Naive Bayes2

Prof. Eric A. Suess

February 24, 2021

Introduction

Today we will work on implementing the Naive Bayes analysis of the SMS data presented in the book.

We will also discuss writing the reports for the class.

Example - filtering SMS

Spam filtering for SMS might be harder than for Email. The messages are shorter.

Working with text data requires a new set of tools for data analysis. In R there are a variety of packages.

- tm
- tidytext
- Tidy Text Mining with R
- Introduction to tidytext
- sentimentr
- rtweet
- text2vec

Text Mining in R

Journal of Statistical Software

- Text Mining Infrastructure in R

The R Journal

- RTextTools: A Supervised Learning Package for Text Classification
- RcmdrPlugin.temis, a Graphical Integrated Text Mining Solution in R

bag-of-words

Today we will go through the code from the book to use naive Bayes to **classify** SMS messages. We will need to read in text data and count words. We will need to apply the naive Bayes algorithm to classify the messages.

The idea with *bag-of-words* is that the words in the messages are considered separately and frequency is used. The *order* of the words is *not taken into consideration*.

For the data preparation we will use the **tm** package to process the messages.

There is a problem with **tolower** and **Dictionary**. We will use the updated commands.

Wordclouds

To compare the training and test datasets we will include wordclouds to see if there is any difference in the commonly used words in ham and spam.

Using the **wordcloud** package and the **wordcloud** function.

Naive Bayes

To implement the naive Bayes algorithm we need to load the **e1071** package and use the **naiveBayes()** and **predict()** functions.

Does the Laplace estimator help?

The last part of the code tries to improve the model performance. To try and improve the model the **Laplace estimator** is used. In the book

laplace = 1

is used.

Can you use 1.5?

Does 2 help more?

Code Writing

Google's R Style Guide

- google R code
- R style guide
- The Tidyverse style guide

Reports

- CS 6375
- CS 391L Machine Learning Project Report Format
- CS 229 Machine Learning Final Reports

Sentiment Analysis of Twitter Data using R

Here are a few interesting blog post about connecting to Twitter and performing Sentiment Analysis.

- Mining Twitter Data with R
- Sentiment Analysis on Twitter Data : Text Analytics Tutorial