Sentiment Analysis and Movie Reviews

By: Donovan Ambler

Overview

- Problem Description
- Sentiment Analysis and Naive Bayes
- Experimental Design and Procedures
- Results

Problem Description

- For this problem, I will be taking a dataset consisting of 2000 movie reviews from IMBD, and seeing if the machine is capable of using sentiment analysis to properly predict whether the reviews are positive or negative by looking at the words used in each review. I will be using Naive Bayes classification to classify the reviews according to their overall sentiment (positive/negative).

What is Sentiment Analysis?

- The use of statistics, natural language processing and machine learning to extract or categorize the sentiment content of a piece of sample text.
- Generally used to gauge emotional reactions
- Shortcomings of machines in sentiment analysis

The Naive Bayes Classifier

- Based on the Bayes' theorem in statistics
- The theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event
- All naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable

The Basic Plan

- Import and randomize Pang and Lee’s data set
- Create a corpus of the documents
- Clean and simplify the corpus
- Create a document term matrix
- Create 75:25 partitions of data frame, corpus, and DTM
- Restrict the DTM to only the most frequent words
- Train the machine with 1500 of the reviews, test with the other 500.

Results of the First Attempt

	Actual	Actual
Predictions	Negative	Positive
Negative	81	68
Positive	179	172

Accuracy: 50.6%

Results of the Second Attempt

	Actual	Actual
Predictions	Negative	Positive
Negative	69	58
Positive	188	185

Accuracy: 50.8%

Results of the Third Attempt

	Actual	Actual
Predictions	Negative	Positive
Negative	88	93
Positive	153	166

Accuracy: 50.8%

Conclusions

- Not much more accurate than a coin toss
- Predicted positive much more than negative
- Potential future work:
  - Refine the code
  - Categorize on a 1-10 scale

Sources

- An Intuitive (and Short) Explanation of Bayes' Theorem. (n.d.). Retrieved November 15, 2017, from https://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/
- Brownlee, J. (2017, August 15). A Gentle Introduction to the Bag-of-Words Model. Retrieved November 15, 2017, from https://machinelearningmastery.com/gentle-introduction-bag-words-model/
- Katti, R. (2016, April 30). Naive Bayes Classification for Sentiment Analysis of Movie Reviews. Retrieved November 15, 2017, from https://rpubs.com/cen0te/naivebayes-sentimentpolarity
- Movie Review Data. (n.d.). Retrieved November 16, 2017, from http://www.cs.cornell.edu/people/pabo/movie-review-data/
- Naive Bayes Classifier. (n.d.). Retrieved November 15, 2017, from http://www.statsoft.com/textbook/naive-bayes-classifier