MS&E 226: “Small” Data

Time and place

Tuesdays and Thursdays, 1:30-2:50 PM
Gates B3

Discussion sections Fridays, 12:30-1:20 PM
Skilling Auditorium

Note: *The discussion section was moved to the time and location above so that we can record it. In place of the 1:30-2:50 discussion section, we will hold office hours.*

Instructor

Ramesh Johari
Associate Professor
Management Science and Engineering
Computer Science (by courtesy)
Electrical Engineering (by courtesy)
Huang Engineering Center, Room 311
E-mail: rjohari@stanford.edu
Admin: Amy Kao, amykao1@stanford.edu
Office hours: Tuesdays, 3-4:20 PM in Huang 311; Fridays, 1:30-2:50 in Thornton 110
Additional office hours by appointment only (contact Amy Kao)

Teaching assistants

Amelia Lemionet
E-mail: lemionet@stanford.edu
Office hours: Tuesdays 5:00-7:00 PM in <<TBA>>

Carlos Riquelme Ruiz
E-mail: rikel@stanford.edu
Office hours: Mondays 3:00-5:00PM in <<TBA>>

Sven Schmit
E-mail: schmit@stanford.edu
Office hours: Wednesdays 4:00-6:00PM in Y2E2 335

David Walsh
E-mail: dwalsh@stanford.edu
Office hours: Wednesdays 1:00-3:00PM in Huang B008.

Course websites

Public (lecture notes, resources): http://web.stanford.edu/class/msande226
Piazza (communication, announcements, materials): piazza.com/stanford/fall2016/mse226
Gradescope (grading, problem set submission): https://gradescope.com/ Entry code: **94Y46M**

Catalog course description

This course is about understanding "small data": these are datasets that allow interaction, visualization, exploration, and analysis on a local machine. The material provides an introduction to applied data analysis, with an emphasis on providing a conceptual framework for thinking about data from both statistical and machine learning perspectives. Topics will be drawn from the following list, depending on time constraints and class interest: approaches to data analysis: statistics (frequentist, Bayesian) and machine learning; binary classification; regression; bootstrapping; causal inference and experimental design; multiple hypothesis testing. Class lectures will be supplemented by data-driven problem sets and a project.

Detailed course description

Before figuring out what a “small” data class is about, it’s important to understand what “big data” means.

It’s hard to walk down the street these days (at least at Stanford) without hearing the phrase “big data”. A casual Google search for headlines including “big data” turns up the following:

- Big oil turns to Big Data as oil prices plummet
- Can Big Data help us fight rising suicide rates?
- When accounting meets Big Data
- Big Data: Searching for drug side effects
- Using Big Data, IoT to track & control building energy efficiency

Informally, the phrase “big data” is often used to refer to the information explosion that is providing data across every walk of human endeavor at finer granularity than ever possible before. More formally, “big data” refers to data that *cannot be stored or analyzed on a single machine*, and instead requires a cluster of machines (or more) to be processed and studied.

Big data has prompted a revolution in algorithms for managing, storing, and operating on data: MapReduce, Hadoop, Spark, etc. These are all fast becoming household names (at least in the technology industry) to make working with massive datasets easier.

But there are two points that should give us pause in the rush to big data. First, even as datasets grow ever larger, computational technology is also advancing -- so the size of a dataset that can be processed on a single machine is becoming much larger than before (for example, it is now common to find 32GB of RAM on a typical household PC, and much more on enterprise machines). This is formally what we mean when we say “small data”: in contrast to big data, it is any dataset that can be stored, processed, analyzed, and visualized on a single machine, typically interactively.

Second, and more importantly, often the most important part of the analytical pipeline in understanding a large dataset is actually understanding small datasets first. It is in the realm of small data where we first understand techniques for summarizing and exploring data, for drawing inferences from data, for making predictions, and ultimately for making decisions. This course is about building that foundation.

Our approach is built on the old cliché: "Give a man a fish and you feed him for a day; teach a man to fish and you feed him for a lifetime." This is not a vocational course: you will not learn a large number of different tools for analyzing data, and we will not be spending a lot of time on the details of good data processing practice. Among other things, this means (if you intend to be a good data scientist) you should also find ways to learn SQL, R, Python (or your favorite language), and a range of different algorithmic techniques for analyzing data. There are plenty of ways to do this now, both through other courses at Stanford and through a large number of resources online. Data processing tools change fast, and more than anything else you will need to *learn how to learn.*

With that in mind the primary goal of this course is to give you a footing in which to ask critical questions about different methods you will encounter over a lifetime of working with data. We will teach you: how to be careful in defining your objective; how to compare and contrast different approaches to the same problem; the differences and similarities between frequentist statistics, Bayesian statistics, and machine learning approaches; checking model assumptions; and how to use data to draw causal inferences. We use linear regression as a playground for our study, but a successful student in this class should be able to use the insights in a class like this to become a more educated consumer of an entire "toolbox" of data analysis methods.

Course outline

1. *Summarization* (2 weeks). Given a single data set, how do we summarize it? Basic sample statistics. Using models to succinctly summarize data. The algebra of linear regression and logistic regression. In-sample measures of fit: R^2 and residuals.
2. *Prediction* (2-3 weeks). How do we generalize our understanding of a data set to *new* samples? Formalizing the prediction problem. Binary classification. Linear regression and logistic regression as approaches to prediction. Model complexity and the bias-variance tradeoff. Training vs. test sets and cross validation.
3. *Inference* (2-3 weeks). How do we generalize our understanding of a data set to draw inferences about the population or system from which the data came? The basics of frequentist estimation and hypothesis testing. Application to linear regression. The bootstrap. The multiple hypothesis testing problem. Comparison to Bayesian estimation and hypothesis testing.
4. *Causality* (2 weeks). How do we determine the effect that *changing* a system will have? The Rubin causal model, potential outcomes, and counterfactuals. The "gold standard":

randomized experiments. The basics of causal inference from observational data. From causal inference to data-driven decisions.

Prerequisites

This is a quantitative, mathematical, and computational course. Accordingly, these are the prerequisites.

(1) *Math 51.* A hard constraint; do not enroll if you have not had multivariable calculus and linear algebra. I will ***not*** be recapping basic concepts in class, so this should be technology you have at your fingertips.

(2) *Probability at the level of MS&E 120 or Stats 116.* Also a hard constraint, just like Math 51. We will not be reviewing probability in this class.

(3) Computational methodology. You will need to be able to work with datasets. While you don't need to be a master, some familiarity (going into the class) with one of R, Matlab, or Python is important. The most significant computational work in the class will happen in the second half. Therefore, concurrent enrollment in Stats/CME 195 is an acceptable way to fulfill this requirement.

(4) Quantitative curiosity. Though we will do few proofs, this class is about learning how to think critically about quantitative methods. Therefore getting the most out of the class requires "opening the hood", and being curious about why things work the way they do. If you are uncomfortable thinking quantitatively, this is not the class for you.

Evaluation

Your evaluation will be based on a combination of five problem sets, a midterm exam (including both an in-class and a take-home component), an in-class final exam, and a guided mini-project that will run through the final three problem sets. The grade will be determined as follows:

- Problem sets: 50%
- In-class component of the midterm: 10%
- Take-home component of the midterm: 10%
- In-class final exam: 10%
- Mini-project: 20%

Problem sets

There will be a total of 5 problem sets. Problem sets must be submitted online through Gradescope (https://gradescope.com/; entry code 94Y46M). *Note that Gradescope will only be used for homework submission and grading; all announcements and discussions will be handled through Piazza (above).* Problem set sheets will also be posted on Piazza (in the resources section).

Except for medical necessity, no late problem sets will be accepted. All assignments will be posted to the course website. Problem sets are assigned and due as follows:

- Problem set 1: Handed out on 9/27, due on 10/4

- Problem set 2: Handed out on 9/29, due on 10/13
- Problem set 3: Handed out on 10/13, due on 10/27
- Problem set 4: Handed out on 11/3, due on 11/17
- Problem set 5: Handed out on 11/17, due on 12/1

Depending on their length and difficulty, the total number of points in each set might vary. Each part of each problem will be graded from 1 to 3 points. You will receive 3 points if you understand the problem and solve it correctly (except for minor numerical errors); you will receive 2 points if there are substantive issues with your solution, but your general understanding is correct; and you will receive 1 point if there are major issues with your solution.

On any of the non-computational components of the problem sets (see below for more on the computational elements of the class), you can discuss the assignments among yourselves, but everybody must turn in his/her own written solutions in his/her own words. If you do a substantial subset of the work on your problem set with others, document on each assignment the other students that you worked with.

On any computational component of any of the problem sets, you can work in pairs. In this case, each student should turn in a complete problem set, including the computational components.

Mini-project

Starting with Problem Set 2, we will guide you through an applied mini-project, learning to apply some of the methods in this class in greater depth. We will provide more details of the structure of this project early in the quarter.

As with the computational parts of the problem sets, for the mini-project, you can work in pairs. Again, only one copy of your project needs to be turned in per pair.

Course communications: Piazza

As noted above, we will use Piazza to manage course announcements and a discussion forum. You can sign up for the course on Piazza here:
piazza.com/stanford/fall2016/mse226

Please use Piazza for all course-related communication with us. We will aim to respond to questions in a 24-48 hour period, except of course for those of an urgent nature (e.g., typos on problem sets or lecture notes, clarifying course logistics, etc.). Among other things this means you should not wait until the last day before a problem set is due to message us; we will likely not respond in time.

You are *strongly encouraged* to attend office hours to ask questions of a technical nature; these are best discussed in a face-to-face setting, given the quantitative nature of the course.

If you are having difficulty, find help right away— *do not wait until you fall even further behind!* There is an obvious temptation to wait until the day before the due date to do all the work on the problem sets, and I can assure you this approach will almost certainly lead to very poor performance in the class.

Student participation on Piazza

We would love to have students help answer Piazza questions -- particularly in the "waiting period" before course staff answers. Thus we are providing an incentive for you to participate in Q&A! Every time you answer a question that is marked "good answer" by an instructor, you will receive one lottery ticket. At the end of the quarter these tickets will be entered into a lottery for a $200 Amazon gift card.

Exam policy

The midterm exam will be held in two parts. On November 1, there will be an in-class multiple choice exam. In addition, on November 1 at the end of class we will post a take-home component, that will be due back by the beginning of class on November 3. The midterm will cover all topics from the first part of the class.

The final exam will be held at the time set by the registrar, from 12:15-3:15 PM on December 12. It will be a multiple choice exam, of similar length to the midterm. The final exam will cover all topics from the second half of the class.

Except for medical necessity, *there will be no alternate exam dates or times.* You should only register for the class if you are certain you can take the exams on these dates.

Discussion sections

Most Fridays we will hold a discussion section from 12:30-1:20 PM in Skilling Auditorium. This will be an opportunity for students to get additional help on the material, in a more guided session. We will use these sessions to go over material from the lectures, and to walk through problems that will help you solve the problem sets.

Computation

Many of the problem sets, as well as the guided mini-project, will require you to be comfortable carrying out computations on data. For this purpose, the "official" language of the course is R; if you want to develop facility with R, it is sufficient to enroll concurrently in Stats/CME 195.

We will provide links to R on the website, as well as any datasets that are needed through the course of the quarter.

As noted above, *you are welcome to work in pairs on any computational component in the class, and on the mini-project.*

Suggested textbooks

I will distribute notes as the course goes on. In addition, you may find it helpful to have the following textbooks on hand. I'm not requiring them, because they are available online, and we will not be linearly working our way through any of them. At the same time, they are not particularly expensive (by textbook standards). Some problems will also be drawn from these books.

1. *All of Statistics*, by Larry Wasserman. This book is as ambitious in scope as the title suggests. The obvious tradeoff for this breadth is a lack of depth. In particular, this book provides less of the nuances of how data analysis can easily go wrong; on the other hand, it is one of the few places where you can find statistics and machine learning treated together (in a somewhat accessible way).
   The book can be accessed online (for Stanford students) here:
   http://link.springer.com/book/10.1007/978-0-387-21736-9
   It is available through Amazon here:
   http://www.amazon.com/All-Statistics-Statistical-Inference-Springer/dp/0387402721
   (Though note that you can buy the "MyCopy" softcover edition for less through Springer at the first link above.)
   Statistical Models, by David Freedman. This book is a great complement to the Wasserman book. Although targeted nominally at statistical methods for the social sciences, it is unique in its healthy skepticism of statistical methods applied without regard to modeling assumptions. Freedman was well known for his advocacy of "shoe leather" statistics: in his view good data analysis required a substantial investment of on-the-ground effort to understand the context of the data you were analyzing. As a result his book spends a lot of time asking the reader to critically evaluate the methods that are being taught.
   This book can be accessed online (for Stanford students) here:
   http://ebooks.cambridge.org/ebook.jsf?bid=CBO9780511815867
   It is available through Amazon here:
   http://www.amazon.com/Statistical-Models-Practice-David-Freedman/dp/0521743850
2. *Data Analysis Using Regression and Multilevel/Hierarchical Models*, by Andrew Gelman and Jennifer Hill. Though primarily about methods for dealing with multilevel models (e.g., models with random effects, fixed effects, and mixed effects), this book is also one that provides great insight into how to think about data more generally. I will be drawing on the first two parts of this book to help describe various modeling and interpretation issues in applying linear regression. Unfortunately this book is not available online; here is the book site:
   http://www.stat.columbia.edu/~gelman/arm/
   It is available through Amazon here:
   http://www.amazon.com/Analysis-Regression-Multilevel-Hierarchical-Models/dp/052168689X

Other references

Here is an assortment of other books that you may find useful to consult.

1. Elements of Statistical Learning, by Hastie, Tibshirani, and Friedman. This is a comprehensive reference on prediction models for classification and regression.
2. Bayesian Data Analysis, by Gelman, Carlin, Stern, Dunson, Vehtari, Rubin. A great reference on applied Bayesian statistics.
3. Statistics for Experimenters, by Box, Hunter, and Hunter. A classic reference on experimental design.
4. Causal Inference for Statistics, Social, and Biomedical Sciences, by Imbens and Rubin. A recent and comprehensive book on methods for causal inference from data, especially with deep coverage of methods for observational data analysis.
5. A Guide to Econometrics, by Kennedy. A good companion to the Imbens and Rubin text, Kennedy's book helps you understand common ways in which causal inference and econometric analysis can go awry.
6. Mostly Harmless Econometrics, by Angrist and Pischke. A succinct and accessible treatment of the basics of causal inference.
7. Theory of Point Estimation, Lehmann and Casella; and Testing Statistical Hypotheses, by Lehmann and Romano. These are advanced theoretical texts for the foundations of estimation and hypothesis testing (often used for Stats 300A).
8. Computer Age Statistical Inference, by Efron and Hastie. This brand new text covers a range of topics relevant to statistical inference in the age of big data. It is notable for being written by two of the most well-known living statisticians, and for the breadth of scope. Highly recommended.