STAT 5525 Homework 2

This homework involves classifying emails as spam or not spam based on the frequency of occurrence of words in them. The data were taken from here: <https://archive.ics.uci.edu/ml/datasets/spambase> These data were also used in a Kaggle competition.

There are 3 files that are relevant to this homework:

* spambase.csv which is the data to analyze in tidy format
* spambase.DOCUMENTATION the original document that explains the data (a text file that can be opened with any simple text editor)
* spambase.names.csv which gives the correspondence between the names of the variables in the spambase.csv file and the original names that are given in the documentation. I changed the names when tidying the data for you.

Assignment (see due date below):

Part 1 Logistic Regression (worth 5 points out of 15):

1. Set aside a 20% sample to be a test dataset
2. Fit a logistic model to all of the training data and display the summary.
3. Evaluate the fit and simplify the model by eliminating some predictors. Fit the simplification. (You may use techniques from the Advanced Regression course.) Explain why you chose to eliminate certain predictors and the approach or any technique(s) you used to do so. (Hint: a simple approach based on significance is adequate for this homework.) Display the summary.
4. Using the refitted model, the estimated fitted values and a threshold of 0.50 for deciding spam, display a confusion matrix and calculate the overall error rate and the false positive rate.
5. Zero false positives are the goal. Using the fitted values find the threshold value that produces approximately zero false positives and re-display the confusion matrix

Part 2: LDA (5 points):

1. Perform an LDA on the data using only the predictors you decided upon for the simplified logistic regression model. Display the summary results.
2. Interpret the summary results as best you can. We did not discuss this in class so this is a challenge step. A round of brews or “mocktail” to those who try this, once we can get back together.
3. As with the logistic model, use the fitted probabilities to establish a threshold that achieves near zero false positives.
4. Display the confusion matrix from step 3. Using only this and the confusion matrix from step 5 above, which model would you recommend to use? Explain.

Part 3: Cross-Validation (5 points)

1. Perform an iterated 5-fold cross validation of 500 iterations each using both models only the original training data as follows. Use total error rate as the measure or “cost” function, not false positives.
2. Similar to what I did in the lab calculate the error rate on the non-fold data (i.e. training error rate) and the validation error on the fold for each fold for each of the 500 iterations. However, at the end do not take the mean of the iterations. Rather keep the separate results. Therefore, at the end you should have 500 values each for non-fold training errors and validation (fold) error rates for both the logistic regression and the LDA fit. In other words, you should have produced 500 estimates of 2 overall error rates for each of 2 models. If you build a dataframe from this you should have one consisting of 500 rows and 4 columns.
3. Display a ggplot consisting of 4 side-by-side boxplots from the dataframe described in step 2.
   1. Challenge: Try discussing the similarities between this and doing a bootstrap. Is this like doing a bootstrap?
4. Based on the boxplots and consideration of the false positive goal, which model would recommend? Explain.
5. Using the fits from both models, classify the observations in the held-out test data and display the corresponding confusion matrices. (You should have 2 matrices.) Based on this do you stand by your recommendation in step 4? Explain.

You may consult online resources, but if you do you need to cite those resources as any good analyst or researcher would do. You may reuse any of the code from class.

Submission: Prepare your assignment as an RMarkdown file with all code present and a narrative that explains your approach. Generate a pdf from the markdown. In the pdf you may hide any code that is not central to your narrative. I’m looking for a report written in a professional style.

Due date: End of day July 22.