Rochester Institute Of Technology

Foundations of Artificial Intelligence

Report

Lab 2

Rahul Golhar

April 28, 2020

Contents

List of Figures

List of Tables

1 Data Collection

For data collection, I used data from different resources for taking variety of data like literature, academic documents, Wikipedia, etc.

2 Description Of Features

For classification I have defined the following features.

- **hasZ:** This feature tells whether there is a $'z'$ in the sentence. This feature is important because letter $'z'$ occurs more in the Dutch language as compared to the English language.

- **avgWordLen:** This feature tells whether the average word length of a sentence is more than 5.5 letters/word or not. This feature is important because words in the Dutch language are longer as compared to the English language.

- **dutchDiphtongs:** This feature tells whether the sentence has any $DutchDiphtongs$ or not. If it has then it's most probably a Dutch word else it could be an English word.

- **englishStopWords:** This feature tells whether the sentence has any $EnglishStopwords$ or not. These words are provided by NLTK package in python. If it has then it's most probably an English word else it could be a Dutch word.

- **dutchStopWords:** This feature tells whether the sentence has any $DutchStopwords$ or not. These words are provided by NLTK package in python. If it has then it's most probably a Dutch word else it could be an English word.

- **englishCommonWords:** This feature tells whether the sentence has any $CommonEnglishWords$ or not. I collected these words from the internet.If it has then it's most probably an English word else it could be a Dutch word.

- **dutchCommonWords:** This feature tells whether the sentence has any $CommonDutchWords$ or not. I collected these words from the internet. If it has then it's most probably a Dutch word else it could be an English word.

- **repeatedVowels:** This feature tells whether there are repeated vowels one after the other in a word for example $voor(Dutch)$. This feature is important because words in the Dutch language usually have a lot of words which have repeated vowels one after the other.

- **repeatedConsonants:** This feature tells whether there are repeated consonants one after the other in a word for example $opgetrokken(Dutch), schilddak(Dutch)$. This feature is important

because words in the Dutch language usually have a lot of words which have repeated consonants one after the other.

- **ratioVowelsConsonants:** This feature tells if the vowels to consonants ratio in a sentence is more than 0.725. This feature is important because words in the Dutch language usually have a lot more vowels than consonants in a sentence.

- **language:** This is not a feature but the classification of the sentence whether it is *Dutch* or *English*.

3 Decision Tree

3.1 Basic steps followed for coming up with decision tree

1. Read the data from the file
2. Send the data for extracting attributes
3. Get the extracted attributes into a data frame
4. Now use this data to form a decision tree
5. Save the decision tree into a file

3.2 Decision tree process in detail

1. Initially, we have a data frame with info about all the 10 attributes.
2. Now we find the attribute that has the highest information gain(Explained in details later)
3. Once we have found the most important attribute A, we will make 2 sub-trees with one side denoting the case when the attribute A is $True$ and other side for the case when attribute A is $False$.
4. While making these sub-trees we need to note than we remove the attribute A from the data frame.
5. We do this for all the nodes. However, there are 3 base cases for stopping the recursion.
    - When there are not attributes to drop: in such case we return the plurality value of the parent frame.
    - When there is just one result that will be returned by the sub-tree no matter which attribute is true or false: in such case the value is directly returned without further recursion
    - When there is no more column to drop: in such case the plurality value of current data is directly returned

6. Once we have the true and the false sub-trees we add those as branches to the current node we found

3.3 Process of finding the most important feature in detail

1. Initially, we need to calculate some values on the whole data frame received like no of cases when the output is *English* and entropy value(explained later) of result as *English*.
2. Once we have these values, we iterate over all the features we have in the data frame received.
3. For every feature we need to find some values which are required to find the information gain of the feature. These are given below:
   (a) No of time we get *English* as result when the current feature being examined is true.
   (b) No of time we get *Dutch* as result when the current feature being examined is true.
   (c) No of time we get *English* as result when the current feature being examined is false.
   (d) No of time we get *Dutch* as result when the current feature being examined is false.
   (e) Total no of cases when the current feature being examined is true.
   (f) Total no of cases when the current feature being examined is false.
4. Using these, we find the probability for each case given above.
5. Now, we need to find the remainder for the current feature. It can be given as below:

$$\begin{aligned} Remainder = {} & Entropy(Probability\ of\ case\ a) \times (Probability\ of\ case\ e) \\ & + Entropy(Probability\ of\ case\ c) \times (Probability\ of\ case\ f) \end{aligned}$$

$$\begin{aligned} Remainder = {} & Entropy(\frac{true\ Values\ giving\ English}{total\ True\ values}) \times (probability\ of\ true\ values\ ) \\ & + Entropy(\frac{false\ Values\ giving\ English}{total\ false\ values}) \times (probability\ of\ false\ values\ ) \end{aligned}$$

6. However, in a case where there are only true values or only true values, we just assign the entropy of *English*(calculated in the initial step) as the remainder value.
7. Once we have the Remainder, now we find the gain using the following formula

$$Informationgain = Entropy\ of\ English - Remainder\ of\ English$$

8. We do the steps 2 for every attribute and find the maximum gain and the attribute that gives it.
9. Now, just return the attribute that had the maximum gain.

3.4 Other functions used

1. **Entropy:** The entropy for a value given can be calculated using the formula given in the book and is mentioned below.

$$Entropy(q) = -qlog_2q + (1-q)log_2(1-q)$$

2. **Plurality value:** The Plurality value for the data frame is the result that appears most of the time i.e. *English* or *Dutch*.

3. **Classification:** The Classification for the data frame is the number of times the result is *English*.

3.5 Predictions using Decision Tree:

1. For prediction using a Decision tree, we read the file specified and get the model to be used in a Root class variable.

2. Also, convert the data from test file into attributes as we did before.

3. Now, we have root node of the tree and the data frame having the data about the lines in the testing file.

4. Now, we do the following steps for attributes of every line:

   (a) If the value of the root node is a result value i.e. *English* or *Dutch* then we print the result and return else we continue.

   (b) If the value of the attribute at root node is true then we recursively traverse the true sub-tree of the node.

   (c) If the value of the attribute at root node is false then we recursively traverse the false sub-tree of the node.

5. This is how we will print the result for every line.

4 AdaBoost Model

4.1 Basic steps followed for coming up with decision AdaBoost Model

1. Read the data from the file
2. Send the data for extracting attributes
3. Get the extracted attributes into a data frame
4. Now use this data to form a AdaBoost Model
5. Save the decision tree into a file

4.2 AdaBoost Model in detail

1. Initially, we have a data frame with info about all the 10 attributes.
2. We need to find a number of stumps which will be used for predictions later.
3. For every stump we get the hypothesis and its hypothesis weight. This is how we get a number of stumps.
4. First, we find a stump with attribute that has the highest information gain(section 3.3) using learning algorithm(Explained in details later)
5. Once we have found the stump with most important attribute we now get the data with updated weights and this will help us in updating the data for next stump.
6. Now we update the data by keeping more entries that were wrongly classified and less entries that were correctly classified.
7. Once we have a new data we do the previous 2 steps for all stumps.
8. When we have all the stumps and their hypothesis weights, we assign those to a AdaBoost Model class object and save it.

4.3 Learning Algorithm in detail

1. Initially, we have a certain data frame with the initial weights.
2. Now we find the most important feature(section 3.3) of the current data.
3. Once we have the most important feature, we find the incorrect predictions for the current feature i.e. the places where the value was $True$ but it predicted $Dutch$ or $False$ but it predicted $English$.
4. Once we have the incorrect entries, we add the weights of all those entries.

5. Now, we find the amount of say of the current stump which is given by following formula

$$Amount\ of\ say = \frac{1}{2} log(\frac{1 - totalError}{totalError})$$
$$where,\ totalError = sum\ of\ weights\ of\ all\ incorrect\ entries$$

6. Once we have the amount of say we can use it to determine the new weights to be assigned to the values that were incorrectly classified and values that were correctly classified.

7. The factor by which we multiply the previous weights are given using the following formula

$$Factor\ to\ multiply\ for\ correctly\ classified = e^{-amount\ of\ say}$$
$$Factor\ to\ multiply\ for\ incorrectly\ classified = e^{amount\ of\ say}$$

$$The\ New\ weight\ for\ an\ entry = previous\ weight \times Factor\ to\ multiply\ for\ correctly/incorrectly\ classified$$

8. Now we have the new weights and we will now normalize them.

9. Now we return the data related to the stump.

4.4 Tackling incorrect classification

1. Once we have got a stump and the weights based on that stump we need to make a better data ready for the next stump so that it predicts better.

2. We need to take the values that have more weight i.e. that were incorrectly classified previously for this we use cumulative sum of all weights.

3. For doing this we use following steps no of times equal to length of the total data frame.
   (a) First, we select a random number between 0 and 1.
   (b) Now, we cumulatively add the sum of every entry in data frame and when the sum exceed the random number, we add that entry to the new data frame.
   (c) Since the weights of incorrect values are high, there is more chance of incorrect ones being added.

4. Once we have the new data frame, we re-assign the weights to $1/N$ where N is size of data frame.

5. Now we return this new data frame which is then used for next stump.

4.5 Predictions using AdaBoost Model:

1. For prediction using a AdaBoost Model, we read the file specified and get the model to be used in a $AdaBoostModel$ class variable.

2. Also, convert the data from test file into attributes as we did before.

3. Now, we have the *hypothesis vector* and *hypothesis weight vector* of the model.
4. Now, we do the following steps for attributes of every line and for every attribute in *hypothesis vector*:
   (a) Get the value of the feature or the current row and get the *hypothesis weight* of the feature.
   (b) Follow the following table to add the values of the weights to possibility of english or dutch variables.

Table 1: Where to add weight to predict

Attribute Value = True		**Attribute Value = False**	
Hypothesis Weight		Hypothesis Weight = Positive	
Positive	Negative	Positive	Negative
Add weight to English variable	Add weight to Dutch variable	Add weight to English variable	Add weight to Dutch variable

   (c) If the value of the english variable is more than dutch variable then the given line is *English* else it is *Dutch*.
5. This is how we will print the result for every line.

4.6 Testing Results on the Trained Models:

For testing purpose I used the following 30 lines and their actual results along with the predicted results are given in the figure. Every column has the number of lines classified at the top. The Yellow marked boxes are the ones that are incorrectly classified and the ones with Gray color are the one which were classified wrongly before with smaller data set but are now classified correctly with larger data set. The image **1** shows the results of **Decision Tree Model** and the image **2** shows the results of **AdaBoost Model**.

5 How to use the code

The files supplied contain:

1. **train.py:** This file is used to train a model and save the model to a file. To run this file use command as follows:

   train <examples> <hypothesisOut> <learning-type>

   - examples is a file containing labeled examples.
   - hypothesisOut specifies the file name to write your model to
   - learning-type specifies the type of learning algorithm you will run, it is either "dt" or "ada".

2. **predict.py:** This file is used to classify each line in a file as either English or Dutch using the specified trained model file. To run this file use command as follows:

   predict <hypothesis> <file>

   - hypothesis is a trained decision tree or ensemble created by your train program
   - file is a file containing lines of 15 word sentence fragments in either English or Dutch.

3. **attributes.py:** This file is used to assign attributes for every line read from a file.

4. **tree.py:** This file is used to store a class *Root* used for storing a node of a decision tree.

5. **trainDecTree.py:** This file contains functions that are required for generating a decision tree model.

6. **trainAdaBoost.py:** This file contains functions that are required for generating an Adaboost model.

7. **predictDecTree.py:** This file contains functions that are required for predicting a result using a decision tree model.

8. **predictAdaBoost.py:** This file contains functions that are required for predicting a result using an Adaboost model.

9. **trainedDecTreeModel_rg1391:** This file contains a trained Decision Tree Model. When we access this file it returns an object of class *Root* specified in *tree.py* file. It has 3 attributes:

   - attributeVal: This is the attribute at the root node of the decision tree.
   - root.true: Root node of subtree to traverse when the root attribute is true
   - root.false: Root node of subtree to traverse when the root attribute is false

10. **trainedAdaBoostModel_rg1391:** This file contains a trained Adaboost Model. When we access this file it returns an object of class $AdaModel$ specified in $trainAdaBoost.py$ file. It has 2 attributes: $hypothesisVector$ and $hypothesisWeightVector$, which are lists. The names of the attributes suggest what they actually have inside. We can now use these 2 lists to predict data.

	Sentence	Language	100 Line	1000 Line	5000 Line	10000 Line	15000 Line
2	Parliament and the Commission clearly share the same vision and the same goal for SEPA.	en	en	en	en	en	en
3	As usual, in EU parlance, 'simplification' means more standardisation and more control by the Commission.	en	en	en	en	en	en
4	Thus, with regard to political and social issues, the Maastricht promises have not been kept.	en	en	en	en	en	en
5	Unfortunately, a failure to recognise this reality is evident in the current EU 2020 strategy.	en	en	en	en	en	en
6	Our opponents are egotism, nationalism and protectionism, not more Europe and not a strong Commission.	en	en	en	en	en	en
7	In this context, there are one or two ideas I would like to put forward.	en	en	en	en	en	en
8	Let us use these six months to make Europe and the community of Europe stronger.	en	en	en	en	en	en
9	The motion for a resolution before us deals with cross-border cooperation in the Union.	en	nl	en	en	en	en
10	It is unfortunate that no ministerial meeting on this policy has been organised since 2008.	en	en	en	en	en	en
11	Madam President, today's resolution comes at a turning point in work in this area.	en	en	en	en	en	en
12	From the point of view of this Parliament, this type of arrangement cannot be allowed.	en	en	en	en	en	en
13	On the second issue: you are aware of the position of the Commission on labelling.	en	nl	nl	en	en	en
14	The Petersberg Tasks can become the launching pad for a European security and defence identity.	en	en	en	en	en	en
15	I would also ask for the earliest possible tightening of the Stability and Growth Pact.	en	en	en	en	en	en
16	I believe that to be an extremely important point, particularly with regard to the future.	en	en	en	en	en	en
17	Anders hebben we vier stemmingen gehad en het amendement verworpen, waarvoor vijf stemmingen nodig zijn.	nl	nl	nl	nl	nl	nl
18	Laten ze maar oppassen, want op een dag lachen we ze alleen nog maar uit.	nl	nl	nl	nl	nl	nl
19	Met dit programma kunnen we rechtstreeks waarde toevoegen aan de levensk waliteit van de Europese burgers.	nl	en	nl	nl	nl	nl
20	Op het ogenblik hebben we een bereik van 1300 km, wat geen kleinigheid is.	nl	en	en	en	en	en
21	Zoals veel sprekers eerder al hebben gezegd, moeten wij afstand nemen van de open coördinatiemethode.	nl	nl	nl	nl	nl	nl
22	Europa moet veel meer zijn dan dit, omdat de volkeren en de burgers dat willen.	nl	nl	nl	nl	nl	nl
23	Ik verlaat op 15 oktober het Europees Parlement dat ik vanaf februari 1981 heb meegemaakt.	nl	nl	nl	nl	nl	nl
24	Het woord is aan de heer Robert J.E. Evans voor een motie van orde.	nl	nl	nl	nl	nl	nl
25	Ik denk in dit verband met name aan Georgië dat 46 miljoen euro zou krijgen.	nl	nl	nl	nl	nl	nl
26	De relatie tussen de EU en Rusland mag geen inzet zijn van deze historische discussies.	nl	nl	nl	nl	nl	nl
27	Het zal boeiend zijn om te zien hoe dit netwerk zich verder zal gaan vertakken.	nl	nl	nl	nl	nl	nl
28	Dat tijdsbestek is wel heel erg klein en ik vrees dat daardoor problemen zullen ontstaan.	nl	nl	nl	nl	nl	nl
29	Hoe nauwkeuriger en betrouwbaarder de verschafte informatie is, des te beter ons milieubeleid zal worden.	nl	nl	nl	nl	nl	nl
30	Helaas zijn de meeste aanbiedingen op het gebied van de financiële dienstverlening tot dusver teleurstellend.	nl	nl	nl	nl	nl	nl
31	De strijd mag echter niet leiden tot onveiligheid, sociale dumping en milieuverwilling in de havens.	nl	nl	en	en	en	en
32		Accuracy	0.87	0.9	0.93	0.93	0.93

Figure 1: Decision tree Results

	Sentence	Language	100 Line	1000 Line	5000 Line	10000 Line
2	Parliament and the Commission clearly share the same vision and the same goal for SEPA.	en	en	en	en	en
3	As usual, in EU parlance, 'simplification' means more standardisation and more control by the Commission.	en	en	en	en	en
4	Thus, with regard to political and social issues, the Maastricht promises have not been kept.	en	en	en	en	en
5	Unfortunately, a failure to recognise this reality is evident in the current EU 2020 strategy.	en	en	en	en	en
6	Our opponents are egotism, nationalism and protectionism, not more Europe and not a strong Commission.	en	en	en	en	en
7	In this context, there are one or two ideas I would like to put forward.	en	en	en	en	en
8	Let us use these six months to make Europe and the community of Europe stronger.	en	en	en	en	en
9	The motion for a resolution before us deals with cross-border cooperation in the Union.	en	en	en	en	en
10	It is unfortunate that no ministerial meeting on this policy has been organised since 2008.	en	en	en	en	en
11	Madam President, today' s resolution comes at a turning point in work in this area.	en	en	en	en	en
12	From the point of view of this Parliament, this type of arrangement cannot be allowed.	en	en	en	en	en
13	On the second issue: you are aware of the position of the Commission on labelling.	en	en	en	en	en
14	The Petersberg Tasks can become the launching pad for a European security and defence identity.	en	en	en	en	en
15	I would also ask for the earliest possible tightening of the Stability and Growth Pact.	en	en	en	en	en
16	I believe that to be an extremely important point, particularly with regard to the future.	en	en	en	en	en
17	Anders hebben we vier stemmingen gehad en het amendement verworpen, waarvoor vijf stemmingen nodig zijn.	nl	nl	en	en	en
18	Laten ze maar oppassen, want op een dag lachen we ze alleen nog maar uit.	nl	nl	nl	nl	nl
19	Met dit programma kunnen we rechtstreeks waarde toevoegen aan de levenskwaliteit van de Europese burgers.	nl	en	nl	nl	nl
20	Op het ogenblik hebben we een bereik van 1300 km, wat geen kleinigheid is.	nl	en	en	en	en
21	Zoals veel sprekers eerder al hebben gezegd, moeten wij afstand nemen van de open coördinatiemethode.	nl	nl	nl	nl	nl
22	Europa moet veel meer zijn dan dit, omdat de volkeren en de burgers dat willen.	nl	nl	nl	nl	nl
23	Ik verlaat op 15 oktober het Europees Parlement dat ik vanaf februari 1981 heb meegemaakt.	nl	nl	nl	nl	nl
24	Het woord is aan de heer Robert J.E. Evans voor een motie van orde.	nl	nl	nl	nl	nl
25	Ik denk in dit verband met name aan Georgië dat 46 miljoen euro zou krijgen.	nl	nl	en	en	en
26	De relatie tussen de EU en Rusland mag geen inzet zijn van deze historische discussies.	nl	nl	nl	nl	nl
27	Het zal boeiend zijn om te zien hoe dit netwerk zich verder zal gaan vertakken.	nl	nl	nl	nl	nl
28	Dat tijdsbestek is wel heel erg klein en ik vrees dat daardoor problemen zullen ontstaan.	nl	nl	nl	nl	nl
29	Hoe nauwkeuriger en betrouwbaarder de verschafte informatie is, des te beter ons milieubeleid zal worden.	nl	nl	nl	nl	nl
30	Helaas zijn de meeste aanbiedingen op het gebied van de financiële dienstverlening tot dusver teleurstellend.	nl	nl	nl	nl	nl
31	De strijd mag echter niet leiden tot onveiligheid, sociale dumping en milieuverwilling in de havens.	nl	en	en	en	en
32		Accuracy	NONE	0.9	0.93	0.93

Figure 2: AdaBoost Results