Spam Data Analysis

Krishna Padayachee

January 8, 2019

The Spam Data

The Spam Data is taken from the Spam Database that can be obtained from ftp.ics.uci.edu and in addition, the UCI Machine Learning Laboratory (https://archive.ics.uci.edu/ml/datasets/spambase) has a repository that contains the data together with a description of the data. The Spam Datatabase (SPAMBASE) contains the Spam Dataset, Spam Documentation, and an description of the Spam Dataset names. We will briefly mention the following:

1. The Spam Data
2. Spam Data names

The Spam Data

Each row entry in the Spam Dataset represents an email message - the contents of the message are analyzed and word counts are performed for various designated words or keywords including punctuation marks such as ‘square’, ‘curly’, and’round’ brackets; ‘exclamation, and ’question’ marks. Sequences of capital letters are also recorded. Each row is designated as spam or non-spam and is marked as a 1 (spam) or -1 (non-spam). The Spam Data contains 57 ‘explanatory variables’ and 1 response variable named ‘spam_indic’.

The Spam Data names

A full description of the variable names is contained in the SPAMBASE at https://archive.ics.uci.edu/ml/datasets/spambase

We have altered the variable names slightly to make them more meaningful for our purposes.

Spam Analysis

The ADABOOST ML algorithm is applied to the spam data. ADABOOST is a gradient boosting algorithn with a binary response and is the similar to the algorithm for ADABOOST found in the book ‘The Elements of Statistical Learning’ by Friedman, Hastie and Tshibirani (Springer Verlag 2nd Edition)

The following results are shown:

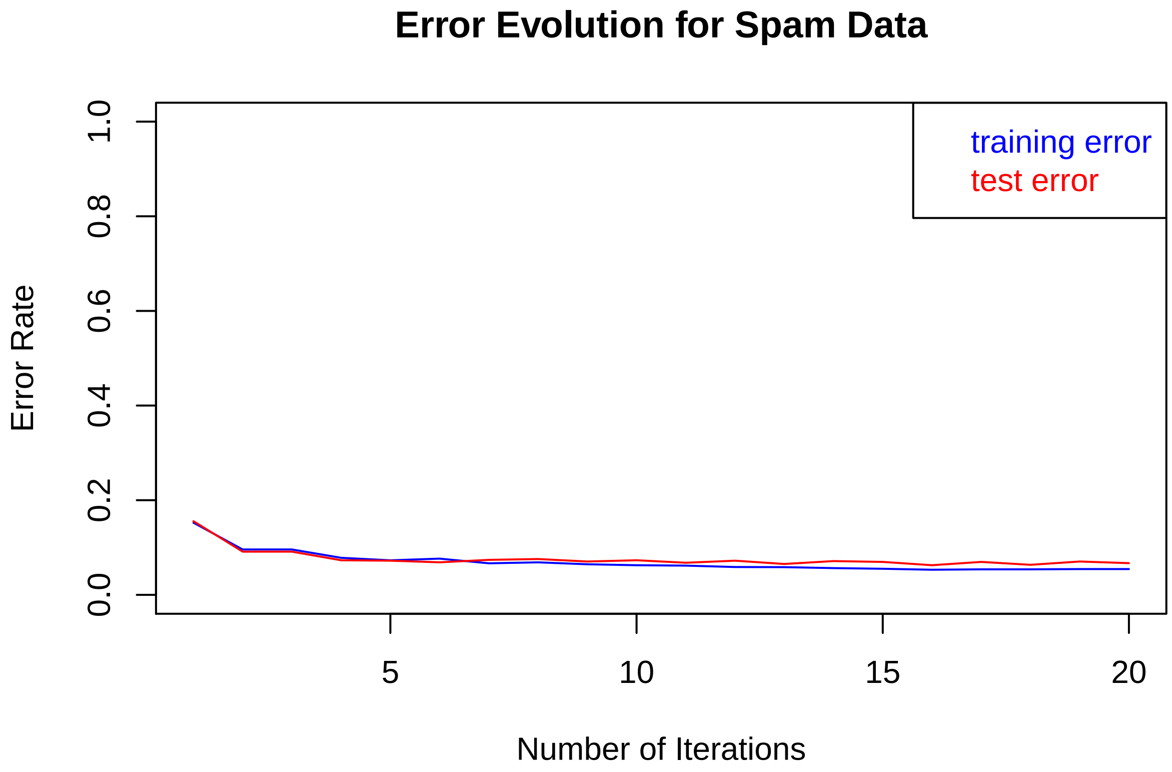
- The misclassification errors of the training and test data are listed as a function of the number of iterations.
- The important variables in the model are displayed on a barchart in increasing order of importance.
- The training and test errors are plotted against the number of iterations performed.

Spam Email Detection

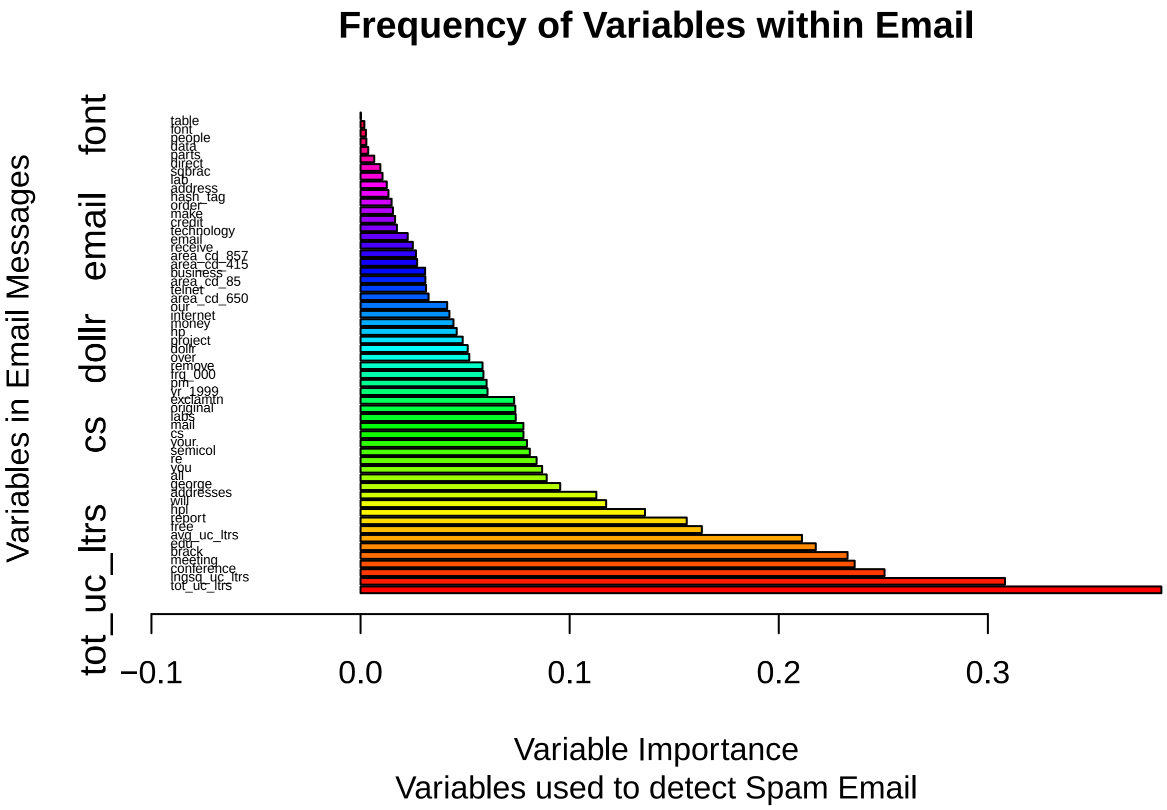
We plot the following:

- Important Variables in Spam Detection

- Error Evolution in applying a ML algorithm (AdaBoost) to the Spam Dataset
  - Error Evolution gives a measure of the predictive accuracy of the algorithm



```
      Training Error Test Error
 [1,]     0.15241959 0.15565217
 [2,]     0.09591423 0.09130435
 [3,]     0.09591423 0.09130435
 [4,]     0.07823819 0.07304348
 [5,]     0.07302231 0.07217391
 [6,]     0.07649957 0.06869565
 [7,]     0.06664735 0.07391304
 [8,]     0.06867575 0.07565217
 [9,]     0.06461895 0.07043478
[10,]     0.06259055 0.07304348
[11,]     0.06172124 0.06782609
[12,]     0.05882353 0.07217391
[13,]     0.05853376 0.06521739
[14,]     0.05650536 0.07130435
[15,]     0.05505651 0.06956522
[16,]     0.05302811 0.06260870
[17,]     0.05389742 0.06956522
[18,]     0.05389742 0.06347826
[19,]     0.05447696 0.07043478
[20,]     0.05447696 0.06695652
```


Frequency of Variables within Email
Variables in Email Messages
tot_uc_ltrs
cs
dollr
email
font
table
font
people
data
parts
direct
sqbrac
lab
address
hash_tag
order
make
credit
technology
email
receive
area_cd_857
area_cd_415
business
area_cd_85
telnet
area_cd_650
our
internet
money
hp
project
dollr
over
remove
frq_000
pm
yr_1999
exclamtn
original
labs
mail
cs
your
semicol
re
you
all
george
addresses
will
hpl
report
free
avg_uc_ltrs
edu
brack
meeting
conference
lngsq_uc_ltrs
tot_uc_ltrs
−0.1
0.0
0.1
0.2
0.3
Variable Importance
Variables used to detect Spam Email