MS&E 226: "Small" Data

Lecture 1: Introduction (v3)

Ramesh Johari
rjohari@stanford.edu

September 28, 2016

What is this class about?

“Big” data

We are collecting data at unprecedented levels of granularity.

- Billions of: Facebook posts, tweets, medical tests, power meter readings...
- Often arriving faster than we can store and analyze it

Key feature of “big” data:
Can’t be analyzed on a single machine.

Requires new algorithms and tools to store, query, and analyze the data.

“Small” data

Data that *can* be analyzed, processed, etc., on a single machine.
Keep in mind:

- Advances in technology means even “small” data is getting bigger
  (e.g., 32GB of RAM even on home PCs)
- Most analysis of “big” data starts by understanding “small” data

This class is a user's manual for “small” data analysis.

In the process you will learn skills that should help you for *any* data analysis.

Key features

- Conceptual rather than vocational: emphasis on how to reason about different approaches to data analysis
- Comparison and contrast between different approaches: machine learning, (frequentist and Bayesian) statistical inference
- Emphasis on articulating your objective carefully

Organization

1. **Summarization** (2 weeks).
   - Given a single data set, how do we summarize it?
   - Basic sample statistics; models; linear and logistic regression; in-sample fit (R^2 and residuals).
2. **Prediction** (2-3 weeks).
   - How do we generalize our understanding of a data set to new samples?
   - Binary classification; linear regression and logistic regression as approaches to prediction; model complexity and the bias-variance decomposition; out-of-sample validation.

Organization

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?
   - Frequentist estimation and hypothesis testing; application to linear regression; bootstrap; multiple hypothesis testing. Comparison to Bayesian approaches.
4. **Causality** (2 weeks).
   - How do we determine the effect that changing a system will have?
   - The Rubin causal model, potential outcomes, and counterfactuals; randomized experiments; causal inference from observational data; data-driven decision making.

Who is this class for?

- Targeted as a *first course* in statistical inference and machine learning.
- Students with either deep backgrounds in one of machine learning *or* statistics tend to benefit from seeing both treated on a common footing, though there may be some redundancy in technical concepts with things you've seen before. You should decide whether the redundancy is worth the conceptual unification.
- Students with deep backgrounds in machine learning *and* statistics should probably not take this class.

Course logistics

Basic info

- Public site: `http://web.stanford.edu/class/msande226`
- Piazza: `http://piazza.com/stanford/fall2016/mse226`
- Gradescope: `https://gradescope.com/` (entry code 94Y46M)
- Details in syllabus on website
- Course assistants:
  Amelia Lemionet, Carlos Riquelme, Sven Schmit, David Walsh
- Discussion sections: Fridays 1:30-2:50 PM

Important dates

No extensions or alternates!

- Problem sets
  - PS1: Out 9/27, due 10/4
  - PS2: Out 9/29, due 10/13
  - PS3: Out 10/13, due 10/27
  - PS4: Out 11/3, due 11/17
  - PS5: Out 11/17, due 12/1
- Exams
  - In-class midterm: 11/1
  - Take-home midterm: Out 11/1, due 11/3
  - Final exam: 12/12, 12:15-3:15 PM

Evaluation

- Each problem set: 10%
- In-class component of the midterm: 10%
- Take-home component of the midterm: 10%
- In-class final exam: 10%
- Mini-project: 20%

Communicating with us

- Use Piazza for all course-related communication
- We will aim to respond within 24-48 hours (or more quickly as needed for urgent inquiries, e.g., logistics or clarification)
- Use office hours for technical questions
- Win a $200 Amazon gift card by answering questions on Piazza
  - Every answer marked "good answer" by an instructor gains one entry into a lottery at the end of the quarter

Notes

- Lecture notes will be posted to the site
- Suggested (but not required!) texts:
  - Wasserman, *All of Statistics* ([AoS]).
  - Freedman, *Statistical Models: Theory and Practice* ([SM]).
  - Gelman and Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models* ([DAR]).