MS&E 226: "Small" Data

Lecture 1: Introduction

Ramesh Johari
rjohari@stanford.edu

What is this class about?

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.