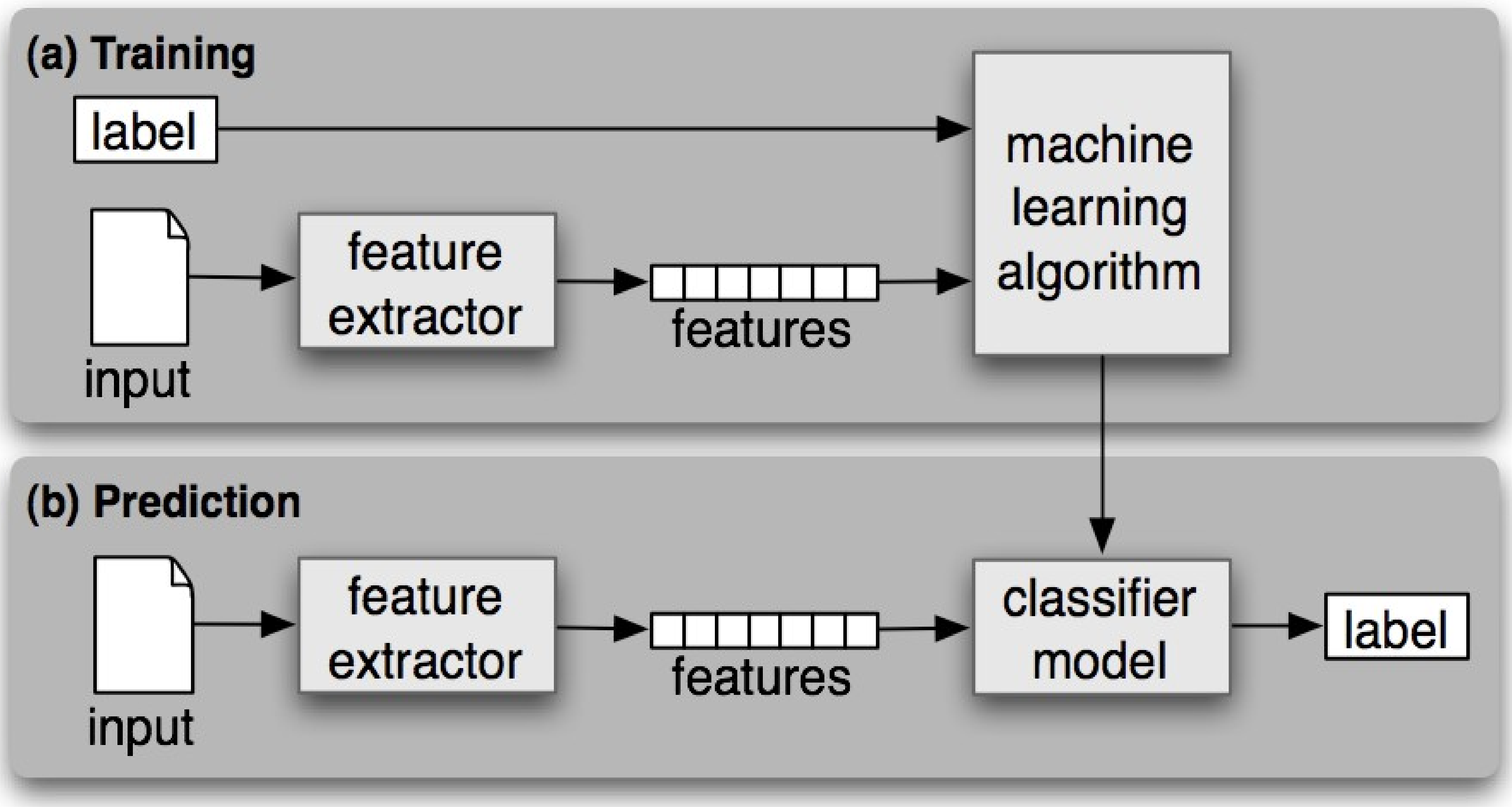
Learning to Classify Documents
Edwin Zhang
Computer Systems Lab 2009-2010

Background and Introduction

I will be learning to classify documents using a Bayesian method to classify documents into certain categories. My program will have two parts: a Learning part and a Prediction part. In the Learning section, I will begin by using a set of training documents to come up with an formula for classifying documents and then begin testing it on documents where I do not know the subject. I will choose a set of features (words) based on my training documents. In the Prediction Part of the program, I will be using the features that I selected previously to predict the category of a document.
My method is very similar to the Naïve Bayes Classifier, albeit with some slight modifications.



Discussion

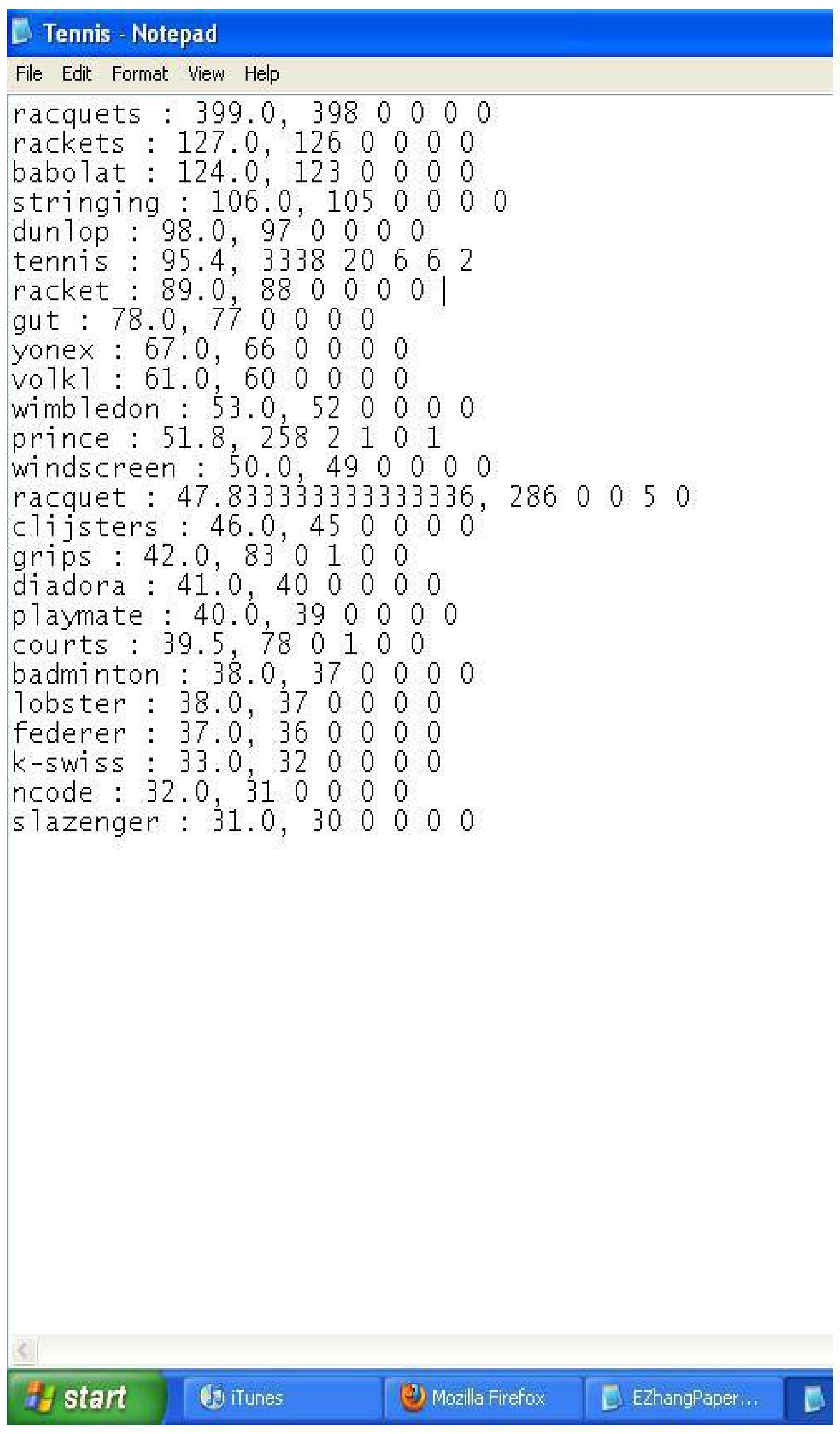
Although my program worked very well on the limited number of the documents, I wish that I had been able to do more tests and test on more documents. In addition, I had hoped that I would have time to add more categories, but that was not the case. However, I am very satisfied with how my program performed and the high level of success my program had. There were, however, several areas which could be further explored. Different methods for calculating the score of the terms would likely produce different results and the different ways of calculating the scores of the categories in the Prediction part would also influence the success of the program. In addition, I expect that doing the counts differently would produce different results. These are only a few of the possible areas that could be explored.
There are several real-life applications of my program and this idea of classifying documents. One example if when we receive email, our emails uses a type of document classification to classify emails as spam or not spam. There are many more examples, but that's just one very common case that affects us everyday, even when we don't know about it.

Development

My program has two parts: a Learning part and a Prediction part. For my Learning Part, I read in training documents and selecting features to use. In addition, I created three separate classes: Category, Document and Terms classes. The Category class deals with the categories and stores an array of documents specific to that category. The Document class deals with the document and the terms in the document. The Terms class deals with all the terms and the number of times each term appears in training documents from each category. I assigned each term a score based on the counts in the respective categories. Then, I sorted my array of Terms by score and for each category, I choose the top 25 terms for each category those terms are my features.
For the Prediction part of the program, I read in a document where I do not know the category and I will read through the documents to look for terms that I selected as features. Each category has a variable and whenever I find a term I multiply each category variable by a number I determine using the counts.

Results and Conclusions

For two categories, my program initially malfunctioned and did not correctly classify the documents. After some corrections, it correctly classified 10 documents. When I moved the five categories, it initially malfunctioned again, but after some tinkering, it worked roughly 90-95% of the time on roughly 30 test documents.



```
racquets : 399.0, 398 0 0 0 0
rackets : 127.0, 126 0 0 0 0
babolat : 124.0, 123 0 0 0 0
stringing : 106.0, 105 0 0 0 0
dunlop : 98.0, 97 0 0 0 0
tennis : 95.4, 3338 20 6 6 2
racket : 89.0, 88 0 0 0 0
gut : 78.0, 77 0 0 0 0
yonex : 67.0, 66 0 0 0 0
volkl : 61.0, 60 0 0 0 0
wimbledon : 53.0, 52 0 0 0 0
prince : 51.8, 258 2 1 0 1
windscreen : 50.0, 49 0 0 0 0
racquet : 47.833333333333336, 286 0 0 5 0
clijsters : 46.0, 45 0 0 0 0
grips : 42.0, 83 0 1 0 0
diadora : 41.0, 40 0 0 0 0
playmate : 40.0, 39 0 0 0 0
courts : 39.5, 78 0 1 0 0
badminton : 38.0, 37 0 0 0 0
lobster : 38.0, 37 0 0 0 0
federer : 37.0, 36 0 0 0 0
k-swiss : 33.0, 32 0 0 0 0
ncode : 32.0, 31 0 0 0 0
slazenger : 31.0, 30 0 0 0 0
```

An Example of a list of features and the corresponding score and counts. This one is the list of features for tennis.