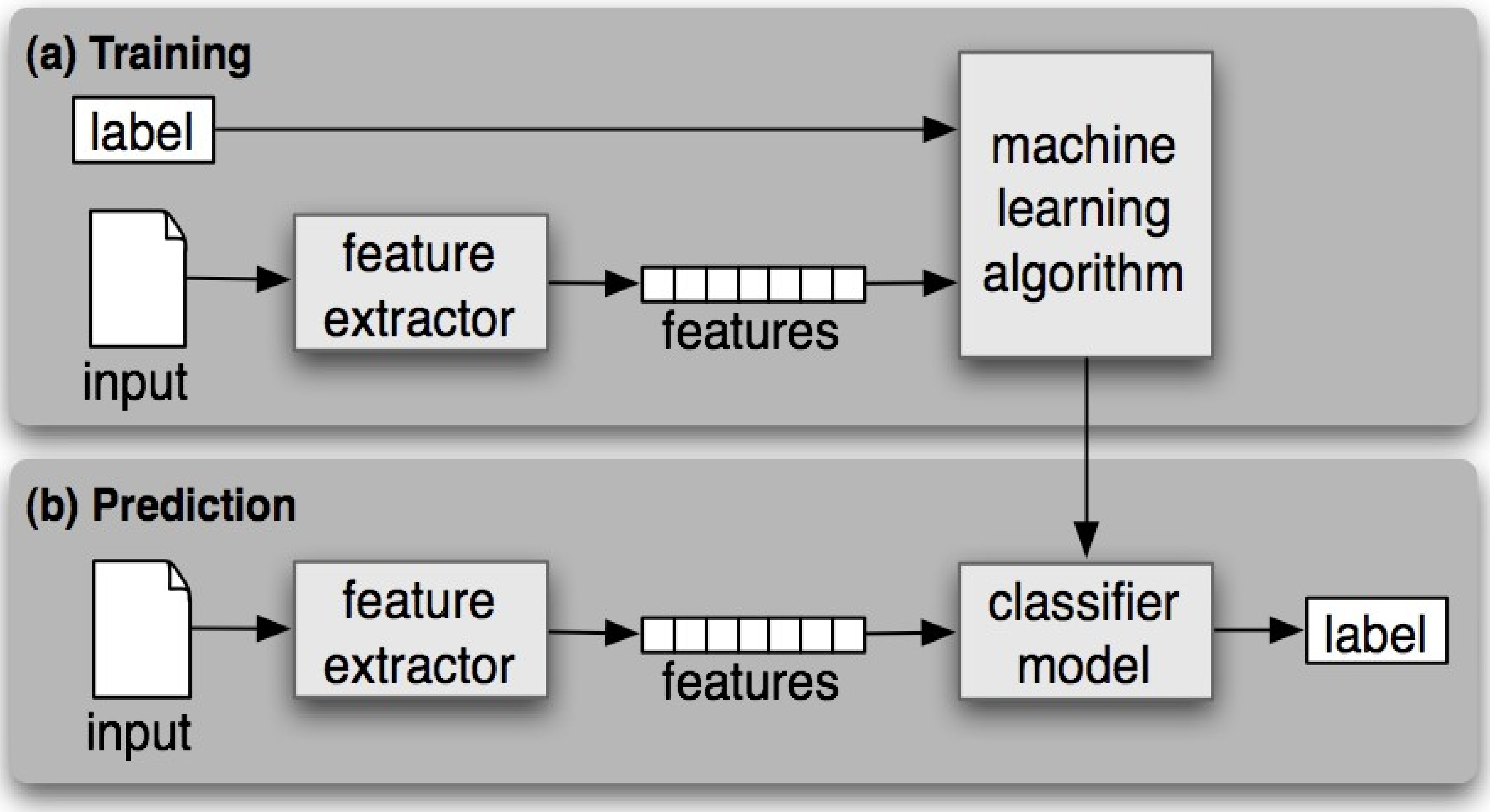
Learning to Classify Documents
Edwin Zhang
Computer Systems Lab 2009-2010

Abstract

I will be learning to classify documents using a Bayesian method to classify documents into certain categories. 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) that are specific to a certain topic and use conditional probability to determine how often the words appear in the training documents and use that to classify other documents

Background and Introduction

In this project, I will be using the Naïve Bayes Classifier. The Naïve Bayes Classifiercomputes the conditional probability p(T|D) for a given document D for every topic T and assigns the document D to the topic with the largest conditional probability. Naïve Bayes Classifier then converts the calculation of the conditional probability into a formula that is easy to calculate using the Bayes rule.



Development

My program has two parts: a Learning part and a Prediction part. For my Learning Part, I have gotten training documents for my program to learn from. In addition, I have 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 also assign each term a score based on the counts in the respective categories. Then, I sort my array of Terms by score and for each category, I choose the top 25 terms for each category and those will be my features. For the Prediction part of the program, I will read in a document where I do not know the category and based on my features, I will see how often my features appear in that specific document and calculate a probability based on this. Then, I will find which category in which the terms had the highest probability and that will be the likely category.

Discussion

Right now, I am starting with only two categories: tennis and other. Once my program predicts correctly for two categories, then I will add more categories and keep testing. In addition, I may modify my score calculating method for the terms. Right now, I am just adding the number of times that terms in appear in a certain category plus one over the number of times the term appears in all the other categories combined plus one to avoid dividing by 0. While this may not necessarily need to be changed, I may play around with that to see which formula produces better results.

Results and Conclusions

I expect that initially, the program may have trouble classifying documents into the correct category but as the program learns more and improves its formulas, it will get better at classifying documents into the correct categories. Since I have not finished my Prediction part, I do not yet have results, but I will keep working on it.