Gov 2018: Intro to ML

Lecture 1: Introduction

Connor Jerzak, Naijia Liu

January 24, 2024

Rich and Complex Data

- Traditional data sources are limited: governments, national survey, etc.
- Today, massive amounts of data are available from diverse sources
  - Fine-grained economic data: stock fluctuations, product-level trade data, ...
  - Text data: party platforms, Twitter, blogs ...
  - Image data: surveillance cameras, bodycams, satellites...
  - Audio data: parliamentary debate, campaign speeches, Supreme Court oral arguments...
  - Network data: Facebook, call records...
  - GIS data: mobile location data, road construction, oil discoveries...

Introduction

- Main course goal: develop an understanding of machine learning models for measurement in complex and high dimensional data
- Broad overview:
  - Begin with tabular data and relevant models
  - Then, non-tabular data
  - Develop statistical underpinnings for these models
  - In parallel, conduct a serious research project applying these models

Rich and Complex Data

Numerous complications with using this data:

- Many observations, high dimension: difficult to visualize and analyze
- Constraints on computational power, working memory, storage
- Unstructured (non rectangular) data: nesting, high missingness, unequal length...

Also numerous opportunities:

- New types of data allow testing new substantive ideas
- Room to develop new methods for data analysis
- Ability to detect more complex patterns than before

Rich and Complex Data: Structuring Unstructured Data

- Raw representations of
  - Text: sequences of discrete words
  - Networks: message sent/received logs
  - Images: pixel intensities
  - Audio: time-series pressure readings
  - Movement: timestamped latitude/longitude
- Often, we want to reorganize (aggregate, summarize, featurize) raw data before analysis

Rich and Complex Data: Structuring Unstructured Data

- Intermediate quantities of interest:
  - Text: relationships between words, grouping into topics, measuring sentiment (e.g. toward taxation) or ideological positions
  - Networks: community detection, inter-group relationships
  - Images: landmarking, object detection, pose estimation
  - Audio: speaker recognition, emotion detection, event localization
  - Movement: participation in political rallies, inter-group mixing
- End goal: Answer to a theoretically motivated puzzle

Broad Categories of Statistical Learning

- Supervised learning
  - Researcher determines the categories in advance
  - Labels/values are available for a subset of observations
  - Learn relationship between features and the label of known observations
- Unsupervised learning
  - No pre-determined categories
  - Discover patterns within the data
- Middle ground: semi-supervised learning, active learning
- Numerous techniques depending on whether quantity of interest is discrete or continuous, dependence structure between observations, etc.

The Role of Machine Learning in Research

A typical research workflow:

1. Define the problem
2. Research design
3. Collect data
4. Represent data
5. Analyze data
6. Validate results

The Role of Machine Learning in Research

Better analysis and representation of high-dimensional data

- Fit more flexible models with fewer functional form assumptions
- Generate better instrumental variables from composite of weak instruments
- Improve imputation of missing data

Improvements over existing approaches can be quantified in terms of estimator bias, estimator variance, predictive accuracy, etc.

The Role of Machine Learning in Research

Measurement of difficult-to-observe characteristics

- Idea: we care about a latent variable that generates the observed (high-dimensional) data
- Use machine learning to infer values of the latent variable for each observation

Final Project

Thoughts on choosing and refining a topic:

- The best research is problem-driven research.
- What is the question? What is the empirical puzzle you're attempting to resolve?
- If you had infinite time and resources, what would you do?
- Some examples of good questions:
  - How do legislators relate to their constituents? (Grimmer 2010, 2013)
  - What explains variation in media slant? (Gentkow and Shapiro 2010)
  - How do campaign consultants disseminate campaign strategy across candidates? (Nyhan and Montgomery 2015)
- Some examples of "bad" questions:
  - Prediction problems that don't connect to a theoretical question: "Can we predict if John Doe voted?"
  - Questions that rely on extrapolation: "Will the United States build a wall at the US-Mexico border?"

Final Project

Thoughts on implementing a project:

- All models are wrong, but some are useful
- Quantitative methods augment humans, not replace them
- There is no globally best method
- Validate, validate, validate
- Communicate your uncertainty in the model

Prerequisites

- Courses:
  Gov 2001, Gov 2002, Gov 2003
  Or equivalent
- Mathematics:
  - Multivariate calculus
  - Linear algebra
  - Probability/statistics
- Programming
  - We will use **R** and Python for most of the class
  - Depending on the task, Python can be far more efficient than R (webscraping, neural networks)

Assignments

- Two problem sets, one take-home midterm and one final project
- Collaboration with classmates is encouraged
  Peer citation: Please cite your classmates if you get help / hints from them.

Assignments

- For your own benefit, you are strongly encouraged to write your own code
- Code that is not your own should be annotated with your own detailed comments to demonstrate understanding
- When code is too long to include directly, a description/pseudocode should be included in text
- Use informative variable names and write/comment code such that a reader (including future you!) can easily understand
- Read and follow the Google R Style guide (`https://google.github.io/styleguide/Rguide.xml`) for best practices

Final Project

- Apply a method from this course to answer a substantive question of interest with your existing data
- Use a method from this course to generate new data
- Replicate an existing study and improve it with better methods
- Coauthoring is encouraged!

Important Deadlines

- **Late February: Descriptive data analysis**
  - Why this data is better than previous data
  - Descriptive figures and tables with informative captions
  - Rough plan for analysis, main contribution
- **Mid March: poster proposal deadline**
  - Preliminary analysis
  - Abstract up to 500 words: summarize the research question, investigative methodology, and (preliminary) findings
  - You are encouraged to submit for **Polmeth summer meeting poster session**!
- **Late April: Poster session**