Homework 4: Principal Component Analysis and Regression

Due September 27, 2018 at 2:10 PM

This homework will investigate PCA and PCR techniques 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.

First, you will explore the **input** space using PCA. Then you will use the principal components to predict percent body fatwith linear regression (PCR).

1. 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 Homework 3.

2. Perform PCA of the standardized input training data. Analyze the results of the PCA. What can you say about the true dimension of the input space? What do the PCA loadings tell you about relationships between the variables?

3. Develop several competing regression models using the PC scores of the data (PCR). First, select enough PCs to explain at least 90% of the information in the input space. Then, choose the significant PCs based on eigenvalues. Finally, select PCs most useful for predicting percent body fat. You should end up with at least three PCR models.

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

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

6. Compare the validation performance of your *best* PCR model with that of your *best* regression model from Homework 3.

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.