STA 4364 Midterm 2

due Wednesday November 23 by 11:59PM on Webcourses

Submission Format: Please submit your midterm as both 1) a HTML or pdf document, and 2) also submit the source file in either R Markdown or Jupyter notebook format (at most one of each type of file).

Problem 1: (25 points) This problem will involve logistic regression on the dataset `midterm_data_2.csv`. The response column is `response` and all other columns are features.

(a) (5 points) Load the dataset. Remove any unnecessary columns. For any columns that have `NA` values, fill in the `NA` values with the median over all non-missing entries in the columns. Format all columns with string entries as categorical variables. Make `response` a categorical variable. Split the dataset into a training set (75% of observations) and validation set (25% of observations).

(b) (5 points) Make a model using all features. Narrow down your features to make a reduced model that uses only the most relevant predictors.

(c) (5 points) Create an ROC curve for your full and reduced model on both the training and validation sets (4 curves in all). Comment on the degree of overfitting for validation performance vs. training performance and the adequacy of your reduced model compared to your full model.

(d) (5 points) Using your reduced model, perform predictions for $P(\texttt{response} = 1|\texttt{features})$ for the validation set. Perform predictions for the binary `response` by thresholding your predicted probabilities $P(\texttt{response} = 1|\texttt{features})$ at two different values: 0.5 and 0.65. Calculate the overall prediction accuracy for both thresholds. Calculate the False Negative Rate for both thresholds.

(e) (5 points) Make two altered copies of your validation set: one where `feat.d` is set to 1 for all rows, and another where `feat.d` is set to 0 for all rows. All other columns should remain the same as your original validation set. Using your reduced model, perform predictions for $P(\texttt{response} = 1|\texttt{features})$ for both altered validation sets, and average the predicted probabilities across all validation observations (end up with 2 average probabilities, one for each altered dataset). Finally, calculate the difference between these average probabilities (either order for the subtraction is OK). How can you interpret the average difference that you have found?