Data Mining and Data Visualization for the Social Sciences

MACS 24000/34000

Benjamin Soltoff Philip Waggoner

2020-08-04

Syllabus

Contact information

	Dr. Benjamin Soltoff	Dr. Philip Waggoner
Email	soltoffbc@uchicago.edu	pdwaggoner@uchicago.edu
GitHub	bensoltoff	pdwaggoner
Office hours	TBD	TBD

Course description

This course introduces students to techniques for extracting and communicating knowledge from data. In the first half, students study visualizations as a method for summarizing information and reporting analysis and conclusions in a compelling format. This introduces the ideas and methods of data visualization, with emphasis on both why you are doing something as well as how to produce optimal visualizations. In the second half, students are introduced to the rapidly developing world of data mining. Focus will be on knowledge discovery and pattern recognition in the context of social science problem solving. From partitioning and anomaly detection to text clustering, high-dimensional mining, and deep learning, students will be given a thorough introduction to prominent techniques for exploring and discovering patterns in data. Throughout the course, class sessions will combine lecture, coding challenges, and computational problem solving to encourage wide engagement with the techniques using the R programming language.

Prerequisites

MACS 20500, CS 10121, or a similar introductory programming course. Experience in R is required. STAT 23400 or similar introductory statistics course is expected. Experience with machine learning is helpful but not required.

Course schedule

Week 1 (Data visualization with Dr. Soltoff)

Date	Topic
27-Jul	Introduction to data visualization
28-Jul	Showing the right numbers
29-Jul	Making plots pretty and clean
30-Jul	Geospatial visualizations
31-Jul	Interactive Shiny applications

Week 2 (Data mining with Dr. Waggoner)

Date	Topic
3-Aug	Foundations of Data Mining
4-Aug	Patterns & Associations
5-Aug	Unsupervised Machine Learning
6-Aug	Mining Labeled Data & Text
7-Aug	Deep Learning

What do I need for this course?

Class sessions are a mix of lecture, demonstration, and live coding. It is essential to have a computer so you can follow along and complete the exercises. Before the course starts, you should install the following software on your computer:

- R - easiest approach is to select a pre-compiled binary appropriate for your operating system.
- RStudio IDE - this is a powerful user interface for programming in R. You could use base R, but you would regret it.
- Git - Git is a version control system which is used to manage projects and track changes in computer files. Once installed, it can be integrated into RStudio to manage your course assignments and other projects.

Comprehensive instructions for downloading and setting up this software can be found here.

All readings (e.g., papers, book chapters) will be open source, with either links or citations provided.

How will I be evaluated?

Students will submit daily problem sets each worth 100 points. Each assignment is due prior to the start of class (10 am CDT) the following day. There is no final exam or project in the course. Assignments will be submitted via GitHub Classroom.

Statement on Disabilities

The University of Chicago is committed to diversity and rigorous inquiry from multiple perspectives. The MAPSS, CIR, and Computation programs share this commitment and seek to foster productive learning environments based upon inclusion, open communication, and mutual respect for a diverse range of identities, experiences, and positions.

This course is open to all students who meet the academic requirements for participation. Any student who has a documented need for accommodation should contact Student Disability Services (773-702-6000 or disabilities@uchicago.edu) and provide us (Dr. Soltoff and Dr. Waggoner) with a copy of your Accommodation Determination Letter as soon as possible.