GR5058 Assignment 3

Due: Tuesday, November 20, 2018 by 6PM

Prediction with Linear Models

Download the (in)famous crime dataset via

```
ROOT <- "https://archive.ics.uci.edu/ml/machine-learning-databases/"
crime <- read.csv(paste0(ROOT, "communities/communities.data"),
                  header = FALSE, na.strings = "?")
colnames(crime) <- read.table(paste0(ROOT, "communities/communities.names"),
                              skip = 75, nrows = ncol(crime))[,2]
```

You can get more information about this dataset from https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime. Split the dataset into training and testing using the `createDataPartition` function in the **caret** package after calling `set.seed()` using the number at the bottom of this page.

Use the following methods via the `train` function in the **caret** package: plsr, lm. Model the `ViolentCrimesPerPop` variable in the `training data.frame`, but you can include interactions, polynomials, and / or new variables that you construct from the other variables. Then use the `predict` function with `newdata = testing` to generate $\widehat{y}_i$ for each observation in the `testing data.frame`. Calculate the mean squared error between $\widehat{\mathbf{y}}$ from $\mathbf{y}$ in the `testing data.frame`. Which function and model produces the lowest mean squared error?

Classification of Binary Outcomes

Dowload the loans.rds file from the course server to your working directory and load it into R via

```
loans <- readRDS("loans.rds")
str(loans, max.level = 1)
```

In these data, the outcome of interest is whether a personal loan was approved by a bank. The variables are

- `Amount.Requested`: The proposed amount for the loan
- `Debt.To.Income.Ratio`: The ratio of the applicant's debt (excluding mortgages and the proposed loan) payments each month to the applicant's stated monthly income
- `Zip.Code`: The 3-digit zip code of the applicant
- `State`: The state where the applicant lives
- `Employment.Length`: The number of years that the applicant has worked at the same job. `10` indicates at least ten years, `0` indicates less than one year, and `-1` indicates unemployed.
- `y`: A binary variable indicating whether the loan was approved

Use the `createDataPartition` function in the **caret** package to split the data into a training set and a testing set. Use the following R functions: glmnet, glm. Estimate classification models for `y` in the training data as a function of other variables in the dataset, possibly including interactions, polynomials, and / or variables you construct. Predict `y` in the testing dataset. You can use a threshold of 0.5 to classify observations in the testing dataset as being approved for a loan or not. Using the proportion of correct classifications in the testing dataset as your criterion, which function and model performs best?

seed = 1184245352