Homework 3: Regression

Due September 20, 2018 at 2:10 PM

This homework will investigate various regression techniques to predict the percent body fat given a variety of information about body measurements and estimated body fat percentage.

The file hwkdataNEW.mat contains data related to body measurements and estimated body fat percentage. **MAKE SURE YOU DOWNLOAD THE NEW HOMEWORK DATA FILE - Five suspicious observations have been removed.** The matrix x contains 14 predictor variables:

1. Age (yrs)
2. Weight (lbs)
3. Height (inches)
4. Adiposity Index (kg/m^2)
5. Neck circumference (cm)
6. Chest circumference (cm)
7. Abdomen circumference (cm)
8. Hip circumference (cm)
9. Thigh circumference (cm)
10. Knee circumference (cm)
11. Ankle circumference (cm)
12. Extended bicep circumference (cm)
13. Forearm circumference (cm)
14. Wrist circumference (cm)

The column vector y contains the output variable, percent body fat. The correct citation for these data is: Penrose, K., Nelson, A., and Fisher, A. (1985), "Generalized Body Composition Prediction Equation for Men Using Simple Measurement Techniques", *Medicine and Science in Sports and Exercise*, 17(2), 189.

1. Divide the data into training, test, and validation data sets. In your report, describe how you divided the data into these three sets and *why it is appropriate.*

2. Develop a linear regression model to predict the percent body fat using the 14 other variables. Then, select a subset of these variables and develop a competing model. Use the correlation coefficients to explain why this representation works. Remember to pad your inputs with a column of ones to allow for a non-zero y-intercept.

3. Find the **single** best predictor of percent body fat and the **pair** of predictors that performs best.

4. Try to identify or derive any additional inputs that may be good predictors of percent body fat based on nonlinear relationships of the available predictors. Use linear-in-parameters regression with your non-linear term(s) to evaluate any model improvement.

5. Compare the performance of all models using the root mean squared error (RMSE) of the test data set. Select the *best* model. Explain why it is the best.

6. Find the validation error of your *best* model.

For this homework, prepare a written report in IEEE format. Include any plots and tables that will support your findings. Make sure you correctly label your figures and tables and refer to them in the text. Include an appropriate citation for the data, both in the text and in the list of references. Your report should include **at a minimum** an abstract, introduction, methodology, results (and discussion!), conclusions, and references. Note that the methodology section of this report (and every report!) should describe the algorithm that you're using – not the implementation in MATLAB. Include all your code in an appendix (single column) at the end of the report. Convert your report to .pdf before submitting it through Canvas.