BSDS 100: Intro to Data Science with R
Final Class Project

by James D. Wilson (University of San Francisco)

Objective

Each of the projects written below are computational tasks that will require tools from what you've learned in this course. The aim of this final project is to provide everyone in the class useful computational tools in R. A major component of this project is to make any code efficient, well-documented, and easy to use. Also any plots or output should be crisp, easily understood, and properly labeled.

You can use *some* code from online resources but do not simply copy and paste some one else's code for the entire product. If you are having trouble finding a data set, you may use data from repositories such as `http://archive.ics.uci.edu/ml/`. Be creative! **Don't** use pre-loaded data in R.

Grading Rubric

You will work in groups of 4. Choose one of the projects below. Each group will be required to turn in the following:

- A presentation deck: each group will present a 15 minute presentation on Thursday, November 30th or Tuesday, December 4th in class. You **must** put the final presentation in either powerpoint or latex beamer. Include the following in your presentation:
  - Description of the problem
  - A brief description of any data that you analyze
  - An analysis of your data including any decisions you made along the way. Creativity counts here!
- Any R code used for the project, which is knit into a .pdf file.

Project Choices

1. [**Regression Software**] Suppose that you are given a predictor matrix (or data frame) X and a continuous-valued response y. Create an easy-to-use function that takes as input X and y and performs the following tasks:

   - Linear regression of y on X via a prespecified model-selection algorithm (best subset, forward or backward)
   - Lasso, Ridge, and Elastic Net
   - K-NN for regression
   - Regression trees

   All models should be chosen using cross-validation (if needed), and grid-searches should be run where appropriate. Consider normalization of X and possible transformations for y. Consider also checking the assumptions on any model specified. Explanatory plots should be reported that provide a reason for any parameters chosen (K, λ, or α). For each of the above methods, the function should provide explanatory plots that will help the user analyze the data. Include plots that compare the methods, and inform the user which model to use and why.

   In your presentation, you will be expected to run the code in person to show how it operates and how to understand the output. Note that this function should be a ”one-stop shop” for applying these regression techniques, and should assume that the user does not automatically provide you with correct input.

   For help on getting started for this project, read Chapter 23 of the *R for Data Science* book.

2. [**Additional ML Topics**] Machine learning is an expansive field with many topics that we have not yet covered in class (though many more topics will be discussed next module). In this project, you will first choose one of the following popular areas in machine learning:

   - Clustering
   - Classification
   - Regression
   - Natural language processing
   - Deep Learning
   - Image Segmentation
   - Neural Networks
   - Semi-supervised learning

   The goal of this project is to research the topic chosen from above, present and implement at least one method in this area. Describe the topic, any challenges inherent in the area, and relationships with other topics that we have covered in the class. Apply your chosen method(s) to a data set of your choice, including any analyses and data-driven decisions from machine learning that can help you analyze the data. Remember you are the instructor of this topic, so present in a way that you wish someone would have taught you.

3. [**Case Study**] Choose your favorite data set or one from a Kaggle competition at Kaggle.com to which you can apply computation techniques in `R` described in class. Discuss the challenges in the problem and the data set, and how you circumvented these problems. Consider issues of, for example, sparsity in the features and response, high dimensions, and the scalability issues of BIG data. For any problem, apply any method that you see appropriate and discuss the advantages and disadvantages of each method and why you found them appropriate. Thoroughly explore and assess any inference that you make on the data and what lead to your analysis. In your presentation, explain the data, why it interested you, and your step-by-step analyses that lead to any final conclusions.

4. [**R Shiny Application**] One way to provide a user-friendly environment to apply `R` code and any other coded functions is to create a graphical user interface with `R` Shiny. In this assignment, create a aesthetically pleasing application that performs a task of your choice. Your interface must contain the following components:

   - Data input and output
   - A Function that was written by you (that contains at least 20 lines of code and is properly written and well-documented).
   - Visualization of the data using *ggplot2*
   - Concise summary of Results

   For the presentation, show how your app works and give an example with a data set of your choice. To get started, see `https://shiny.rstudio.com` and `https://shiny.rstudio.com/tutorial/` for a tutorial of how to create a Shiny app.

Due Dates

1. Enter your names, project choices (1, 2, 3, or 4), and your preferred date (not guarenteed) in the following Google Sheet

   Final Project Signup

   by **Tuesday, October 31st** by 9:00 AM.

2. Presentations are to be submitted on Canvas by **Thursday, November 30th** by 9:00 AM.