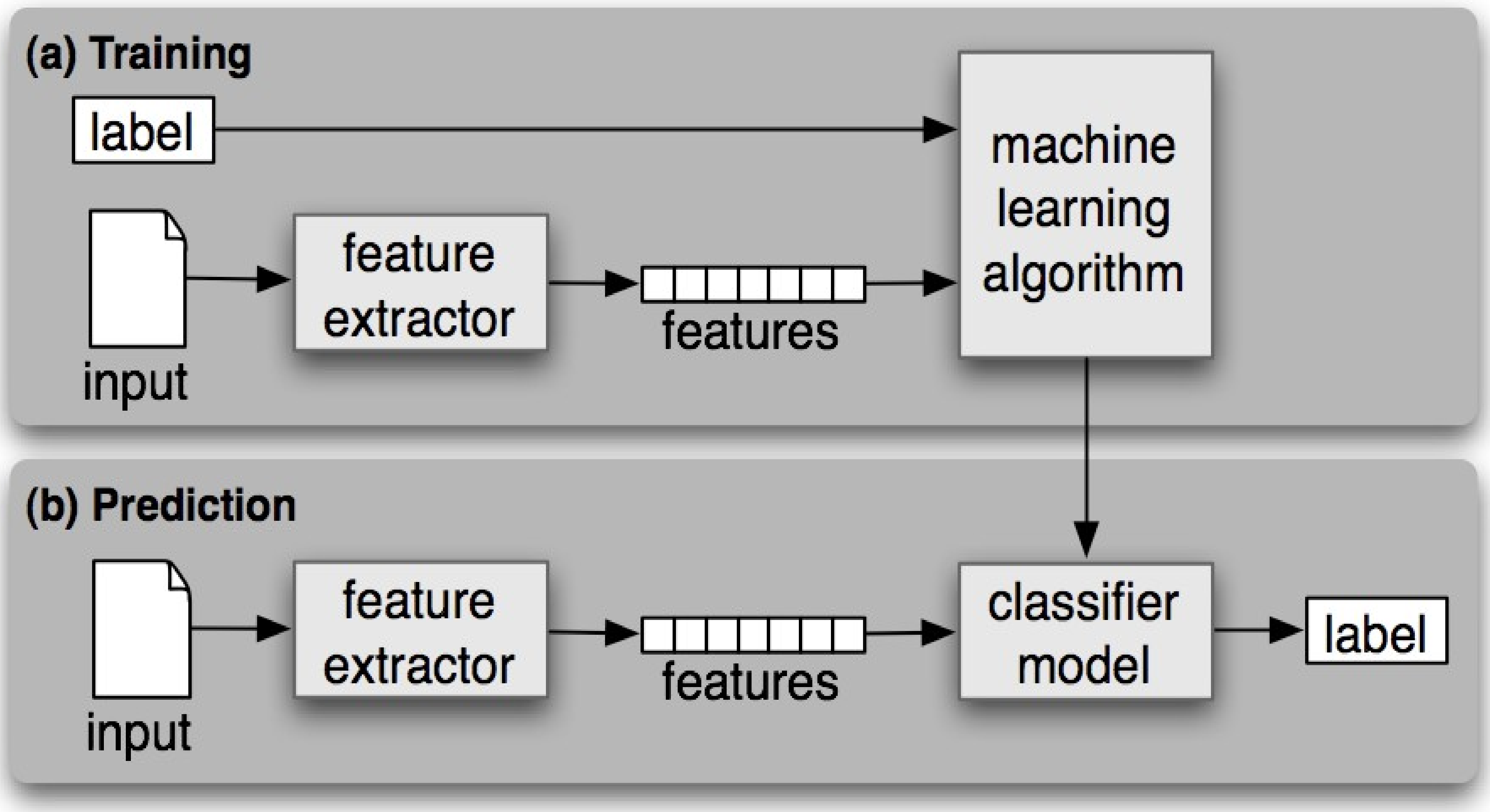
Learning to Classify Documents
Edwin Zhang
Computer Systems Lab 2009-2010

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

After I choose the topics I am going to be classifying documents in, I am going to choose certain features that apply only to that topic or apply mainly to that topic. I will also choose certain words that are in all documents, such as “the” and “and.” Then, my program will begin learning and I will take a set of training documents and figure out how often my features appear in each document, regardless of the topic. I will store my answers and use them for later documents. Then I will use documents and see how often certain features appear in that document to determine what the topic probably is.

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 have created two classes: the Document class, which deals with my documents, and the Category class, which deals with all my categories. What I need to do right now is to get documents for each of my categories so that my program can start learning and I can start testing. In addition, I need to develop the learning part of the program and the formula.

Results and Conclusions

I expect that inittially, the program may have trouble classifying documents into the correct cateogry but as the program learns more and improves its formulas, it will get better at classifying documents into the correct categories. Since I do not have documents yet, or yet coded my formula, I do not have any results thus far.