Homework 6: Ridge Regression

Due October 16, 2018 at 2:10 PM

This homework will investigate ridge regression to predict the percent body fatgiven 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. 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.

This assignment will look at the predictive ability of partial least squares regression (PLS) and compare it to the methods we've investigated previously.

1. Briefly describe ridge regression (including the benefits for ill-posed problems - talk about the condition number!).

2. Highlight the value of ridge regression with an extremely ill-conditioned simulated data set (such as the one shown below).

```
t = [1:100]';
x1 = sin(t) + 0.01*rand(size(t));
x2 = 10*sin(t) + 0.01*rand(size(t));
x = [x1 x2];
y = 0.1*x(:,1) + 0.9*x(:,2) + rand(size(t));
```

   Note that the discussion in (1) and the demonstration in (2) should comprise the bulk of your methodology section!

3. Divide the data into training, test, and validation data sets. You **must** use the same training, test, and validation data sets that you used in previous homeworks.

4. Determine the appropriate regularization coefficient through two methods: the L-curve method and cross-validation. Compare and explain any difference in the results of the two methods.

5. Select the *best* ridge regression model from the two you trained in the prior step. Explain why it is the best model in terms of both accuracy and stability.

6. Compare the validation performance of your ridge regression model to the PLS, PCR, and regression models. Comment on the results.

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.