Data Mining for Social Science (GR5058), Fall 2018

Instructor: Ben Goodrich (benjamin.goodrich@columbia.edu)
Verify that the date below is recent! Syllabus subject to change!

September 2, 2018

Course website: https://courseworks2.columbia.edu/courses/60267
Course Time: Tuesdays 6:10-8:00 in IAB 413.
Office hours: Thursday afternoons. Sign up for a slot on the course website
Teaching Assistants: Rui Lui and Terry Zhang
Optional Computer Session: As needed

Ben Goodrich’s office is in IAB room 270 I (near 270B)

Course Description

The class is roughly divided into two parts:

1. programming best practices, exploratory data analysis (EDA), and unsupervised learning
2. supervised learning including regression and classification methods

In the first part of the course we will focus writing R programs in the context of simulations, data wrangling, and EDA. Unsupervised learning is focused on problems where the outcome variable is not known and the goal of the analysis is to find hidden structure in data such as different market segments from buying patterns or human population structure from genetics data. Supervised learning deals with prediction problems where the outcome variable is known such as predicting a price of a house in a certain neighborhood or an outcome of a congressional race.

Prerequisites

Any QMSS student is presumed to have sufficient background. Any non-QMSS students interested in taking this course should have sufficient background in quantitative methods.

Grading

- 20% homeworks
- 20% in class midterm
- 20% final group project
- 20% final during finals week
- 20% class participation

**Books**

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, 2013, *Introduction to Statistical Learning with Applications in R*, Springer-Verlag. Available as a PDF.

**CampusWire**

CampusWire is a beta version of a tool that is available [https://campuswire.com/p/G8F62DF08](https://campuswire.com/p/G8F62DF08) using code 0010. Rather than emailing questions directly to the professor or TAs, you should post on CampusWire. That way, other students can answer your question, benefit from an answer that the professor or TA provides, ask follow-up questions, etc. There is also Reddit-style upvoting and the statistics collected by CampusWire go into the participation portion of your grade.

If your question pertains to an ongoing homework assignment, your grades, or similar, then you should click on the option to make your post only visible to “Instructors and TAs”. Otherwise, you should post to “Everyone in the class” and avoid direct messaging the instructor and TAs. There is an option to post in Stealth Mode, in which case no one will know it was you that asked the question, but doing so obviously cannot count toward the class participation component of your course grade.

**Outline**

The following outline describes the topics that will be covered along with anticipated associated readings.

**I. Programming Best Practices, Exploratory Data Analysis, and Unsupervised Learning**

**Week 1: Introduction to the Course**

- ISLAR, Chapters 1 and 2. You do not need to read the section of Chapter 1 entitled “Notation and Simple Matrix Algebra” yet.

**Week 2: Introduction to R**

- Grolemund and Wickham, chapters 1, 2, 4, 26, 27, 29, 30
Week 3: Intermediate R
- Grolemund and Wickham, chapters 5, 6, 9, 10, 11, 12, 15

Week 4: Matrix Algebra
- ISLR: Read the section of Chapter 1 entitled “Notation and Simple Matrix Algebra”

Week 5: Exploratory Data Analysis
- Grolemund and Wickham, chapters 3, 7, and 28

Week 6: Unsupervised Learning
- ISLAR, chapter 10

Week 7: Text Analysis
- https://www.tidytextmining.com/ chapters 1 – 6

Week 8: Midterm, in class

II. Supervised Learning

Week 9: Linear Regression
- ISLAR, chapters 3 and 6
- APM, chapters 5 and 6

Week 10: Classification and Logit Models
- ISLAR, chapter 4
- APM, chapters 11 and 12

Week 11: Nonlinear Models
- ISLAR, chapter 7

Week 12: Tree Methods
- ISLAR, chapters 5 and 8
- APM, chapters 8 and 14

Week 13: Neural Networks and Other Stuff
- APM, chapter 7