Graded Assignment 4
(Due Thu 12/07)

ISOM 674, Fall 2017

This is a team assignment.

In this assignment use un-coded spam data set. As in the previous assignment, the data is split these data into a training sample and a validation sample using a 60-40 split and we are not going to use a test sample. I have provided a "helper" R script file that loads the data and splits into the training and validation data sets using the permutation provided. This file also includes a function to compute the log-likelihood that solves a numerical problem.

Questions

(1) Using the spam data with the original continuous feature (i.e., the un-coded data), perform ridge regression using logistic regression (family="binomial"). Use AUC (calculated on the validation data) as the performance criterion. Make a plot of AUC vs complexity. You will have to experiment to find a grid of λ values that will give you a useful plot.
(2) Repeat (1) using the lasso instead of ridge regression.
(3) For the lasso regression, make the plots of AUC against the number of included variables (i.e., variables with non-zero coefficients). Since none of the coefficients is likely to be exactly 0 numerically, you will have to think about what this means and how to make the plot.
(4) Are you getting the behavior you expect? Why or why not? In answering this question, address both the results of the ridge regression and the lasso regression.
(5) To see if you get the same behaviors using a difference criteria that the AUC, repeat (1)-(4) above using the log-likelihood computed on the validation data as the performance criterion.

Note: This assignment can be split up between the team members. I would estimate that there is about 4 hours of work needed here.

Submit your solutions as an R notebook using the link provided. Name the file

GradedHW4-TeamX.Rmd

where you substitute your team number for the X. Put the names of your team members at the top of the file.