CS 207 - Data Science and Visualization
Spring 2016

Professor: Sorelle Friedler
sorelle@cs.haverford.edu

An introduction to techniques for the automated and human-assisted analysis of data sets. These "big data" techniques are applied to data sets from multiple disciplines and include cluster, network, and other analytical methods paired with appropriate visualizations.
Course cap: 24 students.
Includes a required lab section.

Textbooks

Selections from the following textbooks will be used. They are all available either for free or in the science library.

- *Mining of Massive Datasets* by Anand Rajaraman and Jeffrey D. Ullman. Available free at: `http://infolab.stanford.edu/~ullman/mmds/booka.pdf`. We'll call this book "Data Mining."
- *Visualization Design and Analysis: Abstractions, Principles, and Methods* by Tamara Munzner. We'll call this book "Visualization."
- *The Elements of Statistical Learning* by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Available free at: `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`. We'll call this book "Statistical Learning."
- *Interactive Data Visualization for the Web* by Scott Murray. Available free at: `http://chimera.labs.oreilly.com/books/1230000000345/index.html`. We'll call this book "D3."

Prerequisites

All three of the following categories satisfied, or permission of the instructor.

- CS 105 with a grade of 2.0 or better
- CS 106 with a grade of 2.0 or better
- CS 231 with a grade of 2.0 or better

Topics

There will be approximately one lab per main topic listed below, on which students will apply the analysis techniques from that section to a given data set.

1. Statistical Background (2 weeks)
   - probability distributions
   - random numbers
   - regression analysis
2. Visualization Background (1 week)
   - data abstractions

- visual encoding principles

3. Cluster Analysis (5 weeks)
   - nearest neighbors
   - hierarchical clustering
   - centroid-based clustering (k-means and variants)
   - visualization - hierarchical visualizations, heat maps, matrix views
4. Network Analysis (3 weeks)
   - graph theory basics (matrix representation, etc.)
   - centrality
   - PageRank
   - visualization - layout options, force-directed placement, color
5. Supervised Learning (2 weeks)
   - Model evaluation
   - Decision trees
   - Overfitting
   - Ensembles and boosting
   - Naive Bayes

Labs

0. HTML / CSS basics. Make a personal website using bootstrap. You may pass out of this part of the lab by sending a link to a previously created website. Set up d3.
1. Data cleaning.
2. d3 visualization basics via linear regression graphing.
3. Nearest neighbor searching.
4. Hierarchical clustering.
5. k-means clustering and finding k.
6. PageRank.

Schedule - week by week

This schedule is *tentative*. Students should expect ***at least* 10 hours of work each week.** For the most up-to-date dates and deadlines see the course Google Calendar. Reading should be done during or before the week in which it is listed. The topics of the week will be based on, but not exclusive to, the reading.

1. Introduction to data science and visualization. Introduction to specific data sets (with visits by some professors).
   Reading: Chapter 1 in Statistical Learning, Chapters 1 - 4 in D3, Chapter 1 in Visualization, Chapter 1 in Data Mining
   Lab work (lab 0): Set up course repository. Make a personal website. Set up d3.

2. Probability basics, Gaussians, Linear Regression.
   Reading: Chapter sections 2.1, 2.2, and 2.3.1 in Statistical Learning (linear regression), Chapters 3 - 6 in D3
   Lab 0 Due: HTML / CSS basics with bootstrap.
   Due Thursday: an email to Sorelle listing (in order) your top three data set preferences. If you are a scientific computing concentrator, note this in your email.
   Lab work: Unix basics, scripting basics.

3. Nearest neighbors and k-nearest neighbors
   Reading: Chapter 3.1(nearest neighbors) and 3.5 (distance measures) in Data Mining, Chapter sections 2.3.2, 2.3.3 (nearest neighbors), and 13.3 (kth nearest neighbors) in Statistical Learning. For information about nearest neighbors also see this book (especially the introduction): `http://tripod.brynmawr.edu.ezproxy.haverford.edu/find/Record/.b3282286`
   Lab work: Work on lab 2. d3 basics.
   Lab 1 Due: data cleaning.

4. Visualization frameworks
   Reading: Chapters 2 and 3 in Visualization, Chapters 3 - 6 in D3.

5. Hierarchical clustering and labeling
   Reading: Chapter 7.2 in Data Mining
   Lab 2 Due: d3 basics and linear regression graphing.

6. Visualizations - overview first, zoom and filter, details on demand.
   Reading: Visualization mantras chapter in Visualization textbook.

7. Midterm exam week
   Lab 3 Due: nearest neighbor searching for missing data
   Wednesday, March 2nd - **Midterm exam**.

8. SPRING BREAK!

9. Clustering overview, some clustering via proximity (k-center, etc.), k-means clustering, Lloyd's algorithm
   Reading: Chapter 7.1 in Data Mining, Chapter 13.1 in Statistical Learning, 7.3 in Data Mining, Suresh's clustering series posts 2 - 5 (`http://geomblog.blogspot.com/p/conceptual-view-of-clustering.html`)

10. Choosing k and visualization with filtering, correlation clustering, force-directed layout
    Reading: Chapter 10 in Visualization, correlation clustering Wikipedia page and this blog post:
    `http://blog.computationalcomplexity.org/2006/08/correlation-clustering.html`
    Lab 4 Due: Clustering lab 2 - hierarchical clustering.

11. Network analysis intro: adjacency matrices, adjacency lists, graph theory intro, network visualizations, color, network analysis basics. Dijkstra's algorithm.
    Reading: 7.2 (link marks) and 7.3 (color) in Visualization

12. Betweenness centrality and PageRank.
    Reading: Chapter 5.1 and 5.2 (PageRank) in Data Mining and 14.10 (PageRank) in Statistical Learning
    Lab 5 Due: Clustering lab 3 - k-means clustering and finding k.

13. Supervised learning: Evaluating training vs. test data. Linear regression as a model. Decision trees. Over fitting. Ensembles. Boosting.

14. Naive Bayes. Visualization case studies (from the NY Times). Memory limitations. Use lab time to work on your posters and your lab.
    Lab 6 Due: Network analysis lab - PageRank

15. Advanced topics. Data set discussions and poster session.
    Poster printing appointments at the KINSC office
    Due Wednesday: Poster session during class time.

Total grade breakdown

Participation and Attendance	5%
Labs	35%
Midterm	20%
Final Project	40%

Grades will be awarded based on the number of points earned and according to the percentage breakdowns shown. Students will not be graded on a curve.

Final Project

Students will work to analyze a data set throughout the semester. They will be responsible for choosing an appropriate analysis method and creating an associated visualization. Based on their findings, they will write a research paper including a description of their methods and the analysis performed, an explanation of their findings, and the visualization produced.

See the separate project details description for more information.

Late work policy

All extensions must be requested **at least 24 hours in advance** of the deadline. Extensions will be granted based on individual circumstances. Work handed in late without a previously granted extension may not be accepted.

Rules and Pet Peeves

- **Be on time.** This includes class, lab, office hours, and appointments.
- **No computer use in class without approval.** Computers should only be used in class to take notes. This includes laptops, tablets, phones, etc..
- **Expect 24 hours before an email response** and read all emails within 24 hours.
- **Attend all classes and labs.**

Collaboration

You are encouraged to discuss the lecture material and the labs and problems with other students, subject to the following restriction: the only "product" of your discussion should be your memory/understanding of it - you may not write up solutions together, or exchange written work or computer files. Any group projects are the only exception to this - in these cases, these collaboration rules apply to students outside of your group and you may freely work closely with students within your group. Collaboration is not allowed on examinations or quizzes.

As usual, anything taken from outside sources should be cited. Code should not be copied without permission from the instructor. If permission is given, code should be cited at the location it is used with a comment.

Learning Accommodations

Haverford College is committed to supporting the learning process for all students. Please contact me as soon as possible if you are having difficulties in the course. There are also many resources on campus available to you as a student, including the Office of Academic Resources (`https://www.haverford.edu/oar/`) and the Office of Access and Disability Services (`https://www.haverford.edu/access-and-disability-services/`). If you think you may need accommodations because of a disability, you should contact Access and Disability Services at `hc-ads@haverford.edu`. If you have already been approved to receive academic accommodations and would like to request accommodations in this course because of a disability, please meet with me privately at the beginning of the semester (ideally within the first two weeks) with your verification letter.