ST563 - Homework 4

Wenbin Lu

1 Instructions

Please follow the instructions below when you prepare and submit your assignment.

- Include a cover page with your homework. It should contain
  1. Full name
  2. Course #
  3. HW #
  4. Submission date
- Assignments should be submitted using **Gradescope**.
- Neatly typed work should be submitted. All R code/output should be well commented on, with relevant outputs highlighted and discussed.
- When you solve a particular problem, do not only give the final answer. Instead, show your work and the steps you used (with proper explanation) to arrive at your answer to get full credit.
- Submission in the PDF format is preferred; Please convert other formats such as doc or docx into PDF.

Problem 1 (30 points)

This question uses the variables `dis` (the weighted mean of distances to five Boston employment centers) and `nox` (nitrogen oxides concentration in parts per 10 million) from the Boston data. We will treat `dis` as the predictor and `nox` as the response.

(a) Use the poly() function to fit a cubic polynomial regression to predict `nox` using `dis`. Report the regression output, and plot the resulting data and polynomial fits.

(b) Plot the polynomial fits for a range of different polynomial degrees (say, from 1 to 10), and report the associated residual sum of squares.

(c) Perform cross-validation or another approach to select the optimal degree for the polynomial, and explain your results.

(d) Fit a smoothing spline model to predict `nox` using `dis` using four degrees of freedom. How did you choose the knots? Plot the resulting fit.

(e) Now fit a smoothing spline for a range of degrees of freedom, and plot the resulting fits and report the resulting RSS. Describe the results obtained.

(f) Perform cross-validation or another approach in order to select the best degrees of freedom for a smoothing spline on this data. Describe your results.

Problem 2 (20 points)

This question relates to the College data set in the ISLR2 library.

(a) Split the data into a training set and a test set. Using out-of-state tuition as the response and the other variables as the predictors, perform forward stepwise selection on the training set in order to identify a satisfactory model that uses just a subset of the predictors.

(b) Fit a GAM on the training data, using out-of-state tuition as the response and the features selected in the previous step as the predictors. Plot the results, and explain your findings.

(c) Evaluate the model obtained on the test set, and explain the results obtained.

(d) For which variables, if any, is there evidence of a non-linear relationship with the response?

Problem 3 (10 points)

Problem 6 of Chapter 4.8 Exercises.

Problem 4 (10 points)

Problem 7 of Chapter 4.8 Exercises.

Problem 5 (10 points)

Problem 9 of Chapter 4.8 Exercises.

Problem 6 (30 points)

Problem 14 of Chapter 4.8 Exercises. The `Auto` data is in `ISLR2` library.