Data Science

About the Course

Instructor

- Ben Baumer (bbaumer@smith.edu, Burton 308, 413-585-3440)
  Office hours: Mondays 2:30 - 4:00, Fridays 10:00 - 11:30, and by appointment

Description Computational data analysis is an essential part of modern statistics. This course provides a practical foundation for students to compute with data, by participating in the entire data analysis cycle (from forming a statistical question, data acquisition, cleaning, transforming, modeling and interpretation). This course will introduce students to tools for data management, storage and manipulation that are common in data science and will apply those tools to real scenarios. Students will undertake practical analyses using real, large, messy datasets using modern computing tools (e.g. R, SQL) and learn to think statistically in approaching all of these aspects of data analysis.

Prerequisites CSC111 or MTH205/CSC205 plus an introductory statistics course (e.g. MTH245, ECO220 or AP Statistics), CSC107 recommended, but not required. Some programming experience is required.

Textbooks

- Required:
  1. *The Visual Display of Quantitative Information*, Edward Tufte, Graphics Press, 2001. [$27-40]
  2. *Introduction to Data Technologies*, Paul Murrell. Chapman Hall, 2013. Available as PDF: `https://www.stat.auckland.ac.nz/~paul/ItDT/itdt-2013-03-26.pdf`
  3. *Introduction to Data Mining*, Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Pearson Addison-Wesley, 1st edition, 2006. Chapters 4, 6, 8 available as PDF: `http://www-users.cs.umn.edu/~kumar/dmbook/index.php`
  4. *The elements of statistical learning*, Trevor Hastie, Robert Tibshirani, and J Jerome H Friedman. Springer New York, 2nd edition, 2009. Available as a PDF: `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`
  5. *Mining of massive datasets*, Anand Rajaraman and Jeffrey David Ullman. Cambridge University Press, 2011. Available as a PDF: `http://infolab.stanford.edu/~ullman/mmds.html`
- Supplementary:
  1. *Introduction to Data Mining*, Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Pearson Addison-Wesley, 1st edition, 2006. [$110]
  2. *An Introduction to Data Science*, Jeffrey Stanton. Free, available as an eBook: `http://jsresearch.net/groups/teachdatascience/`
  3. *Visual Explanations*, or any other book by Edward Tufte
  4. *Visualize This!* or *Data Points*, by Nathan Yau

Classes Class meets Tuesday and Thursday from 9:00-10:20 am in Sabin-Reed 301. I expect you to attend class. Your participation is an important part of the learning process. If you cannot attend a particular class I would appreciate the courtesy of advanced notice and an explanation for your absence. Class participation and attendance contribute 5% to your final grade.

I hope it goes without saying that during class, you should not use your computer or cell phone for personal email, web browsing, Facebook, or any activity that's not related to the class.

Policies

Attendance Your attendance in class is crucial, as is your punctuality. We are all going to learn this material together, so we need to have everyone present and working. I will make accommodations for an unavoidable absence if you notify me. Our Honor Code means that you will be the judge of whether or not an absence was unavoidable. (For instance, staying in bed because you had the flu would be an unavoidable absence, but oversleeping because you stayed up late to write a paper would be an avoidable absence.) One necessary absence during the semester is not unusual; having more than two is uncommon.

Collaboration Much of this course will operate on a collaborative basis, and you are expected encouraged and to work together with a partner or in small groups to study, complete homework assignments, and prepare for exams. However, every word that you write must be your own. Copying and pasting sentences, paragraphs, or blocks of R code from another student is not acceptable and will receive no credit. All students are bound by the Smith College Honor Code.

Academic Honor Code Statement

> Smith College expects all students to be honest and committed to the principles of academic and intellectual integrity in their preparation and submission of course work and examinations.
>
> Students and faculty at Smith are part of an academic community defined by its commitment to scholarship, which depends on scrupulous and attentive acknowledgement of all sources of information, and honest and respectful use of college resources.

Assignments

Homework [25%] There will be several problem sets over the course of the semester. Problem sets will involve computational assignments in R with written explanations. In order to streamline your workflow and my ability to comprehend your work, you should complete all of your homework assignments in R Markdown. All homework will be submitted electronically by 2 am ET of the due date. Late assignments will lose points at the rate of 20% per day.

Project [25%] You will work on a term project in a group of three over the course of the semester. This is an opportunity for you to exercise your creativity and create something meaningful. This project will be wildly open-ended, and its evaluation will emphasize originality and ingenuity in addition to sophistication and complexity. More details about the project will follow.

Exam 1 [20%] The first exam will be a traditional, in-class, closed book written exam. You may bring a calculator and one piece of paper of hand-written notes (double-sided). Smith College has had an academic honor code since 1944, and all students, staff and faculty are bound by this code. Cases of dishonesty, plagiarism, etc., will be reported to the Academic Honor Board.

Exam 2 [25%] For the second exam, you will compete in an internal data mining competition. The competition will have several stages. I will present you with a large data set that I have already split (uniformly at random) into three pieces. I will give you the first piece (the *training* set, which will likely comprise about 60% of the data). During the first stage, you will work to analyze the training data and build a model that will make predictions about the value of the response variable. You will be free to use whatever techniques you deem appropriate. Once you have submitted your model, I will test it on the second piece of the data (the *testing* set, which will likely comprise about 20% of the data). During the second stage, your original team will combine with another team to form a superteam. You will then work together to improve your model, which will then be tested on the remaining piece of the data set (the *validation* set). Your grade will be based on how well

your models perform, and how well you explain the logic behind the choices you have made. More details will follow as we get closer.

Participation [5%] There will be many opportunities to participate in class during the semester.

Extra Credit [?%] Extra credit is available in several ways: attending an out-of-class lecture (as will be announced) and writing a short review of it; pointing out a substantial mistake in the book or a homework exercise not already present in the errata; drawing my attention to an interesting data set, data science project, or news article; etc. The extra credit is applied when a student is near the boundary of a letter grade.

Grading When grading your written work, I am looking for solutions that are technically correct and reasoning that is clearly explained. Numerically correct answers alone are not sufficient on homework, tests or quizzes. Neatness and organization are valued, with brief, clear answers that explain your thinking. If I cannot read or follow your work, I cannot give you full credit for it.

Resources

Moodle The Moodle site will be regularly updated with homework, project information, assignments, and other course resources. Please check it regularly.

Computing The use of the R statistical environment with the RStudio interface (downloadable from `rstudio.org`) is thoroughly integrated into the course. You have two options for using RStudio:

1. The *server* version of RStudio on the web at (`http://rstudio.smith.edu:8787`). The advantage of using the server version is that all of your work will be stored in the cloud, where it is automatically saved and backed up. This means that you can access your work from any computer with a web browser (Firefox is recommended) and an Internet connection. The downside is that there are only 2 CPUs and 8 GB of RAM allocated to this machine, and you have to share those resources with each other and with all of the MTH220 students!

2. A *local* version of RStudio installed on your machine. This is recommended due to the computational resources this course demands and the aforementioned resource allocation problem. Your laptop likely has at least 2 CPUs and at least 4 GB of RAM. The downside to this approach is that your work is only stored locally, but I get around this problem by keeping all of my work in a Dropbox folder.

Note that you do not have to choose one or the other – you may use both. However, it is important that you understand the distinction so that you can keep track of your work. Both R and RStudio are free and open-source, and are installed on most computer labs on campus. Several reference books for R are on reserve in the Young Science Library in Bass Hall.

Unless otherwise noted, you should assume that it will be helpful to bring a laptop to class. It is not *required*, but since there are only three workstations in the classroom, we will need a critical mass (i.e. at least 12) computers in the classroom pretty much everyday.

Writing Your ability to communicate results, which may be technical in nature, to your audience, which is likely to be non-technical, is critical to your success as a data scientist. The assignments in this class will place an emphasis on the clarity of your presentation and writing.

Extra Help The Spinelli Center for Quantitative Learning (2nd Level of Neilson Library) supports students doing quantitative work across the curriculum, and has a Statistics Counselor available for appointments. Your fellow students are also an excellent source for explanations, tips, etc. – especially since some of you have more programming experience and deeper prior exposure to R.

Schedule

Tentative Schedule The following outline lists each class date and gives the topic that will be discussed in that class. A reading assignment is also given for each class date. Please complete the reading assignment *before* coming to class so that you can participate fully in the discussion. I reserve the right to revise this schedule – updates will be posted on Moodle.

N	Date	Day	Reading	Topic	Due
1	Sep-3	T		Introduction to Data Science	
2	Sep-5	Th	VDQI, Ch. 1-3	Data Graphics & Presentation	
3	Sep-10	T	VDQI, Ch. 4-6	Elements of Visualization	HW 1
4	Sep-12	Th	VDQI, Ch. 7-9	Theory of Visualization	
5	Sep-17	T	IDT, Ch. 9.3-9.6	Data Management in R	HW 2
6	Sep-19	Th	IDT, Ch. 9.8.7, 9.11.3	Functions and Vectorized Operations	
7	Sep-24	T	IDT, Ch. 7.1-7.2	Beyond the Flat File: SQL	HW 3
8	Sep-26	Th	IDT, Ch. 8.1-8.2	More with SQL	
9	Oct-1	T	IDT, Ch. 9.8.8	Merging and Aggregating	HW 4
10	Oct-3	Th		**Exam 1**	
11	Oct-8	T		Spatial Analysis & Mapping	**Project Groups**
12	Oct-10	Th	IDT, Ch. 9.9, 11	Text Mining & Regular Expressions	HW 5
-	Oct-15	T		**Fall Break**	
13	Oct-17	Th		Smith Data Expo	
14	Oct-22	T	OI, Ch. 4	Statistical Inference Revisited	**Initial Proposal**
15	Oct-24	Th		Simulation	HW 6
16	Oct-29	T	OI, Ch. 7-8, ESL, Ch 3.1-3.2	Regression Revisited	
17	Oct-31	Th	ESL, Ch. 3.3	Variable Selection	**Final Proposal**
18	Nov-5	T	MMD, Ch. 1.1-1.2; ESL, Ch. 1-2.3	Data Mining & Cross-Validation	HW 7
19	Nov-7	Th	IDM, Ch. 4	Classification	
20	Nov-12	T		More Classification	
21	Nov-14	Th	IDM, Ch. 6	Clustering	HW 8
22	Nov-19	T		**Exam 2**	
23	Nov-21	Th	MMD, Ch. 10.1-10.2, 10.7.1-10.7.3	Network Science	**Project Update**
24	Nov-26	T	MMD, Ch. 5.1, 9.5	Network Science in the Wild	HW 9
-	Nov-28	Th		**Thanksgiving Break**	
25	Dec-3	T	MMD, Ch. 2.1-2.3	How big is Big Data?	HW 10
26	Dec-5	Th		**Project Presentations**	**Presentation**
27	Dec-10	T		**Project Presentations**	**Presentation**
-	Dec-16	T			**Project Write-up**

Table 1: Last Modified: September 2, 2013