IST 707 Applied Machine Learning

HW5: Use a Decision Tree to Solve a Mystery in History

In this homework assignment, you are going to use the decision tree algorithm to solve the disputed essay problem from last week.

Organize your report using the following template:

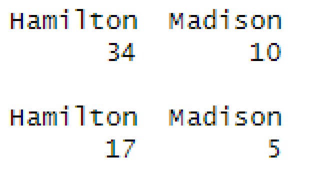
Section 1: Data preparation

You will need to separate the original data set to training and testing data for classification experiments. Describe the contents of your training and test data. What steps did you take?

Steps:

- **Open up the csv file as a data frame: fedPapers. When our dataset was loaded, I discovered that it contained data on 85 writers (72 words), 11 of whom were dispt values, or authors who were in dispute. Hamilton and Madison, as well as Jay and HM, were additional writers.**

- **I saw 'filename' column in our data frame that did not have any relevance to finding true author of these diputed essays. Therefore I removed them and only took author column and the words with frequency percentage.**
- **Since we only want to know if the disputed authors were Hamilton or Madison, (not Jay or HM) I filtered the data set with Hamilton and Madison as authors ; and the subset of data set that contained 'disp' as our final test dataset.**
- **The subset with Hamilton Madison is further partitioned as train.set and test.set data through createDataPartition function.**
- **I partitioned our Hamilton, Madison data with 0.66 split, creating partion (1/3rd for testing and 2/3rd for training)**



```
Hamilton  Madison 
      34       10 

Hamilton  Madison 
      17        5 
```

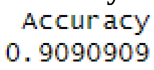
-
- **The above rows contain Hamilton and Madison count for train set and the below**

  Row contains count for test set. Train has 44 authors (H and M), Test.set has 22 authors.

Section 2: Build and tune decision tree models

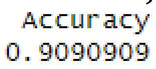
First, build a DT model using the default settings, then tune the parameters to see if a better model can be generated. Compare these models using appropriate evaluation measures. Describe and compare the patterns learned in these models.

- **1st: using default settings, I trained the model , tested accuracy on test data, then predicted final test data set.**
- **Accuracy of this method: 0.9090909**



```
 Accuracy 
0.9090909 
```

-
- **2nd method : using grid: by using appropriate values for minimum samples for a leaf in the DT, and pruning confidence numbers. Using Expand.grid method.**
- The final values used for the model were C = 0.01 and M = 2.
- **This model, when trained also had accuracy of 0.9090909.**



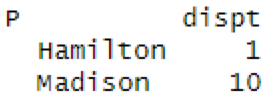
```
 Accuracy 
0.9090909 
```

-

Section 3: Prediction

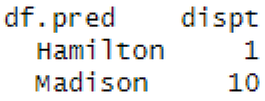
After building the classification model, apply it to the disputed papers to find out the authorship. Does the DT model reach the same conclusion the clustering algorithms did? Explain any differences.

- **For Default method: Once predicted the unknown authors, I found that our model predicted/recognized 10 out of 11 dispt values as Madison, and one Hamilton. This strongly suggests that the disputed essays were written by Madison.**



```
P           dispt
  Hamilton      1
  Madison      10
```

- **For Cross validation method: with accuracy of 0.9090909, this model also predicted 10 out of 11 dispt values as Madison, and one Hamilton. This strongly suggests that the disputed essays were written by Madison.**



```
df.pred    dispt
  Hamilton      1
  Madison      10
```

-
- **All in all I believe that the essays where written by Madison as both holdout and crossvalidation method suggests same.**
- **This result matches with the results of clustering algorithm where we also found that Madison was author of disputed essays.**

Provide your code in a separate script.