Introduction to Data Science
PPOL 670-01. Monday 3:30 pm - 6:00 pm.

Instructor: Gaurav Sood
Office: Rm. 402, Old North
Office Hours: Monday 1:00 pm – 3:00 pm
Email: gaurav.sood@georgetown.edu
Website: http://gsood.com/teach/ds/

The class focuses on building skills for collecting, managing, and analyzing large data sets. As part of the course, we will cover two prominent ways of building large original datasets: scraping data from the web, and conducting large n experiments. We will also cover how to manage and analyze data using a popular relational database, SQLite. As part of the discussion around managing and analyzing large datasets, we will also cover cloud based solutions, and basics of mapReduce. In the data analysis segment, we will cover basics of supervised and unsupervised learning, including techniques like cross-validation, and bootstrapping, before learning popular supervised techniques, such as SVM, Ridge Regression, and Elastic Net, and popular unsupervised methods, such as k-means clustering. As part of the course, the students will be expected to complete an independent project in addition to three assignments.

Prerequisites

Some familiarity with programming, and an undergraduate or graduate course in statistics.

Books

An Introduction to Statistical Learning: with Applications in R
By Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
ISBN: 1461471370
Edition: 2014
Required or Optional: Required

Python Programming: An Introduction to Computer Science
By John Zelle
ISBN: 1590282418
Edition: 2nd
Required or Optional: Optional

ggplot2: Elegant Graphics for Data Analysis (Use R!)
By Hadley Wickham
ISBN: 0387981403
Publisher: Springer
Edition: 2010 Required or Optional: Optional

Evaluation

Students will be evaluated on how well they perform on the programming assignments, the final project, and their attendance. The breakdown is as follows:

– Programming Assignment - Scraping (15%)

– Programming Assignment - SQL (15%)

– Programming Assignment - Modeling (25%)

– Final Project (45%)

Assignments will be due Monday at 12 noon. Submit homework by email to me. Anything submitted after the deadline but within 72 hours of the deadline will be docked 30%. Anything beyond 72 hours will get a 0. No excuses barring doctor's note. No extensions. You can work together in groups, but must write code etc. individually. Assignments must be presented in plain text files (no rtf, doc etc.), must be neatly commented, and should have a ReadMe file.

Final project is an original research paper ($\sim$ 2500 words) (or in some cases, a 'data product'). The project must use some of the techniques covered in the class. The data product can be a web service or similar such application. The research paper can be a journal article, or part of a thesis project. The paper must defend the method used, and must include a detailed and careful interpretation of all the numbers and figures. No long literature reviews– anything more than a page– should be part of the paper. Clearly commented code and, where possible, data should accompany all final projects. Students are also expected to give a 10-15 minute in-class presentation. Students will be judged on both the presentation (15% of 45%) and the paper (85% of 45%).

The project has to be done in teams of 2. (If there are odd number of students, one team will be of 3.) You get to choose your partner(s). Alert me when you have found your partner(s) via email, cc'ing your partner(s). Due date for choosing your team is January 26th at 12pm. Each person in the team will get the same grade. Any conflicts with partners must be resolved internally.

Proposals for the final projects are due February 23rd at 12pm. You are welcome to consult me on the project (though not required to) via email before the deadline. Proposals do not need to be more than a paragraph long. They should clearly explain the problem and the proposed methods.

Planned Schedule

01/07: Introduction to "Introduction to Data Science"

- Big Data: New Tricks for Econometrics. By Hal Varian
- Detecting influenza epidemics using search engine query data By Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant.
- The Parable of Google Flu: Traps in Big Data Analysis By David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani.
- Web-scale pharmacovigilance: listening to signals from the crowd. By Ryen W White, Nicholas P Tatonetti, Nigam H Shah, Russ B Altman, and Eric Horvitz.
- 3V Big Data. By Doug Laney.
- What is Data Science? By Mike Loukides.

01/12: R and Python Bootcamp

- R, RStudio, Python, and SQLite should be installed.
- Introduction to R. By Gaurav Sood.
- Chapters 1 and 2 of Python Programming by Zelle.
- Python Tutor. By Philip Guo.

01/26: Getting and Cleaning Data

- Team submissions due.
- Data from the Web into R. By Simon Jackman.
- Scraping the Web for Arts and Humanities. by Chris Hanretty.
- For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights. By Steve Lohr at the NY Times.
- Unix for poets. By Ken Church.

02/02: Working with Social Networking Data

- Scraping Assignment Due.
- Guest lecture by Pablo Barberá

02/09: TBD

02/23: Introduction to Databases

- Likely an extended class.
- SQL for Web Nerds. By Philip Greenspun.
- Codd Turing Award Profile.

03/02: Advanced Databases

- Project proposals due.
- Scalable SQL and NoSQL Data Stores. By Rick Catell.
- New Analysis Practices for Big Data. By Jeffrey Cohen et al.
- A Co-Relational Model of Data for Large Shared Data Banks. By Erik Meijer and Gavin Bierman.

03/16: Statistical Learning

- Database assignment due.
- Chapters 1, 2 and 3 of ISLR.
- A Few Useful Things to Know about Machine Learning. By Pedro Domingos.

03/23: Cross-validation

- Chapter 5 of ISLR.

03/30: Model Selection and Regularization

- Chapter 6 and 9 of ISLR.
- Text Categorization with Support Vector Machines: Learning with Many Relevant Features. By Thorsten Joachims.

04/13: Tree Based Methods

- Modeling assignment due.
- Chapter 8 of ISLR.
- Top Ten Algorithms in Data Mining. By Wu et al.

04/20: Unsupervised Methods

- Chapter 10 of ISLR.
- Follow Your Ideology: A Measure of Ideological Location of Media Sources By Pablo Bareberá and Gaurav Sood.

- Repeated observation of breast tumor subtypes in independent gene expression data sets. By Therese Sorlie et al.

04/27: Class Presentations
Presentation of final projects in an extended 3 hour class.

05/04: Final Papers Due at 12pm.