STA 4364 Midterm 1

due Wednesday October 14 by 11:59PM on Webcourses

Submission Format: Please submit your midterm as **a single file in either R Markdown or Jupyter notebook format**. R and Python are both acceptable languages for your exam.

Note: All data in this exam is synthetic data.

Problem 1: (25 points) This problem will involve linear regression on the dataset `midterm_data_1.csv`. The response column is `response` and all other columns are features.

(a) (5 points) Load the dataset. Remove any unnecessary columns. Remove any rows that have an `NA` value. Format the columns for `feat.c` and `feat.g` as categorical variables. Split the dataset into a training set (80% of observations) and validation set (20% of observations). Make pairwise plots showing the relations between all columns. Compute the pairwise correlations between all numerical columns.

(b) (5 points) Make a linear model using all features. How can you interpret the coefficients of `feat.c`? What does R^2 signify? How can you interpret the value of the residual standard error? What does the F-statistic say about your model? Make a residual plot of the residuals vs. fitted values and comment on what this says about validity of the linearity and constant-variance error assumptions of the model.

(c) (5 points) Make a linear model that includes all interactions between features and all quadratic terms for numerical features. From this model, identify a reduced set of coefficients that are the most relevant predictors. Look at the residual plot of your reduced model and comment on any observed differences between this plot and the residual plot from part b).

(d) (5 points) Calculate the MSE and the R^2 value on the validation set using your full quadratic model and your reduced model. Comment on the degree of overfitting compared to the model performance on training data and the adequacy of your reduced model compared to your full model.

(e) (5 points) Using your reduced model, calculate a 95% confidence interval for each validation set prediction (you can do this using the `predict` function in R). Calculate the percentage of true observations from your validation set that fall within your prediction interval. (For this problem, you don't need to print all of the confidence intervals. Please only print the final value of the number of true observations that fall within your confidence interval).

Problem 2: (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 (60% of observations) and validation set (40% 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?