MVP for Project 3 - Spam Detection

Belle Peng 07/2018

Domain Description and Motivation

I want to build a spam-detection model to separate out "Spam" vs. "Ham". I chose this topic because there are lots of applications for detecting anomalies, not only in the spam area but also in detecting fraud or outliers for instance.

Data

I found a dataset on Kaggle of 4601 emails with 39.4% spam, 58 features. Sampling methodology won't be an issue for me in this case since I have close to 40% spam. Each observation contains the frequency of 48 words (one word frequency per column). Some examples of these words are "money", "edu", "direct", "free", etc. I also have 6 columns of frequency counts on the following characters: [;, (, [, !, $, #], a column on "capital run length average", "capital run length longest", "capital run engths total", and an indicator whether the email was spam (1) or not (0). That makes up a total of 58 features. I plan to use all of the features on word and character frequency, and pick only one of the capital run length features to predict the Spam (0 / 1).

48 continuous real [0,100] attributes of type word_freq_WORD: percentage of words in the e-mail that match WORD.

6 continuous real [0,100] attributes of type char_freq_CHAR: percentage of characters in the e-mail that match CHAR

1 continuous real [1,...] attribute of type capital_run_length_average: average length of uninterrupted sequences of capital letters.

1 continuous integer [1,...] attribute of type capital_run_length_longest: length of longest uninterrupted sequence of capital letters.

1 continuous integer [1,...] attribute of type capital_run_length_total: sum of length of uninterrupted sequences of capital letters or total number of capital letters in the e-mail.

1 nominal {0,1} class attribute of type spam: denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail.

For a complete list of my features available, please see the second page of this document.

Known Unknowns

I also found a second dataset on Kaggle containing 2500 unprocessed emails with labels (Spam or Ham). If I can somehow parse through them and get the same 58 features as my previous dataset, then I will have additional training data, will also have additional opportunity of finding signal not in the other dataset.

Feature	Type		Feature	Type
word_freq_make	continuous.		word_freq_lab	continuous.
word_freq_address	continuous.		word_freq_labs	continuous.
word_freq_all	continuous.		word_freq_telnet	continuous.
word_freq_3d	continuous.		word_freq_857	continuous.
word_freq_our	continuous.		word_freq_data	continuous.
word_freq_over	continuous.		word_freq_415	continuous.
word_freq_remove	continuous.		word_freq_85	continuous.
word_freq_internet	continuous.		word_freq_technology	continuous.
word_freq_order	continuous.		word_freq_1999	continuous.
word_freq_mail	continuous.		word_freq_parts	continuous.
word_freq_receive	continuous.		word_freq_pm	continuous.
word_freq_will	continuous.		word_freq_direct	continuous.
word_freq_people	continuous.		word_freq_cs	continuous.
word_freq_report	continuous.		word_freq_meeting	continuous.
word_freq_addresses	continuous.		word_freq_original	continuous.
word_freq_free	continuous.		word_freq_project	continuous.
word_freq_business	continuous.		word_freq_re	continuous.
word_freq_email	continuous.		word_freq_edu	continuous.
word_freq_you	continuous.		word_freq_table	continuous.
word_freq_credit	continuous.		word_freq_conference	continuous.
word_freq_your	continuous.		char_freq_;	continuous.
word_freq_font	continuous.		char_freq_(	continuous.
word_freq_000	continuous.		char_freq_[	continuous.
word_freq_money	continuous.		char_freq_!	continuous.
word_freq_hp	continuous.		char_freq_$	continuous.
word_freq_hpl	continuous.		char_freq_#	continuous.
word_freq_george	continuous.		capital_run_length_average	continuous.
word_freq_650	continuous.		capital_run_length_longest	continuous.