Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2016:9**
http://www.math.su.se

Logistic Regression for Spam Filtering

Niclas Englesson*

June 2016

Abstract

Unsolicited bulk emails, also known as spam emails, are a regular occurrence for anyone who uses email. *Spam filtering* is a way to distinguish between spam emails and regular emails. The goal with spam filtering is to determine whether an email is spam or not spam, then filtering out the spam emails, resulting in a spam-free in-box for the user.

Logistic regression is a statistical method that can be utilized for spam filtering. It is sensible that spam emails typically share a certain type of characteristics. Words that recurrently show up in spam emails can be used as predictor variables in the logistic regression model. Other email characteristics, such as special formatting, tables, links, may also be used as predictor variables. More on this in section **??**.

This report looks into what determines the probability of an email being a spam email by using logistic regression. We will examine if certain characteristics alter the probability of an email being a spam email or not. We will also test which model best predict the probability of an email being spam.

The study initially contain 12 variables that potentially could alter the probability of an email being spam. After variable selection, the number of explanatory variables decrease to 5 in the chosen model. The variables shown to be insignificant are excluded from the model while all the significant variables are kept in the model. It appears that factors such as *winner*, *mult_sent*, *prior_corr* and *sender_spam* have a significant effect on determining the probability of an email being spam, and thus should be used in a spam filter.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: niclas.englesson@gmail.com. Supervisor: Gudrun Brattström and Jan-Olov Persson.