TEXAS McCombs

MASTER OF SCIENCE IN BUSINESS ANALYTICS

Introduction + The data scientist's toolbox

Roadmap for our course

This class is about gaining knowledge from raw data. You'll learn to use large and complicated data sets to make better decisions.

A mix of practice and principles:

- Solid understanding of essential statistical ideas
- Concrete data-crunching skills
- Best-practice guidelines

We'll learn what to trust, how to use it, and how to learn more.

There are two main parts to our course: **supervised learning** and **unsupervised learning**.

Supervised learning

Given past data on outcomes y paired with features x, can we find patterns that allow us to predict y using x?

Key characteristic: there is a single privileged outcome y.

Example: a house has 3 bedrooms (x_1), 2 bathrooms (x_2), 2100 square feet (x_2), and is located in Hyde Park (x_4). What price (y) should it sell for?

In real life, there might be hundreds or thousands of features. If you know regression: this is like regression on steroids!

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."

An alphabet soup of labels...

Statistical learning, data mining, data science, ML, AI... there are many labels for what we're doing!

Econometrics, statistics: focused on understanding the underlying phenomena and formally quantifying uncertainty.

Business analytics, data science, data mining: traditionally focused on pragmatic data-analysis tools for applied prediction problems.

Machine learning, pattern recognition, artificial intelligence: focused on algorithms with engineering-style performance guarantees.

About "data mining"...

Among economists, "data mining" is a dirty word. Example: the "Lucas critique":

- Fort Knox has never been robbed.
- Thus historically, there's a zero correlation between security spending at Fort Knox (x) and the likelihood of being robbed (y).
- Naive "data mining" conclusion: leave Fort Knox unguarded!

Thus historically, there's a zero correlation between security spending at Fort Knox (x) and the likelihood of being robbed (y).

This is a total caricature. We'll strive to give data mining a better reputation :-)

What does it mean for data to be "big"?

Big in either or both:

- the number of observations (size n)
- the number of features or predcitor variables (dimension p).

In these settings, you cannot:

- Look at each individual variable and make a decision (t-tests).
- Choose amongst a small set of candidate models (specification tests from stats or econometrics).
- Plot every variable to look for interactions or transformations.

Good data mining = inference at scale

Some data-mining tools are familiar, or familiar with a twist:

$\rightarrow$ linear regression
$\rightarrow$ p-values
$\rightarrow$ automatically select a set of relevant feature variables, then fit a linear model

Some are totally new:

$\rightarrow$ PCA
$\rightarrow$ K-means

All require a different approach when n and p get really big.

People use these tools everywhere

- Mining client information: Who buys your stuff, what do they pay, what do they think of your new product?
- Online behavior: Who is on what websites, what do they buy, how do/can we predict or nudge behavior?
- Collaborative filtering: predict preferences from people who do what you do; recommender engines.
- Text mining: Connect blogs/emails/news to sentiment, beliefs, or intent. Parsing unstructured data.
- Big regression: mining data to predict asset prices; using unstructured data as controls in observational studies.

The four pillars of data science

1. Data collection
2. Data cleaning (pre-processing/hacking)
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

You will analyze a lot of data in this course. 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.

Data analysis and summary: principles

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: R, Markdown, and Git

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?

Data analysis and summary: principles

The basic recipe for writing a statistical report:

1. Make the key figures and tables first.
2. Write detailed, self-contained 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.

This helps avoid “fear of the blank page”!

Our software toolkit

- R: for data analysis
- Markdown and RMarkdown: for writing reports
- GitHub: for collaboration and dissemination or results. The location of our course website, code, and data.

R: an immensely capable, industrial-strength platform for data analysis.

It's used everywhere:

- Academic research (stats, marketing, finance, genetics, engineering)
- Industry (Google, Microsoft, eBay, Boring, Citadel, IBM, New York Times)
- Governments/NGOs (Rand, DOE, National Labs, US Navy)

R is free and looks the same on all platforms, so you'll always be able to use it.

R

A huge strength of `R` is that it is open-source. `R` has a *core*, to which anyone can add contributed packages.

- >18,000 packages, as varied as the people who write them
- Some are specific, others general
- Some are great, some decent but unpolished, some are crappy

`R` has flaws, but so do all options (e.g., Python is great, but the community of stats developers is smaller, interactive data analysis is less slick, and you need to be a more careful and sophisticated programmer.)

Most prefer to use R via an IDE. We'll use *RStudio*.

Markdown

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

`Rmarkdown` allows you to write up data analyses easily within `R` to make reproducible reports. You can install the package directly in `R` by running the following command:

```
install.packages("rmarkdown")
```