Assignment 2: SPAM Detection with Decision Trees

MEF - BDA 503

Nov 21, 2017

Original library is in UCI Database. See documentation on the website for further detail.

Your assignment consists of buiding a CART model to detect spam mail using UCI's Spambase data and analyze it. You performance depends on correct specification of spam/non-spam mails in the test subset. You are going to use the RData file associated with your assignment on Moodle. Report your way of thinking, methodology, code and results.

You can load the data by using `load` command from your working directory or anywhere if you specify the path. For some installations, you can also double click the on the RData file to load. Name of the data frame is `spam_data` (same as the file name).

```
load("spam_data.RData")
head(spam_data)
```

Column names and short explanations are given below. For further details see the UCI documentation given in the above link.

train_or_test - 0 train, 1 test

spam_or_not - 0 not spam, 1 spam

V1 - word_freq_make

V2 - word_freq_address

V3 - word_freq_all

V4 - word_freq_3d

V5 - word_freq_our

V6 - word_freq_over

V7 - word_freq_remove

V8 - word_freq_internet

V9 - word_freq_order

V10 - word_freq_mail

V11 - word_freq_receive

V12 - word_freq_will

V13 - word_freq_people

V14 - word_freq_report

V15 - word_freq_addresses

V16 - word_freq_free

V17 - word_freq_business

V18 - word_freq_email

V19 - word_freq_you

V20 - word_freq_credit

V21 - word_freq_your

V22 - word_freq_font

V23 - word_freq_000

V24 - word_freq_money

V25 - word_freq_hp

V26 - word_freq_hpl

V27 - word_freq_george

V28 - word_freq_650

V29 - word_freq_lab

V30 - word_freq_labs

V31 - word_freq_telnet

V32 - word_freq_857

V33 - word_freq_data

V34 - word_freq_415

V35 - word_freq_85

V36 - word_freq_technology

V37 - word_freq_1999

V38 - word_freq_parts

V39 - word_freq_pm

V40 - word_freq_direct

V41 - word_freq_cs

V42 - word_freq_meeting

V43 - word_freq_original

V44 - word_freq_project

V45 - word_freq_re

V46 - word_freq_edu

V47 - word_freq_table

V48 - word_freq_conference

V49 - char_freq_;

V50 - char_freq_(

V51 - char_freq_[

V52 - char_freq_!

V53 - char_freq_$

V54 - char_freq_#

V55 - capital_run_length_average

V56 - capital_run_length_longest

V57 - capital_run_length_total