3 and maxdepth=5)

* “upon” is considered key word to split the tree into Hamilton vs Madison, upon >= 0.015 resulting in root node error as 0.29412.
* There are additional branches with function words (no < 0.021), (an>=0.032), (a>=0.28).
* Out of 51 samples from training set
  + 36/36 mapped correctly to Hamilton’s writing
  + 10/10 mapped correctly to Madison’s writing
  + 2/2 mapped correctly to HM’s writing
  + 3/3 mapped correctly to Jay’s writing
* CP split with relative error is 0 with number of splits at 4
* Overall, all assignments made successfully.

1. Tune Decision tree model with control measures (cp, minsplit, maxdepth and minbucket)

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated

From above, DT with additional parameters (cp=0, minsplit=3, maxdepth=5 and minbucket = 1)

* “upon” is still considered key word to split the tree into Hamilton vs Madison, upon >= 0.015 resulting in root node error as 0.29412.
* There are additional branches with function words (no < 0.021), (an>=0.032), (a>=0.28).
* Out of 51 samples from training set
  + 36/36 mapped correctly to Hamilton’s writing
  + 10/10 mapped correctly to Madison’s writing
  + 2/2 mapped correctly to HM’s writing
  + 3/3 mapped correctly to Jay’s writing
* CP split with relative error is 0 with number of splits at 4
* Overall, all assignments made successfully.

## **Prediction**

With the latest DT models performing better with an accuracy, lets run it against test data.

1. Prediction on test data

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

From above, prediction on test data with trained tree model,

* + Model is predicting at 82.6%
  + Confusion matrix,
    1. confirming all of Hamilton’s mapped correctly.
    2. 1/2 of Jay’s mapped correctly and remaining mapped to Madison’s
    3. 3/5 of Madison’s mapped correctly. remaining mapped to Hamilton and HM’s respectively. May be showing influence on Hamilton’s writing in those.
    4. Only 1 of HM’s mapped to Madison – confirming influence on Madison’s writing on that paper.
    5. There is still improvement needed. However, with volume of split on training vs test data, 82.6% remains the best so far.

1. Prediction on disputed data – lets spin it against the disputed papers.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

From above, prediction on disputed data with trained tree model,

* + based on the confusion matrix,
    1. 8/11 mapped to Madison’s – high influence on his writing
    2. 2/11 mapped to HM – still confirming influence on Madison’s writing
    3. Only 1/11 is mapped to Jay’s – drawing conflict with the writing. May be Jay had influence over writing on that paper.
    4. Remember, training set had 0% error on the mapping. Unsure about overfitting, may be feature reduction could result better predictions

## **Other Analysis**

Let’s perform random forest algorithm on this dataset to help compare with DT models.

A screenshot of a cell phone

Description automatically generated

A picture containing bird, tree, flower

Description automatically generated

From above, random forest results

* + 10/11 disputed papers mapped to Madison’s, having clear winner here.
  + 1/11 goes to Hamilton’; still drawing confusion. May be with the usage of “upon”
  + All of Hamilton’s actual papers (51) mapped correctly to Hamilton
  + 3 out of 5 Jay’s mapped correctly; with 1 referred to Madison and other being disputed. Showing patterns of Madison’s writing in Jay’s papers.
  + All 3 HM’s mapped to Madison’s further confirming his influence over Hamilton’s
  + Feature importance highlights “upon”, “were”, “which” and few other words having influence over the predictions going towards Hamilton on the only paper that favored him. So, re-training DT models based on the feature importance might help get an accuracy closer to 90%.

A close up of a device

Description automatically generated

## **Summary**

* Based on DT models
  + Jay –possibly could have written one of the papers or could have influenced by Madison’s writing.
  + Hamilton – possibly did not write those essays. None of them mapped to Hamilton’s in the final predictions
  + HM (Co-authors) – likely possibility with 2/11 favoring HM’s. Mostly, an influence on having Madison’s writing.
  + Madison – high possibility with 8/11 aligning to Madison. In addition, other 2 also favoring Madison’s influence.
* Based on Random Forest model
  + Madison – most likely possibility with 10 out 11 papers confirming him
  + Hamilton – least likely with 1 out of 11 favoring him.
  + overall, suggesting Madison as clear winner.

## **Conclusion**

In comparison with clustering analysis involving K-means, Euclidean and Hierarchical clustering methods –

“None of these methods could not conclude any one author over other. However, strongly ruling out Jay being author of those disputed essays. Whereas favoring “Madison” as author having most influence with the words used. It is also possible that Hamilton & Madison could have co-authored adapting to Madison’s style of writing.”

DT models along with random forest – confirming Madison’s impact on majority of the papers. So, it can be determined that Madison as an author to those mystery papers with one or two exceptions on HM.

In addition, instead of just analyzing function words – may be sentence or phrases used in that essay analysis or feature reduction to include only those with higher importance could rule out the exceptions from the analysis.