

Learning to Classify Documents

TJHSST Senior Research Project

Computer Systems Lab 2009-2010

Edwin Zhang

October 29, 2009

Abstract

This project uses a Bayesian method to classify documents into certain categories. A set of training data will be used to derive a formula for probability. A set of features (words) specific to a certain topic and the conditional probability of the appearance of these features (the formula), will be used to determine the classification of unknown documents.

Keywords: bayesian probability, document classification

1 Introduction

In this project, I will be using the Naive Bayes Classifier to classify documents based on content. The Naive Bayes Classifier computes 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. The Naive Bayes Classifier then converts the calculation of the conditional probability into a formula that can be easily calculated using Bayes rule.

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.

2 Development

The program developed consists of two major steps: Learning and Prediction. The Learning part makes use of training documents to develop a formula for conditional probability, to be based on the probability that certain features appear in documents of similar topic. We will go through the training documents and look at how often a certain feature appears in a document that is about a certain topic. For example, if our topic is "tennis" and our feature is "unforced error" we would go through all the documents and see how often "unforced error" occurs in documents about tennis and other documents. The Prediction part uses the results from the Learning portion to predict and classify the topic of an unknown document.

3 Expected Results

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.

4 Discussion

References

- [1] Eyheramendy, Susana, and David Madigan. "A Flexible Bayesian Generalized Linear Model for Dichotomous Response Data with an Application to Text Categorization", *Lecture Notes-Monograph Series*, 54 (2007): 76-91. JSTOR. Web. 25 Oct. 2009. <http://www.jstor.org/stable/20461460>.