Naive Bayes Classfier

YANG LIU

September 25, 2014

This Naïve Bayes Classifier accomplishes the job of classifying the gender of the writers of blogs with the accuracy about 70%.
I have some variables to record the different types of parameters generated in the process of training. There are some arrays, such as parameters, labels, possibilities-of-features, etc. The different index of these array means different classes, such as male and female. In the parameters and possibilities-of-features there is a dictionary that is used for recording the appearance of different features (or words in this specific assignment). In this way, the program is able to record the parameters of different features in different classes.
The variable ?type? is for select which kind of classifier, multinomial or Bernoulli. The program has too kinds of methods to deal with the different parameters, which is controlled by the variable 'type'.
Even it is not complex for computers to do multiple computations, but I used log to compute the final probability because the probability of each feature is so small that the result could be 0.

1 Smoothing

I deployed to kinds of smoothing method, Laplace smoothing and Generalized Laplace smoothing, but there not much difference between when we looking at the result of classifying.

2 Features

There are 4 ways of choosing features in this program: BagOfWords, BagOfWordsImproved, GenderPrefer, NGrams. First, BagOfWrods is a given method. Second, the BagOfWordsImproved get rid of various kinds of punctuations and empty elements. GenderPrefer is a method I referred from the paper which select certain kind of words that could be used to distinguish between male and female. And the last is the N-gram method which combine n words together to be a feature. Even it has a worse result in the blog classify but it really helpful in the name classify. I used the first two letters and the last letter of a name to be the feature and it gain an accuracy of 79%. Unfortunately, the test shows that the best result comes up with the simplest method: BagOfWords. Another method called BagOfWordsImprovedForBernoulli is used for Bernoulli classifier.

3 Classifier

I implemented both two kinds of Naïve Bayes classifier. The Bernoulli model could get the accuracy about. With the Multinomial model, the program could gain a highest accuracy about 72%. The number of items that used for train and test also has an effect of the accuracy, the test results listed below.

2750	2800	2850	2900	2950	3000	3050	3100	3150	3200
70%	70%	70%	70%	69%	71%	69%	69%	72%	70%