Report of Project: Predict Stress in English Words

Author: 陈卓艺(Joey Chen)

Student id: 15331047

In this project, we are required to predict the primary stress in the English words with a model we trained with a lot of data. And I think the steps to build the classifier of this project is as follows:

1. find the common rules of the primary stress in English words
2. select an efficient classifier with the best performance

I will show what I got in each steps I mentioned above.

Common rules of stress in English

At first, I found some papers about the stress location in English words on the Internet. For example, when there are only two vowels in a word then usually the former one may be the primary stress. It is right and this accuracy seems to reach almost 90%. However, with more than 2 vowels the rules didn't work well. The primary stress location seems to be depend on the property of the English word which means we need to distinguish whether the word is a noun, an adjective or anything else.

I didn't know how to make it so I tried another way.

I used some special prefixes and suffixes to classify the words. Like **-ly**, **-tion**, **-sion** and so on. But the accuracy was quite low at about 70%. So I gave it up.

Then I thought there may be some special order of the vowels so that we can know the primary stress location. It is difficult to make it out by myself. So I built a data frame with the vowels list as columns and classified the train data into the data frame. I train a decision tree classifier with the data frame and the cross validation can reach around 88% accuracy. So I think there must be some relation between the primary stress location and the order of vowels of an English word.

After all the tries, I think it is better to use the order of vowels in an English word as the rule to decide the primary stress location. The values of each columns store the sequence number that the corresponding vowel appear ascending. For example, a word **CAMBODIAN:K AE2 M B OW1 D IY0 AH0 N** will be something like

AA	AE	AH	AO	AW	AY	EH	ER	EY	IH	IY	OW	OY	UH	UW
0	1	4	0	0	0	0	0	0	0	3	2	0	0	0

The best classifier for this project

This is a supervised learning and we don't need to build the label from data. Since there is a linear relation between the data I chose and the results, it is better to use a decision tree classifier to be the model. I also considered to try with the NB classifiers and obviously it didn't work well. So at last I used decision tree classifier.

Conclusion

In this project, what impresses me most is that machine learning is amazing and interesting since it can predict the results with high accuracy from the data with some models. I can't image this ever before and it really inspired me. However, my elementary knowledge about the statistical and math is not enough and there are still many things I don't really understand about the machine learning. I am glad to be taught about this and thank you very much.