Homework 7

MSCS 6520 Business Analytics
Spring 2018
Assigned: April 23, 2018
Due: April 30, 2017 (by beginning of class)

Exercises

For this assignment, you're going to build a model to classify email as spam or ham. We're going to use data from a competition called trec07p. I have already parsed the messages of the email bodies to create a "bag of words" model. I selected ~3,000 predictive words, excluding those with non-ASCII characters and numbers. Except for the label column, each column corresponds to a single word. A value of 1 indicates that the word is present in the email, while a value of 0 indicates that the word is not present in the email. A label of 1 indicates that the message is spam, while a label of 0 indicates that the message is not. You do not need to do any centering or scaling (no preprocessing, basically.)

Note that this data set has too many features to do forward or backward feature selection. We will rely on Logistic Regression's ability to select features.

1. Download the data set from D2L. The zip file contains a single file (spam_data.csv).
2. Build a single model to predict spam vs ham. Include all of the features – do not attempt to perform forward or backward feature selection
3. What were the accuracy and confusion matrix for your model?
4. Using the confusion matrix, calculated the percentage of spam messages erroneously classified as ham and ham messages erroneously classified as spam. Are the percentages equal? Often times, one type of error is preferably for a business or user. In this case, it would be worse for an important email to end up in the spam folder than for a spam message to end up in the inbox.

Prepare a document containing the answers to the above questions. Submit the document as a PDF to D2L. You may work in pairs, in which case, you should only submit one PDF per group.