MSiA 400 Lab Assignment 4

Due Dec 3rd at 11:59pm

Instructions: Please submit a report file that includes: a short answer, related code, printouts, etc. for each problem (where necessary). Push your answers to Github or Canvas. All programming must be in R (or R Markdown).

Problem 1

For this problem, you will evaluate the `redwine.txt` dataset. For 1599 variants of red wine, it contains 1 response variable, wine quality (`QA`), and 11 explanatory variables: fixed acidity (`FA`), volatile acidity (`VA`), citric acid (`CA`), residual sugar (`RS`), chlorides (`CH`), free sulfur dioxide (`FS`), total sulfur dioxide (`SD`), density (`DE`), pH (`PH`), sulphates (`SU`), and alcohol (`AL`). Note: some variables contain missing values, you will handle them in Problem 2; for now, you may ignore them.

This dataset comes from the UCI repository: https://archive.ics.uci.edu/ml/datasets/wine+quality

Note: you must include all plots, either in the body of your assignment solutions or as an appendix.

Problem 1a

For each of the 12 variables, plot the distribution using either a histogram or stem-and-leaf plot, whichever you deem more appropriate (for each variable).

Problem 1b

For each variable, plot the distribution using a box-and-whisker plot. Are there any significant outliers? Note: some variables have comparable scales, while others do not.

Problem 1c

For each variable, compute the skewness and kurtotsis. Which variables are skewed left, skewed right, and not significantly skewed? Which are mesokurtic (kurtosis around 3), leptokurtic (kurtosis > 3, fat tails), and platykurtic (kurtosis < 3, thin tails)?

Problem 1d

For each variable, display the Q-Q plot. Do they confirm your observations from previous parts of Problem 1?

Problem 2

Your goal is use the `redwine.txt` dataset (the same dataset evaluated in Problem 1) to predict the wine quality given the other variables. You will use various methods to fill missing values.

Problem 2a

Use `is.na` to determine which variables have missing values. How many missing values are there in each variable (Hint: use `colSums`)? How many samples have missing values (Hint: use `rowSums`)?

Problem 2b

Split the dataset into 5-folds. For each fold, use random sampling (from the training set) to fill in the missing values. Train a linear regression model. Compute the Mean Squared Error (MSE), i.e. compute $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ (where N is s the size of the dataset, y_i is the actual wine quality, and $\hat{y}_i$ is the predicted wine quality). Record the average MSE for both the training and test datasets.

Problem 2c

Repeat the process in Problem 2b, using the most common value method to fill missing values. Use the same CV folds.

Problem 2d

Repeat the process in Problem 2b, using the average value method to fill missing values. Use the same CV folds.

Problem 2e

Repeat the process in Problem 2b, using the k-NN method (with 5 nearest neighbors) to fill missing values. Use the same CV folds.

Problem 2f

Repeat the process in Problem 2b, using the MICE method (with PMM) to fill missing values. Use the same CV folds.

Problem 2g

Repeat the process in Problem 2b, omitting samples with missing values. Use the same CV folds.

Problem 2h

Which method for handling missing values performs best? Why may this be?