STOR 390: Introduction to Data Science

Syllabus, Spring 2017

Class

- TuTh 5:00-6:15
- Greenlaw 101

Instructor: Iain Carmichael

- iain@unc.edu
- Hanes B30
- Office hours: TBA

Grader: Brendan Brown

- bb@live.unc.edu
- Hanes B26
- Office hours: TBA

Graduate Research Consultant: Varun Goel

- varung@live.unc.edu
- Office: TBA
- Office hours: TBA

All course material can be found on the course website (https://idc9.github.io/stor390/) and github repository (https://github.com/idc9/stor390).

Course description

This course is an application-driven introduction to data science. Statistical and computational tools are valued throughout the modern workplace from Silicon Valley startups, to marine biology labs, to Wall Street firms. These tools require technical skills such as programming and statistics. They also require professional skills such as communication, teamwork, problem solving, and critical thinking.

You will learn these tools and hone these skills through hands-on experience working with datasets such as: Museum of Modern Art records, TCGA Gene Expressions and the text script of Beauty and the Beast. The first half of the semester will cover R programming skills. The second half will cover a number of topics: exploratory data analysis, web scraping, text processing, and effective visualization through a series of modules.

Textbook

The primary reference for the course is R for Data Science by Hadley Wickham (free online).

We will use a number of other references (all free)

- Data Carpentry R for data analysis for Ecology
- R Programming for Data Science by Roger Peng
- Introduction to Statistical Learning with Applications in R by Tibshirani et al.
- Advanced R by Hadley Wickham
- Tidy Text Mining with R by Julia Silge and David Robinson

I will provide additional resources in the lecture nodes.

Goals and Learning Objectives

- learn core R programming skills (sections 1 - 21 of R for Data Science)
- practice statistical and programming best practices e.g. comment code, identify possible sources of biases
- make use of literate programming with R Markdown
- develop professional skills: clear communication, creative thinking, critical thinking, self-directed learning, and effective teamwork

Topics

Many of the topics will follow R for Data Science (RDS) fairly closely. These topics are subject to change

- Visualization with ggplot2 (RDS ch3)
- Data manipulation dplyr (RDS ch5)
- R Markdown (RDS ch 29-30)
- Programming e.g. functions, loops, if/else, comments (RDS ch 4, 6, 17, 19, 20)
- Tidy data, relational data and data import (RDS ch 11-13)
- Strings and regular expressions (RDS ch 14)
- Exploratory Data Analysis
- Modeling: classification, clustering and regression
- Web scraping
- Text data and Natural Language Processing
- Additional programming topics (if there is time)
  - reproducibility
  - interactive graphics with shiny
  - effective visualization for communication
  - date/time data
  - github
  - GIS data
  - data privacy/ethics
- Three rules of effective communication (from Trees, Maps and Theorems)

Prerequisites

Some coding or some statistics (STOR 155 or COMP 110).

We will use concepts from statistics, machine learning and computer science that rely on more advanced math, however, we will cover these topics from a heuristic perspective. STOR 390 is meant to augment many fanatic existing courses such as: probability, theoretical/applied statistics, algorithms, databases etc. Following R for Data Science

> We believe it's important to stay ruthlessly focused on the essentials so you can get up and running as quickly as possible.

Technology

We will use the R programming language and R Studio. Both of these are free. Students must have a laptop and bring it to class. If this is a problem please see the instructor.

Grading

- Labs: 35%
  - start these in class they are due the following class period
- Longer assignments: 35%
  - 4 data analyses
- Class participation: 15%
- Final project: 15%
- Extra credit: up to 5%

The **longer assignments** will be end to end data analyses i.e. get, clean, analyze, communicate. Code will be provided to make these manageable.

These assignments are graded on a 100 point scale and will take into account: statistical rigor, coding accuracy, and communication. Note that communication includes: writing human readable code, clear visualizations and coherent writing. Detailed rubrics will be provided with each assignment.

Some of these assignments will be done in teams (see below for group work policy). One homework may be dropped *if you participate in DataFest.* Assignments handed in late will lose 10 points each day past the due date.

Labs are more frequent and shorter than assignments. They are graded on a 0-3 scale

- 0: not handed in
- 1: partially complete
- 2: complete with mistakes
- 3: complete (possibly minor mistakes)

Class activities will be graded on a 0-1 point scale (i.e. incomplete/complete). One missing class activity will be dropped. If you miss more than one you may make them up by doing an additional assignment of the instructor's choosing with a maximum of 2 make ups. Warning: the more classes you miss the more challenging these assignments will become.

The **final project** is to do a novel data analysis and write a blog post about it. This project will be done in groups. You are more than welcome to reach out to professional researchers and companies to collaborate for this project. See the final project page for more details.

There will be several opportunities for **extra credit** (worth one point each). See the extra credit page for details

If you believe a homework was graded incorrectly you may appeal to the grader – however we reserve the right to regrade the entire assignment (meaning your score may end up lower).

Final grades

I will curve the total grade, but not grades from individual assignments. If your final grade is about 90 you are assured an A- or A. Similarly, if your final grad is about an 80 you are assured a B- or B. I reserve the right to curve grade using more generous cutoffs depending on the overall performance of the class. Once the final grade has been assigned it cannot be changed unless there was a numerical error in computing the final grade.

Group work

The teamwork policy comes from Teaching and Learning STEM and applies to homework and the final project (not class activities).

- the instructor will assign teams
- final grade will be adjusted by peer ratings
- as a last resort a team may fire an uncooperative member

More details will be provided before the first team project.

Getting help

Programming can be incredibly frustrating and take some time to get used to. Before you email the TA please spend some time trying to solve/Google the problem. Once you have exhausted your resources/patience ask some (rule of thumb: spend at least 5 minutes, don't spend more than 20 minutes stuck on one problem).

There is a large number of free R/data science resources online. I've listed a few at the top of this page and in the references page.

FAQ/tips for success

- Code a little bit every day. Even if you can only find 20 minutes – write a little bit of code.
- Google is your best friend for solving programming problems. I will repeat this many times.
- If one of the readings isn't satisfying to you go find a different explanation! It often takes 3 different explanations (read multiple times) before I understand an unfamiliar, tricky concept.
- If you are new to programming the beginning of the course will be a big adjustment. Don't give up, you'll get the hang of it in a couple weeks.
- Coding in general can be very frustrating. Be patient and keep hacking at the problem.
- If you are already familiar with R there is still a lot to learn. For example, if you are not used to the tidyverse the beginning of course will take some adjustment.
- If you have already machine learning/more advanced statistics courses there will be opportunities for you to use what you have learned/learn new things.
- This class may be quite different from other STEM courses you have taken – particularly the amount of class participation. Active Learning is backed up by hundreds of studies that show active learning performs better than traditional lecturing on almost every examined learning outcome (see Felder and Brent below).
- If there is something related to data science you want to learn tell me. I can at least point you to resources and may try to include it in the class!

Honor Code

All students must be familiar with and abide by the Honor Code, which covers issues such as plagiarism, falsification, unauthorized assistance, cheating, and other grievous acts of academic dishonesty. Violations of the Honor Code will not be taken lightly.

For labs/group activities you are encouraged to work with other people, however, you are **not allowed to copy someone else's code**. For homework you are encouraged to use online resources as long as you **cite a resource you borrow code from**.

For informal collaborations, such as labs and activities, you may help each other debug programs and develop approaches to solve a problem, but you should not directly share code. You can post questions about specific

parts of a problem on Sakai if you have made several attempts to solve it. Please do not post the entire solution to a problem – just answer the specific question being asked.

Please report any significant collaborations (just post names in a comment at the top of a script). We may use software to detect cheating and these reporting collaborations will help prevent false positives. You are expected to be able to explain your code by your self in class. These policies are based on COMP 401 taught by Ketan Mayer-Patel.

Students with disabilities

UNC facilitates the implementation of reasonable accommodations, including resources and services, for students with disabilities, chronic medical conditions, a temporary disability or pregnancy complications resulting in difficulties with accessing learning opportunities.

All accommodations are coordinated through the Accessibility Resources and Service Office, Tel: 919-962-8300 or Email: accessibility@unc.edu. Detailed information about the registration process is available at accessibility.unc.edu/.

References

The following books, courses and papers were influential in the design of this course. See the acknowledgements page for a long list of fantastic people who provided lots of input for the class.

- https://idc9.github.io/stor390/course_info/acknowledgments.html
- Teaching and Learning STEM: a Practical Guide by Richard Felder and Rebecca Brent
- Trees, Maps and Theorems by Jean-Luc Doumont
- Data Science Specialization from Johns Hopkins via Coursera
- CS109: Data Science taught by Joe Blitzstein at Harvard
- Data wrangling, exploration and analysis with R taught by Jenny Bryan at UBC
- Introduction to Data Analysis taught by Hadley Wickham at Rice
- A Guide to Teaching Data Science by Stephanie Hicks and Rafael Irizarry
- Data Science in Statistics Curricula: Preparing Students to “Think with Data” by Hardin et al.