Foundations of Intelligent Systems, Lab2

Hitesh Sapkota

December, 02, 2018

1 Data Collection:

For data collection, I used the Wikipedia package available in Python3.6. With this package, first I generated random titles and then for each title I crawled the summary. The total dataset consists of 4068 balanced examples.

2 Feature Selection:

My feature selection involves manual inspection of several Wikipedia random articles written in English and Dutch words. Moreover, I did some research to investigate how the Dutch language differs from the English language. Based on my direct inspection along with the research, I came up with the following features.

2.1 hasY

From research, I found that probability of occurrence of the character 'y' in English language is 0.06% whereas it is 3.97% in the Dutch Language [Refer to the link1 and link2 to verify the provided facts]. Therefore, I decided to use this feature, and is true if at least one word contains y character.

2.2 startD

From the manual inspection, I found that lots of words in Dutch starts with the letter 'D'. However, in the English language this is not a case. Therefore, I added this feature and is true if at least one word starts with the alphabet 'D'.

2.3 repeatVowels

From the manual inspection, I noticed that many Dutch words have a repeated vowel such as 'ee' which is unlikely to happen in the English words. Therefore, I decided to use this feature and is true if at least one word has the same repeated vowel.

2.4 presenceij

In the Dutch language, ij is most frequently but its use is rare in the English Language. Therefore, I decided to use this as a feature and is true if at least one word has consecutive ij characters.

2.5 frequentdutchWords

There are frequent Dutch words such as 'ik', 'je' etc. (refer to following link for full list) that are barely used in the English language. Therefore, I incorporated this feature and is true if there is a word from the top 10 most frequent Dutch words.

2.6 frequentenglishWords

There are frequent English words such as 'the', 'be' etc. (refer to following link for full list) that are barely used in the Dutch language. Therefore, I incorporated this feature and is true if there is a word from the top 10 most frequent English words.

2.7 avgWordLength

From my research, I found that the average English word length is 8.23 whereas that of Dutch word is 9.7. Therefore, I used this feature with the threshold of 8.5.

2.8 startswithAVowel

English sentences seem to start with A vowel than that of Dutch words. Therefore, I decided to incorporate this feature and is true if a sentence begins with a letter A.

2.9 presencedutchDiphtongs

From the research, I found a list of Diphthongs usually found in the Dutch Language such as 'ae', 'ei'. Therefore, I incorporated this feature.

2.10 presenceEnglishWords

From observation, I realized that many Dutch words contain characters other than alphabets (A-Z, a-z), and numbers (0-9) that we use in the English Language. Therefore I used this feature for classification.

2.11 hasEnglishStopWords

Stop words usually vary from one language to another language. Therefore I use English Stopword as a feature and is true if there is at least one stop word in the provided text.

2.12 hasDutchtopWords

It is true when there is at least one Dutch Stopword in the provided text.

3 Decision Tree

In decision tree information gain is used to determine the best split. We continue on exploring the tree until the predefined depth is met or leaf node is reached from which further split is not possible.

3.1 Parameter Selection

We divide the given data into the two categories: train data, test data. Then, for each depth length (starting from 1 to the length equals to a number of predictor variables) we find out the average accuracy. Here average accuracy is a ratio of a total number of correctly predicted instances to total data instances. We select the depth that gives maximum prediction accuracy as our depth of the tree for classification. So, we can consider this as a cross-validation process. Considering all data, the depth of the tree is 9 with accuracy 94.1.

4 Adaboost

We use the decision stumps as weak learners in the Adaboost technique. The split in the decision stump is determined by the Information Gain computed using weighted examples.

4.1 Parameter Selection

We divide given data to the two categories: train data and test data. Then we construct the AdaBoost classifier starting from 1 decision stump to the number equals the number of predictor variables. In each stage, we compute the accuracy using test data. Finally, we select the Adaboost classifier with maximum accuracy as our final classifier. Considering all data, the depth of the tree is 7 with accuracy 94.1.

5 Sample Size Effect

We experimented with three example types: 10-word samples, 20-word samples, and 50-word samples. Table 1 shows the result for Decision Tree Classifier. From

the table, we can say that the model trained in the examples with n-word sample example has better prediction accuracy in m-word sample example than n-word sample example provided that m¿n. For instance, we can observe from the table that model trained on 10 word sample produces better accuracy for the 50 word sample test data than that of 10 word sample test data.

One of the justifiable reason for this would be model trained on the lower word sample example is more generic and hence, able to capture the properties of even longer word sample examples.

In contrast, the model trained on higher word sample example is tend to perform worse in the lower word sample examples. This might be due to the fact that the model is too specific and failed to capture the true generic properties. For instance, in a table, the model trained in a 50-word sample example has a bad performance on a 10-word sample example. For the prediction purpose, we choose one of the highlighted models according to the word length of the test example.

Table 1: Tree depth and accuracy variation for Decision Tree Classifier

Train Set	Test Set	Depth	Accuracy (%)
10 word sample	**10 word sample**	8	**91.80**
10 word sample	*20 word sample*	5	92.03
10 word sample	*50 word sample*	3	95.14
20 word sample	*10 word sample*	7	91.39
20 word sample	**20 word sample**	9	**94.51**
20 word sample	**50 word sample**	11	**98.58**
50 word sample	*10 word sample*	3	78.68
50 word sample	*20 word sample*	8	92.85
50 word sample	*50 word sample*	5	96.60

For the Adaboost classifier, result is presented in the table 2. We get similar, type of observation as that of decision tree. For the prediction, we choose the highlighted models according to the word length of the test example.

Table 2: Decision Stumps and accuracy variation for Decision Tree Classifier

Train Set	Test Set	Decision Stumps	Accuracy (%)
10 word sample	**10 word sample**	8	**92.21**
10 word sample	*20 word sample*	8	92.85
10 word sample	*50 word sample*	8	95.63
20 word sample	*10 word sample*	9	90.98
20 word sample	**20 word sample**	11	**95.05**
20 word sample	*50 word sample*	8	*96.60*
50 word sample	*10 word sample*	5	86.88
50 word sample	*20 word sample*	5	92.30
50 word sample	**50 word sample**	11	**96.60**