title: "Proposal: Classification of BBC News Dataset" author: "Nura Kawa" output: html_document: theme: spacelab

highlight: tango

Introduction

Document classification is the problem of algorithmically assigning text documents to descriptive categories; it is a vital first step in solving larger Natural Language Processing problems, such as text summarization, spam filtering, and sentiment analysis. Document Classification methods fall into two categories: Machine Learning algorithms and Statistical Natural Language Processing Methods. **I propose a project that aims to build language models using n-gram, a statistical language model, to classify BBC news articles into categories, comparing its success to various popular out-of-the-box Machine Learning Algorithms.**

n-gram Models vs. Classification Algorithm

A language model is a probability distribution over a sequence of words. Language models make predictions on class based on calculated probabilities.

The **n-gram model** predicts a letter/word based on a derived probability distribution given a sequence of letters/words. Specifically, the probability of seeing word x in a document can be approximated by the probability of observing it given the previous n-1 words, using a Markov model.

For classification I will use the n-gram model to generate a table of frequencies in the following manner:

Word	P("business")	P("tech")	...
apple	0.2	0.8	
market	0.8	0.1	
money	0.7	0.5	

Classification Algorithms use probabilities to predict a document's class based on feature vectors. I will use both linear classifiers, which work by fitting function to weighted feature vectors, and non-linear classification algorithms that use nearest-neighbor methods for prediction.

Project Outline

I will use the publically-available BBC News Articles Database, downloaded from http://mlg.ucd.ie/datasets/bbc.html. This dataset contains the text of BBC news articles of five categories: business, entertainment, politics, sports, and tech. I will use R to download the text and to clean the data and to split it into training and testing sets.

I will then make visualizations such as word clouds, as well as show summary statistics of properties of the text.

I will use n-gram statistical modeling and machine learning algorithms (support vector machines and LASSO) to classify documents and write conclusions based on results.

My work will be available on my Github repository.