Data624 - Homework8

Amit Kapoor

4/19/2021

Contents

Exercise 7.2 **1**
Models . . . 1
Performance . . . 1

Exercise 7.5 **1**
(a) . . . 1
(b) . . . 2
(c) . . . 2

Exercise 7.2

Friedman (1991) introduced several benchmark data sets create by simulation. One of these simulations used the following nonlinear equation to create data:

$$y = 10sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + N(0, \sigma^2)$$

where the x values are random variables uniformly distributed between [0, 1] (there are also 5 other non-informative variables also created in the simulation). The package `mlbench` contains a function called `mlbench.friedman1` that simulates these data:

Models

Tune several models on these data.

Performance

Which models appear to give the best performance? Does MARS select the informative predictors (those named X1–X5)?

Exercise 7.5

Exercise 6.3 describes data for a chemical manufacturing process. Use the same data imputation, data splitting, and pre-processing steps as before and train several nonlinear regression models

(a)

Which nonlinear regression model gives the optimal resampling and test set performance

(b)

Which predictors are most important in the optimal nonlinear regression model? Do either the biological or process variables dominate the list? How do the top ten important predictors compare to the top ten predictors from the optimal linear model?

(c)

Explore the relationships between the top predictors and the response for the predictors that are unique to the optimal nonlinear regression model. Do these plots reveal intuition about the biological or process predictors and their relationship with yield?