GR5058 Assignment 4

Due: Tuesday, December 4, 2018 by 6PM

Smooth Nonlinear Models for a Continuous Outcome

Use the `College` dataset in the **ISLR** package, which can be accessed by executing

```r
data(College, package = "ISLR")
str(College, max.level = 1)

## 'data.frame': 777 obs. of  18 variables:
##  $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Apps       : num  1660 2186 1428 417 193 ...
##  $ Accept     : num  1232 1924 1097 349 146 ...
##  $ Enroll     : num  721 512 336 137 55 158 103 489 227 172 ...
##  $ Top10perc  : num  23 16 22 60 16 38 17 37 30 21 ...
##  $ Top25perc  : num  52 29 50 89 44 62 45 68 63 44 ...
##  $ F.Undergrad: num  2885 2683 1036 510 249 ...
##  $ P.Undergrad: num  537 1227 99 63 869 ...
##  $ Outstate   : num  7440 12280 11250 12960 7560 ...
##  $ Room.Board : num  3300 6450 3750 5450 4120 ...
##  $ Books      : num  450 750 400 450 800 500 500 450 300 660 ...
##  $ Personal   : num  2200 1500 1165 875 1500 ...
##  $ PhD        : num  70 29 53 92 76 67 90 89 79 40 ...
##  $ Terminal   : num  78 30 66 97 72 73 93 100 84 41 ...
##  $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
##  $ perc.alumni: num  12 16 30 37 2 11 26 37 23 15 ...
##  $ Expend     : num  7041 10527 8735 19016 10922 ...
##  $ Grad.Rate  : num  60 56 54 59 15 55 63 73 80 52 ...
```

The variables are described under `help(College, package = ’ISLR’).`

(a) Use the `createDataParition` function to split the observations into training and testing.

(b) Use the `lm()` function to predict `Outstate` using whatever transformations, polynomials, cuts, and interactions you feel are necessary to predict well in the testing data.

(c) Use caret to fit a Generalized Additive Model where `Outstate` is the outcome using the predictors from your best model for the training data in part (b). Explain what calling `plot()` on the `finalModel` list element tells you.

(d) Which predictors, if any, exhibit a very non-linear relationship with `Outstate`, conditional on the other predictors?

(e) Is the average squared error in the testing data greater, less than, or about the same than with `lm`?

Fused Lasso Additive Model

We have not discussed the Fused Lasso Additive Model (FLAM) directly, but it is described at

https://channel9.msdn.com/Events/useR-international-R-User-conference/useR2016/Flexible-and-Interpretable-Regression-Using-Convex-Penalties

You can install the package the implments this model (once) via

```r
install.packages("flam")
```

(a) In your own words, describe what the Fused Lasso Additive Model does

(b) Use the flamCV function in the **flam** package to find the optimal values of the tuning parameters and estimate the coefficients of a model with the same predictors as in problem 1. Does the Fused Lasso Additive Model predict better in the testing data than the Generalized Additive Model?

Tree-Based Models for a Binary Outcome

In your home directory on the course server, there is a file called payback.rds that you can download to your working directory and then bring into R with

```
payback <- readRDS("payback.rds")
```

There will now be a data.frame called payback that data on people who were given personal loans and the question is whether the loan was or was not paid back on time. It has the following 19 variables:

1. **loan_amnt**: amount of the loan in dollars
2. **term**: how long the borrower has to pay back the loan
3. **int_rate**: the annual interest rate for the loan
4. **installment**: the amount of money the borrower is scheduled to pay each month
5. **emp_length**: the amount of time the borrower has worked at the current job (0 means less than 1 year, 10 means ten or more years, missing means unemployed)
6. **home_ownership**: a factor indicating how the borrower pays for housing
7. **annual_inc**: the stated annual income of the borrower
8. **verification_status**: whether the stated annual income of the borrower has been verified
9. **purpose**: the stated purpose of what the loan is for
10. **zip_code**: the first 3 digits of the borrower's ZIP code
11. **addr_state**: the state that the borrower lives
12. **delinq_2yrs**: the number of times in the last two years that the borrower has been more than a month behind on any payment (not just the loan in question)
13. **earliest_cr_line**: the year in which the borrower first opened a credit line (for a credit card, etc.)
14. **inq_last_6mths**: the number of formal inquiries by the borrower's creditors in the last 6 months (not just for the loan in question)
15. **open_acc**: the number of open credit lines the borrower currently has (for credit cards, etc.)
16. **pub_rec**: the number of derogatory public records the borrower has
17. **revol_bal**: the total revolving balance the borrower has (from credit cards, etc.)
18. **total_acc**: the number of open credit lines the borrower currently has ever had (for credit cards, etc.)
19. **y**: the binary outcome, which is 1 if the loan was defaulted on, charged off, very behind at the time the dataset was created, etc. and is 0 if the loan was (or was being) fully paid on time

(a) Use the createDataParition function to split the observations into training and testing.

(b) Fit a logit model to the outcome in the training data, using whatever transformations, polynomials, cuts, and interactions you feel are necessary to predict well in the testing data.

(c) Use a BART and then a random forest approach to fit the outcome in the training data.

(d) Rank the three approaches in parts (b) and (c) in terms of which is most likely to yield a correct classification in the testing data.