National University of Singapore
School of Computing

SWS3018 Predictive Analytics
Lab 4

Learning Objectives

- Perform logistic regression classification using R

1. In this exercise, we will be using the `Email Spam` (`email_spam.csv`) data set which can be downloaded from Liminus. This data set contains 3921 different email characteristics. The variables are:

   - **spam** : class that determines whether email is a spam (0 – not spam, 1 – spam)
   - **to_multiple** : whether the email is sent to multiple recipients
   - **from** : whether the message was listed as from anyone (usually set by default for outgoing mails)
   - **cc** : whether anyone is CCed
   - **sent_email** : whether the sender has been sent an email in the last 30 days.
   - **image** : whether email contains any image
   - **attach** : whether email contains any attachment
   - **dollar** : whether a dollar sign or the word "dollar" is used in the email
   - **winner** : whether "winner" is used in the email
   - **inherit**: whether "inherit", "inheritance", etc are used in the email
   - **password** : whether "password", etc are used in the email
   - **num_char** : number of characters in the email, in thousands
   - **line_breaks** : number of line breaks in the email
   - **format** : whether the email is written using HTML or plaintext
   - **re_subj** : whether "Re:" is included at the start of the email subject
   - **exclaim_subj** : whether an exclamation point is used in the email subject
   - **urgent_subj** : whether "urgent" is used in the email subject
   - **exclaim_mess** : number of times exclamation point is used in body
   - **number**: whether there is mention of number in the body of the email. (none, small (under 1 million), big)

   In this lab exercise, we will be learning how to classify email to be either spam or non-spam

   a) Generate a logistic regression model using **to_multiple** as the predictor to classify the response (**spam**). Explain the results.

   b) Determine the probability of p(spam=1 | to_multiple=yes) using the following command:

   ```
   predict(model, data.frame(to_multiple=as.factor("yes")), type="response")
   ```

   What is the probability of an email being spam given that it contains multiple

recipients compared to that with only a single recipient?

c) Try using one of the remaining predictors to classify the email spam status.

d) Try using all the predictors to fit into a logistic regression model (`model2`).

e) Using `model2`, is an email with image more likely to be spam? How about email with attachment? What if the email has both image and attachment?

f) Which of the variables are useful for identifying spam? Build another model (`model3`) using just these variables. Which of the predictor is the predictor have the largest effect? Discuss.