Untitled

Zachary Palmore

7/16/2021

## KJ 7.2

Friedman (1991) introduced several benchmark data sets created by simulation. One of these simulations used the following nonlinear equation to create data: y = 10 sin(πx1x2) + 20(x3 − 0.5)2 + 10x4 + 5x5 + N(0, σ2) 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:

> library(mlbench)  
> set.seed(200)  
> trainingData <- mlbench.friedman1(200, sd = 1)  
> ## We convert the 'x' data from a matrix to a data frame  
> ## One reason is that this will give the columns names.  
> trainingData$x <- data.frame(trainingData$x)  
> ## Look at the data using  
> featurePlot(trainingData$x, trainingData$y)  
> ## or other methods.  
>  
> ## This creates a list with a vector 'y' and a matrix  
> ## of predictors 'x'. Also simulate a large test set to  
> ## estimate the true error rate with good precision:  
> testData <- mlbench.friedman1(5000, sd = 1)  
> testData$x <- data.frame(testData$x)

### Part A

#### Question

Tune several models on these data. An example is shown in the code. Which models appear to give the best performance? Does MARS select the informative predictors (those named X1–X5)?