The data scientist's toolbox

STA 380
James Scott

About the course

First half of the course: supervised learning.

- Given past data on outcomes y paired with features x, can we find patterns that allow us to build a model for $(y \mid x)$?
- Key characteristic: there is a single privileged outcome y.

Second half of the course: unsupervised learning.

- We still have multivariate data and want to find patterns.
- But there is no single privileged outcome. ("Everything is y.")
- Example: "Here's data on the shopping basket of every Whole Foods customer at 6th and Lamar last month. Find some patterns that we can use to improve product placement."

About the course

Daily structure

- Each day: 2 × 50 minute modules (roughly)
- Traditional lecture to provide background on a technique
- Hands-on walkthrough of an example data analysis

Evaluation

- 2 homework assignments (40% each, due 8/6 and 8/14)
- Scribing (20%)

Scribing

- Each person is responsible for a 50-minute module.
- Goal: provide a definitive set of notes for that day to share with your classmates.
- Scribe notes will be posted on the website.

About the course

There are many labels for what we're doing:

- Econometrics, statistics: traditionally focused on formally quantifying uncertainty.
- Business analytics, data science, data mining: traditionally focused on pragmatic data-analysis tools for applied problems.
- Machine learning, pattern recognition, artificial intelligence: traditionally focused on algorithms with engineering-style performance guarantees.

How confusing! My working definitions:

- Data science: the *collection*, *processing*, *analysis*, and *summary* of data in an attempt to answer an empirical question or make a decision.
- Business analytics: data science focused on business, industrial, and internet-scale applications

About the course

How our main goals fit into this jungle:

- Infer patterns from noisy, complex data.
- If at all possible, do so using simple, scalable algorithms (machine learning, AI).
- If necessary, provide error bars (statistics/econometrics).
- Always be aware of the problem context or decision at hand.

Among economists, data mining is a dirty word. We'll strive to earn it a better reputation!

About the course

What's different about "big data"?

- Big in the number of observations n. . .
- . . . and in the number of features p.

In this setting, we can't just apply a recipe you learned in Stat 101:

1. Look at individual variables or pairs of variables and make a decision (t tests, χ^2 tests, etc.)
2. Choose among a small set of pre-specified models (F tests, permutation tests)
3. Plot everything to look for interactions, useful transformations, or violations of model assumptions.

When data sets get large, some tools remain the same, while others are completely new.

About the course

The four pillars of data science:

1. Data collection
2. Data cleaning (pre-processing/hacking/"munging")
3. Analysis
4. Summary (figures + prose)

This course focuses a little on 2, heavily on 3-4, and not at all on 1.

Data collection and cleaning: principles

On collection, management, and storage: a full subject unto itself. (I'm happy to provide references, but this isn't the part of data science we cover in this course.)

On cleaning: I defer to Jeff Leek's description of "How to Share Data with a Statistician." (See course readings.) Always provide:

1. The raw data.
2. Tidy data.
3. A variable "code book" written in easily understood language.
4. A complete, fully reproducible recipe of how the clean data was produced from the raw.

Data analysis and summary: principles

Our watchwords are *transparency* and *reproducibility*.

- The end product: you will write a report with beautiful figures, and someone else will marvel at it.
- Data science is hard enough already: there is *zero room* for ambiguity or confusion about data or methods.
- Any competent person should be able to read your description and reproduce *exactly* what you did.

The ideal: “hit-enter” reproducibility.

- Someone hits enter; your analyses and figures are reproduced from scratch and merged with prose, before their eyes.
- We will rely on a handful of easily mastered software tools to put this ideal into practice.

Data analysis and summary: principles

All reports involve three main things:

1. A question: what are we doing here?
2. Evidence: a set of figures, tables, and numerical summaries based on the analyses performed.
3. Conclusions: what did we learn?

The basic recipe for writing a statistical report:

1. Make the key figures and tables first.
2. Write detailed captions for each one.
3. Put these figures and tables in order (question, then answer).
4. Write the story around these main pieces of evidence.

Our basic software tools

Markdown

- A simple markup language for generating a wide variety of output formats (HTML, PDF) from plain text documents.
- Two pillars: (1) a formatting language; (2) a conversion tool.
- Much simpler than, for example, HTML.

R and RMarkdown

- Ideal for real-time exploration of data.
- Has a large ecosystem of libraries.
- RMarkdown: generate Markdown reports with R integration.

git and Github

- git: software for version control; ideal for collaborative work.
- Github: a git repository hosting service.

Our basic software tools

R + Markdown + GitHub: a short demo of each