BOCCONI UNIVERSITY

Spam Detection

20596 – MACHINE LEARNING

1 PROBLEM DESCRIPTION

The dataset consists of information from 4601 **e–mail messages**, in a study to screen e–mail for `spam`. For 3101 of these e–mails you known whether they are `spam` [`response` = 2] or `non-spam` [`response` = 1], and you have additional input variables describing several features of each e–mail. For the other 1500 e–mails, you have only information on the inputs.

Your goal is to construct a classifier which has good performance in labeling the remaining 1500 e–mails as `spam` or `non-spam`. **Note** that in classifying as `spam` an e–mail which is actually `non-spam` you pay a cost of 5. Classifying as `non-spam` e–mails which are actually `spam` is less dangerous and hence you pay a cost of 1.

There are **57 input variables** which are described below.

- 48 quantitative predictors measuring the **percentage of words** in the e–mail that match a given word. Examples include `business`, `address`, `internet`, etc ...
- 6 quantitative predictors measuring the **percentage of specific characters/symbols** found in the e–mail. These are `charSemicolon`, `charRoundbracket`, `charSquarebracket`, `charExclamation`, `charDollar` and `charHash`.
- Average length of uninterrupted sequences of **capital letters** [`capitalAve`]. Length of the longest uninterrupted sequence of capital letters [`capitalLong`]. Sum of the length of uninterrupted sequences of capital letters [`CapitaTotal`].