Learning to Classify Documents
TJHSST Senior Research Project
Computer Systems Lab 2009-2010

Edwin Zhang

January 21, 2010

Abstract

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

Keywords: bayesian probability, document classification

1 Introduction

In this project, I will be using the Nave Bayes Classifier to classify documents based on content. The Nave 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 Nave 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 cateogry 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 if "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. 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. I need 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.

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

Since I do not have my documents yet, which I expect that I will get soon, I have not been able to test yet. I also still need to come up with a formula and once I do that, I can start testing to see if my program correctly classifies the documents.

References

[1] Chai, Kian Ming Adam, Hai Leong Chieu, and Hwee Tou Ng. *ACM Portal.* Assocation of Computing Machinery, 2002. Web. 14 Jan. 2010. ¡http://portal.acm.org/citation.cfm?id=564376.564395coll=Portaldl=ACMCFID=70884224C

[2] 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¿.

[3] Lavine, Michael, and Mike West. "A Bayesian Method for Classification and Discrimination." *Canadian Journal of Statistics* 20.4 (1992): 451-461. JSTOR. Web. 14 Jan. 2010. ¡http://www.jstor.org/¿.