PPOL 670 - Introduction to Data Science for Public Policy, Spring 2017

Jeff Chen (jc2817@georgetown.edu) and Dan Hammer (dan.hammer@georgetown.edu)

Updated: February 27, 2017

Course Description

The Introduction to Data Science for Public Policy is a survey course of the fundamentals of data science. The course is focused on evaluating and analyzing public policy, telling stories with data to make compelling and fact-based arguments.

The objective of the course is to equip students with the skills to tell stories with data and drive action. Public policy is part of a large and sprawling social system. Parsing causality from a system of variables where everything is related requires a scalpel. This refined approach can be assembled from pre-written code and routines; but it still requires skilled assembly. We will teach an approach that leverages analytical routines that have already been written. The value of this course is in the mortar, not the bricks.

Instructors

Jeff Chen is the Deputy Chief Data Officer of the U.S. Department of Commerce. He has led wide ranging initiatives across 30+ fields, from emergency services to international public health to legal affairs to trade economy. Jeff has previously served as the Director of Analytics at the NYC Fire Department leading development of fire prediction algorithms, senior data roles in the NYC Mayor's Office during the Bloomberg Administration focusing on city operations and health + human services, and an advisor to governments, corporations, and non-profits on applied data for strategy and operations.

Dan Hammer is currently a Senior Policy Advisor at the White House, where he works with the U.S. Chief Technology Officer and U.S. Chief Data Scientist on the public finance of data infrastructure. He was previously the Chief Data Scientist at two environmental non-profits. He cofounded Global Forest Watch, a web application to monitor forests from satellite imagery. He is a Fellow at the Berkeley Institute for Data Science and a PhD candidate in environmental economics at UC Berkeley.

Prior to their current positions, Jeff and Dan worked together as White House Presidential Innovation Fellows at NASA.

Time and location

Classes will be held on Mondays from 6:30pm to 9:00pm in Reiss 283.

- January: 11 (Wednesday), 23 (Monday), 30
- February: 6, 13, 27
- March 13, 20, 27
- April 3, 10, 24
- May 1

Website

Students are expected to sign up for a Github account (https://github.com). Readings and materials will be available from the class Github repository (https://github.com/GeorgetownMcCourt/data-science).

Workload and assignments

Students will be evaluated on the basis of five problem sets (60%) and one final project (40%). Late problem sets will be penalized by 10% per day late. All problem sets will be submitted electronically. The final project assignment details will be made available in late March and the final product will be due on Monday May 8th. As this class is quite hands-on, it is expected that students bring their computers to class to partake in computational activities.

Course Outline

Data science is dependent on sound application of computer programming, mathematics/statistics, and communication. This course is thus organized into three units that dive into the fundamentals. Particular emphasis is placed on skilled assembly of empirical ideas, drawing from standard and non-standard data. The section outlines are subject to change up to one week before the lecture. Please continue to review the syllabus throughout the course.

Section 1: Fundamentals

Data science is about designing and building data products that derive insight. This first section will focus on developing fundamental skills required to build effective products.

Lecture 1: Preliminaries

The objective of the first lecture is to overcome the coefficient of static friction in using R for data science. Students will learn to execute simple R scripts to read, write, and extract data elements.

Lecture objectives

1. Data science: What is it? What is the lay of the land?
2. Languages of data science
3. Basics of R progrmming
4. Read data from CSV and JSON
5. Data types and classes, including matrix, data.frame, list, and vectors
6. Extracting rows, columns, and specific elements from a data frame
7. Basic operations (e.g., sum, mean) on rows; useful as consistency checks.
8. Write data to CSV and JSON
9. Getting started with Github

Example application

- Graphing photovoltaic energy data from the National Institute of Standard and Technology's Net Zero Energy Residential Test Facility

Lecture 2: Data manipulation

The objective of this lecture is to present the most important and fundamental elements of data manipulation. These core operations include sort, merge, reshape, and collapse. We will also present loops through multiple rows or columns, and other alternatives to operate on partitions of data frames.

Lecture objectives

1. Sound data manipulation as the basis of good data science
2. Sort data based on column values
3. Subset data frames
4. Reshape data table, wide <–> long
5. Merge data frames

6. Collapse data frames
7. Text processing: capitalization, substring, regex
8. Looping through basic operations (bonus: same idea without loops)

Example application - Parsing and conducting basic text analysis using State of the Union Speeches (2009 to 2016)

Lecture 3: Functions and Control Structures

Building upon basic data manipulation and high level analytical tasks, this session will focus on programming paradigms that are commonly relied upon when practicing data science.

Lecture objectives

1. Custom functions for consistency and efficiency
2. Control structures: Loops, if statements
3. Suitable practices

Example application

- Collaborative filtering with Last.FM music listening data

Homework Assignment

- Batch Extraction of Census Housing Permits data

Lecture 4: Exploratory Data Analysis (EDA)

The objective of this lecture is to handle missing values appropriately and script visual checks to find errors introduced in data input/output. We will also start to view computational optimization techniques, like taking advantage of multiple cores for heavy duty operations (parallel processing).

Lecture objectives

1. Understanding data structures
2. Statistical measures
3. Graph and visual analytics

Example application

- Finding health coverage patterns using the US Census American Community Survey
- Conducting analysis of missing values analysis of weather anomalies from 1880 to Present using the National Oceanographic and Atmospheric Administration's GHCN-M

Section 2: Data Analysis + Modeling

The use case drives the technique. In public policy, data can be used to support evaluation of programs to understand causal mechanisms (e.g. retrospective focus) or enable the creation of data-rooted products that drive action (e.g. deployed applications). Machine learning and data analysis enables both uses of data and will be the focus of the next five courses.

Lecture 5: Ordinary Least Squares, Simulation and Bias

Formal statistics offers methods to calculate closed-form, analytical answers to the limits of OLS regression. Data science offers a more immediate and arguably a more accessible solution: simulate conditions and examine the outcomes. We begin to use the early visualizations techniques taught in a previous lecture for analysis.

Lecture objectives

1. Simulating OLS and identifying p-values
2. For-loops versus `apply` for simulations
3. Visualizing distributions with ggplot

Example application

- Schooling outcomes

Lecture 6: Introduction to Supervised Learning

Supervised learning is the most relied upon class of techniques that enable causal inference but also deployed precision policy. How does changing one variable independently impact another variable? We begin to introduce basic regression analysis, correlation coefficients, ordinary least squares, and the relationship between the concepts. Note that this is a very cursory review, and the deep assumptions are not tested or expounded upon.

Lecture objectives

1. What is supervised learning?
2. Structure of a supervised learning project
3. Target variables, Input variables, Objective function and evaluation measures, model experiment design, Cross validation versus train/validate/test, Regression versus classifiers
4. Ordinary Least Squares (OLS)
5. K-Nearest Neighbors (kNN)

Example application

- Prediction of missing values in satellite imagery using kNN

Homework Assignment

- Lec 6: Satellite imagery for predicting employment

Lecture 7 + 8: Classification techniques

Classification models are one of the workhorses of data science. Classifiers enables data-driven applications such as risk scoring, lawsuit outcome prediction, marketing lead generation, facial detection and computer vision, spam filtering, among other use cases. This session will focus on the fundamentals of classification models, types of models, and daily applications.

Lecture objectives

1. Three common problems using classifiers
2. Structure of a classification project, Target variables, Input variables, Objective function and evaluation measures, model experiment design, Cross validation versus train/validate/test, Confusion matrix, TPR, TNR, AUC
3. Framing dataset
4. Models: statistical assumptions and mechanics, risks/strengths, implementation, non-technical explanation, Decision trees, Logistic Regression, K-Nearest Neighbors
5. Appropriate uses of classification techniques, Scoring, prediction and prioritization, Propensity score matching

Example application

- Healthcare insurance coverage data

Homework Assignments

- Lec 7: Predict activity using smartphone accelerometer data (due Lec 8).
- Lec 8: Hand out class project instructions, one page proposal of what you'll do due by Lec 9.

Lecture 9: Unsupervised learning

No, this is not an independent study session. Unsupervised learning techniques such as clustering and principal components analysis help to identify recognizable patterns when no labels are provided. In sales and recruitment offices, customer segmentation may use current customer data, then use clustering techniques to identify k-number of distinct customer profiles. In resourceful law firms, data scientists may develop topic modeling algorithms to automatically tag and cluster hundreds of thousands of documents for improved search. This session will focus on clustering methodologies that are commonly employed in applied research.

Lecture objectives

1. Three common problems using unsupervised learning
2. Structure of unsupervised learning project, Input variables, optimization methods
3. Framing dataset
4. Models: statistical assumptions and mechanics, risks/strengths, implementation, sanity checks, non-technical explanation, K-means clustering (K-means), Principal Components Analysis (PCA)/Dimensionality Reduction, Hierarchical clustering (if time permits)
5. Appropriate uses of k-means and PCA

Example application

- Univariate clustering application: k-means
- Multivariate clustering application: Customer segmentation using Census American Community Survey

Homework Assignment - Lec 9: Write prototypical functions that will help you do your project. Due Lec 10.

Section 3: Data enhancement and visualization

Beyond the data preparation and modeling, the 'presentation layer' is the glue that will allow a data science project to stick with target audiences. Often times, presentation is graphical and relies upon a rich ecosystem of visualization, web services, and interactive applications to communicate pertinent issues.

Lecture 10: Data storytelling through graphical representation

Often times, the model is not enough to communicate the value of the data analysis. A well-designed visualization can illustrate patterns and allow target audiences to establish a connection with the analytical effort at hand.

Lecture objectives

1. Three examples of the presentation layer
2. Static visualizations: ggplot2
3. Interactive visualizations: dygraphs, plotly, networkd3, threejs

Example application

- Phone gyroscopic data: time series
- Developing a network map of global trade flows

Lecture 11: Web service APIs and spatial data

There are many cases where you will rely on data or services that aren't stored or built on your local machine, but rather are exposed as web service application programming interfaces (APIs). These are the components of modern software development, and we will teach how to find and utilize these services from within the R programming environment. We use this lecture as an opportunity to introduce spatial data within R by interacting with an API for geographic data.

Lecture objectives

1. Three examples of APIs and why they matter
2. Interacting with web service APIs from R
3. Writing a client-side function to simplify the remote interaction
4. Batch request for elevation data from the Google Elevation API
5. Viewing the spatial data in R

Example application

- Extracting elevation data from the Google Elevation API
- Identifying the characteristics of farmers' markets in the Southwest United States.

Lecture 12: Spatial Data and Maps

The state of data is rapidly expanding in two principal directions: transactional-level and spatially. Maps are the principal mode of representing spatial data, which relies upon different types of GIS formats (e.g. shapefiles, raster, GeoJSON) and presentation medium. This lecture dives into spatial considerations in data science.

Lecture objectives

1. Three examples of spatial data
2. A framework for approaching GIS
3. Preliminaries of GIS data, File types (shp, .nc, .tif, .json), Projections, Feature type (point, line, polygons, grids), Visualizations (choropleth, dot density, proportional)
4. Mapping choropleth maps using polygon shapefiles
5. Geoprocessing of points to polygon
6. Displaying multiple layers

Example application

- Chicago crime data

Lecture 13: Export results and interactivity

Often times, users like to interact with a product as opposed to reading curated results. Enter interactivity – a well-proven mode of displaying results.

Lecture objectives

1. Three examples of interactive data science
2. An Intro to R Shiny
3. Building a simple interactive tool

Example application

- TBD

Academic Resource Center/Disability Support

If you believe you have a disability, then you should contact the Academic Resource Center (arc@georgetown.edu) for further information. The Center is located in the Leavey Center, Suite 335 (202-687-8354). The Academic Resource Center is the campus office responsible for reviewing documentation provided by students with disabilities and for determining reasonable accommodations in accordance with the Americans with Disabilities Act (ASA) and University policies. For more information, go to http://academicsupport.georgetown.edu/disability/.

Important Academic Policies and Academic Integrity

McCourt School students are expected to uphold the academic policies set forth by Georgetown University and the Graduate School of Arts and Sciences. Students should therefore familiarize themselves with all the rules, regulations, and procedures relevant to their pursuit of a Graduate School degree. The policies are located at: http://grad.georgetown.edu/academics/policies/

Provosts Policy Accommodating Students Religious Observances

Georgetown University promotes respect for all religions. Any student who is unable to attend classes or to participate in any examination, presentation, or assignment on a given day because of the observance of a major religious holiday (see below) or related travel shall be excused and provided with the opportunity to make up, without unreasonable burden, any work that has been missed for this reason and shall not in any other way be penalized for the absence or rescheduled work. Students will remain responsible for all assigned work. Students should notify professors in writing at the beginning of the semester of religious observances that conflict with their classes. The Office of the Provost, in consultation with Campus Ministry and the Registrar, will publish, before classes begin for a given term, a list of major religious holidays likely to affect Georgetown students. The Provost and the Main Campus Executive Faculty encourage faculty to accommodate students whose bona fide religious observances in other ways impede normal participation in a course. Students who cannot be accommodated should discuss the matter with an advising dean.

Statement on Sexual Misconduct

Please know that as a faculty member I am committed to supporting survivors of sexual misconduct, including relationship violence, sexual harassment and sexual assault. However, university policy also requires me to report any disclosures about sexual misconduct to the Title IX Coordinator, whose role is to coordinate the University's response to sexual misconduct.

Georgetown has a number of fully confidential professional resources who can provide support and assistance to survivors of sexual assault and other forms of sexual misconduct. These resources include:

Jen Schweer, MA, LPC
Associate Director
Health Education Services for Sexual Assault Response and Prevention
(202) 687-0323
jls242@georgetown.edu

Erica Shirley
Trauma Specialist
Counseling and Psychiatric Services (CAPS)
(202) 687-6985
els54@georgetown.edu

More information about campus resources and reporting sexual misconduct can be found at http://sexualassault.georgetown.edu.