INTRODUCTION TO INTELLIGENT SYSTEMS

Lab 2

Wikipedia Language Classification

By: MRUDULA V

Date: 04/25/2018

INTRODUCTION

This project will be investigating the use of decision trees and boosted decision stumps to classify text as one of two languages. Our task was to collect data and train them so that a given 15 word segment of text from either English or dutch. Our code will depict the best of its ability which language the text is in.

FEATURES

I have utilized 10 features which were more leaning towards Dutch language, thus leading us to get "nl" even though there are some True values for English.

1. Contains English words: This feature was used to train the dataset to fetch the most recurring English words used in the file 'frequency_words_en'.
2. Contains Dutch words: This feature was used to train the dataset to fetch the most recurring Dutch words used in the file 'frequency_words_du'.
3. Has a special character: This feature trained the dataset to find the special characters from both the files and find their average number of occurrences in both with the range of 128-255 characters.
4. Length of the word: Gets the length of the word from each attribute of the training data.
5. Consists of the word the: Gets the count of word the and compares it between 2 languages.
6. Consists of double aa: Implementation of the word aa and its functions to find it true existence or not.
7. Consists of word her: The word her is counted and implemented with its count and we get to see which has a higher power with it-English or Dutch.
8. Consists of double kk: Implementation of the word aa and its functions to find it true existence or not.
9. Consists of word ly: The word ly is counted and implemented with its count and we get to see which has a higher power with it-English or Dutch.
10. Has triple vowels: All kinds of triple vowel examples are provided and are randomly chosen to get the most used in the languages to get the appropriate prediction.

DECISION TREE LEARNING

The decision tree includes all the possible consequences with the respective outcomes, costs and utilities. Using this decision tree we see an algorithm that contains conditional control statements. In my program, we see the working of dt as a modified binary tree where the True and False gets sent to either side of the root node to get a more clear picture of the split. We do this learning on all of the features available to get the depth of 1 and understand how the split works.

ADA BOOSTING

A machine learning algorithm which is a modified decision tree implementation as it improves performance. In the code, performance of each one is slightly better than slightly guessing, but

the final model can converge to be the strongest learning. I have used the algorithm from the text book to implement my tree. I ran Adaboost on my decision tree with height of 1 and K values being 100 (which are the number of hypothesis in the ensemble). Since I have provided 100, it successfully reweights every time from the training example. The output seen has the highest implementation from the hypothesis h with weight z.

STEPS

1. Run the train.py and make sure the terminal gives us a print statement of either English or Dutch.
2. Now, open the frequency files to start the comparison of English and dutch.
3. Run the code to get the output with gain and the highest language for each feature.
4. We see that, there is not much difference in output for decision tree and Adaboost, except for the performance.
5. The same step is performed for all the seasons and we see similar paths for all the seasons.
6. The overall steps involved are:
    - Reading the test and train .dat files
    - Parsing through the input and the features
    - Performing decision tree algorithm
    - Performing Adaboost algorithm on the decision tree
    - Getting better output and efieciency.

OUTPUT

For the output, we see a print of the best split performed and which feature has the highest information gain.

ACKNOWLEDGEMENT

The train.dat file was shared from my friend Kunal Nayyar which I've used as my reference too.

NOTE: The output reads nl at times for some kinds of test cases, it is due to the features and has nothing to do with the split functionality or the information gain. Kindly consider this as a minute overlook.