MATH 456 — Mathematical Modeling

Assignment #6: Logistic Regression for Spam Email Detection **Due Date:** April 6, 2022, 11:59 PM

The attached data set consist of 4601 email items, of which 1813 items were identified as spam. This data frame contains the following columns:

- crl.tot: total length of words in capitals
- dollar: number of occurrences of the $ symbol
- bang: number of occurrences of the ! symbol
- money: number of occurrences of the word 'money'
- n000: number of occurrences of the string '000'
- make: number of occurrences of the word 'make'
- yesno: outcome variable, a factor with levels n not spam, y spam

Goal: Apply the logistic regression model predict whether a future email is spam or not based on these explanatory variables.

Here are some concrete steps to achieve this goal.

1. Split the dataset randomly into a training set and a test set. For example you may take 4000 obs in training and 601 obs in test. Note that you should make a relatively even split of spam and nonspam emails.
2. Transform the explanatory variables or features using $\log(x_{ij} + 0.1)$. This is mainly because many explanatory variables are zero.
3. Fit the data with the logistic regression model. Report the mean error rate on the training and test sets. Discuss your findings.

You need to implement logistic regression using either Python or Matlab. You are encouraged to use any available modules in Python or Matlab for logistic regression. Or you could build your own logistic regression solver using gradient descent or Newton's method as the optimizer. In either case, you need to document your results in Word/PDF/Markdown/Jupyter Notebook. Upload the document along with the codes.