Graded Assignment 3

(Due Thu 11/2)

ISOM 674, Fall 2017

In this assignment we will use the coded and un-coded spam data set. I have split these data into a training sample and a validation sample using a 60-40 split. We are not going to use a test sample in this exercise. If this were a "real" problem I was working on, since the sample size is moderate, I would most likely use something like a 75-25 split with the 25% sample used as a test data set and with 10-fold cross validation used to provide both training and validation for the 75% sample. I did not want to add the complexity of K-fold cross-validation to this exercise. However, I also do not want you to think it is OK just to "throw away" the test sample.

Questions

In the following questions, use AUC using the validation data as your final performance criterion. Also use the split of the data into training and validation samples that I used in class (which was based on the random permutation in the file "SpamdataPermutation.RData").

(1) Read questions 2 and 3 below. In finding the best 10 feature models, think about whether or not you want to build the models on the training data set and then test the "built" model on the validation data set or whether you want to use the validation data set to determine what the best variables are. Indicate the strategy you take and explain what the likely effect is on possible overfitting to the validation data set. Indicate the pros and cons of you decision and why you made the decision you did.

(2) Using the coded spam data set, find the best naïve Bayes model that you can find that uses no more than 10 features.

(3) Using the coded spam data set, find the best logistic regression model that you can find that uses no more than 10 features.

(4) Using the un-coded spam data set, find the best logistic regression model that you can find that uses no more than 10 features.

(5) Of these three models which seems to be better? How much of an effect did coding the variables seem to have?

(6) Using one other technique that we have learned about, find the best model that uses no more than 10 features and compare its performance to the other models.

(7) Read Chapter 17 in Alpayden. This chapter discusses ensemble methods (Combining Multiple Learners). You will see that this can become quite complex.

(8) Construct the best ensemble approach that you can based on the naïve Bayes approach, one of the two logistic regression approaches, and the additional approach you selected for item (6). Base your ensemble approach on combining the probabilities from the

models. Keep things very simple. How does the performance of the ensemble approach compare to the performance of the individual models.

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

GradedHW3-TeamX.Rmd

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