ORIE 4741: Learning with Big Messy Data

Introduction

Professor Udell
Operations Research and Information Engineering
Cornell

August 22, 2021
Outline

Logistics

Stories

Definitions

Kinds of learning

Syllabus
want to take this class?

▶ **ASAP:**
  ▶ enroll (or drop) (or get on wait list)
  ▶ fill out course survey
  ▶ sign up for discussion forum
  ▶ sign up for iClicker

▶ **Thursday 9/2/2021:** homework 0

links on course website:
https://people.orie.cornell.edu/mru8/orie4741/
Course staff

- Prof. Madeleine Udell
- TA: Richard Phillips (CS PhD)
- TA: Tao Jiang (ORIE PhD)
- TA: Connor Lawless (ORIE PhD)
- TA: Yuanping Du (ORIE MEng)
- TA: Max de Ledebur (ORIE MEng)
- TA: Jody Zhu (ORIE+CS Undergraduate)
- TA: Tara Khanna (ORIE Undergraduate)
- TA: Kevin Jiang (CS Undergraduate)
Tech stack

- Zoom for lectures
- Course website for course materials (syllabus, schedule, homework, project, etc)
- iClicker REEF for polls
- Zulip for Q&A and announcements
- Gradescope for quizzes, submitting homework, grades, solutions
- Github for code (demos, projects, and hw starter code)
Course requirements and grading

course website:
(grading, course requirements, lectures, homework, etc)
https://people.orie.cornell.edu/mru8/orie4741/

► (15%) Participation: for every lecture (after this one), use
  ► iClicker REEF for sync lectures
  ► participation form for async lectures
► (30%) Homework
  ► due every two weeks or so
  ► first one due next Thursday
► (15%) Quizzes
  ► 30 min quiz every week or so
► (40%) Project
Course requirements and grading

course website:
(grading, course requirements, lectures, homework, etc)
https://people.orie.cornell.edu/mru8/orie4741/

▶ (15%) Participation: for every lecture (after this one), use
  ▶ iClicker REEF for sync lectures
  ▶ participation form for async lectures
▶ (30%) Homework
  ▶ due every two weeks or so
  ▶ first one due next Thursday
▶ (15%) Quizzes
  ▶ 30 min quiz every week or so
▶ (40%) Project

FAQ:
▶ yes, you can take the class online (even async)
▶ yes, you can take section online (even async), or not take the section
ORIE 4741 vs 5741

ORIE 5741

- same course material as 4741
- expectations and rubric for course project somewhat different
  - more business-oriented project
  - more detailed problem formulation
  - project presentation required (in addition to report)
- rubric will reflect MEng learning outcomes:
  1. Mastery and Application of Core Disciplinary Knowledge
  2. Problem Formulation and Organization and Planning of the Solution Process
  3. Collaborative Problem Solving and Issue Resolution
  4. Communication of Knowledge, Ideas, and Decision Justification
  5. Self-Directed Learning and Professional Development

- 5741 students are recommended to work with other 5741 students on their project
- if you work with any 5741 student on a project, all will be graded according to 5741 rubric
Questions

during lecture:

▶ ask out loud
▶ zoom chat (to everyone, or to a TA)

outside of lecture:

▶ ask at office hours
▶ ask on discussion forum
▶ don’t send email
Outline

Logistics

Stories

Definitions

Kinds of learning

Syllabus
Oh, you work with big messy data? Maybe you could help us out...?
My career in big data

academic

▶ B.S. in Mathematics and Physics at Yale
▶ Ph.D. in Computational and Mathematical Engineering at Stanford
▶ postdoctoral fellow at the Center for the Mathematics of Information at Caltech
▶ professor in ORIE at Cornell
My career in big data

academic
► B.S. in Mathematics and Physics at Yale
► Ph.D. in Computational and Mathematical Engineering at Stanford
► postdoctoral fellow at the Center for the Mathematics of Information at Caltech
► professor in ORIE at Cornell

applied work
► finance: Goldman Sachs, BlackRock, Capital One, Schonfeld, Two Sigma, . . .
► tech: Google, Retina.ai, Marketing Attribution
► cybersecurity: DARPA, Expanse (formerly Qadium)
► healthcare: Apixio, Ontario
► clean energy: Aurora
► politics: Obama 2012
## Data table: politics

<table>
<thead>
<tr>
<th>age</th>
<th>gender</th>
<th>state</th>
<th>income</th>
<th>education</th>
<th>voted?</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>F</td>
<td>CT</td>
<td>$53,000</td>
<td>college</td>
<td>yes</td>
<td>Biden</td>
</tr>
<tr>
<td>57</td>
<td>?</td>
<td>NY</td>
<td>$19,000</td>
<td>high school</td>
<td>yes</td>
<td>?</td>
</tr>
<tr>
<td>?</td>
<td>M</td>
<td>CA</td>
<td>$102,000</td>
<td>masters</td>
<td>no</td>
<td>Trump</td>
</tr>
<tr>
<td>41</td>
<td>F</td>
<td>NV</td>
<td>$23,000</td>
<td>?</td>
<td>yes</td>
<td>Trump</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>

**goals:**
- detect demographic groups?
- find typical responses?
- identify related features?
- impute missing entries?
### Data table: politics

<table>
<thead>
<tr>
<th>age</th>
<th>gender</th>
<th>state</th>
<th>income</th>
<th>education</th>
<th>voted?</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>F</td>
<td>CT</td>
<td>$53,000</td>
<td>college</td>
<td>yes</td>
<td>Biden</td>
</tr>
<tr>
<td>57</td>
<td>?</td>
<td>NY</td>
<td>$19,000</td>
<td>high school</td>
<td>yes</td>
<td>?</td>
</tr>
<tr>
<td>?</td>
<td>M</td>
<td>CA</td>
<td>$102,000</td>
<td>masters</td>
<td>no</td>
<td>Trump</td>
</tr>
<tr>
<td>41</td>
<td>F</td>
<td>NV</td>
<td>$23,000</td>
<td></td>
<td>yes</td>
<td>Trump</td>
</tr>
</tbody>
</table>

goals:

- detect demographic groups?
- find typical responses?
- identify related features?
- impute missing entries?
How to get data for politics?

two major sources of data:

▶ census
▶ voter registration

data quality is critical!

▶ for hw0, you’ll respond to census + register to vote
▶ if eligible:
  ▶ census eligibility: college students (including foreign) should be counted at the on-campus or off-campus residence where they sleep most of the time (even if you went home before Census Day last spring, and even if you’re a foreign citizen)
  ▶ voter eligibility: US citizen 18+
## Data table: medicine

<table>
<thead>
<tr>
<th>age</th>
<th>gender</th>
<th>heart disease</th>
<th>statins?</th>
<th>⋮</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>F</td>
<td>yes</td>
<td>no</td>
<td>⋮</td>
</tr>
<tr>
<td>57</td>
<td>?</td>
<td>no</td>
<td>no</td>
<td>⋮</td>
</tr>
<tr>
<td>?</td>
<td>M</td>
<td>no</td>
<td>no</td>
<td>⋮</td>
</tr>
<tr>
<td>41</td>
<td>F</td>
<td>yes</td>
<td>yes</td>
<td>⋮</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>

- find similar patients?
- understand systemic healthcare needs?
- use symptoms to detect which patients have COVID-19?
- detect patients who had series of mini-strokes?
Application areas

- health
- politics
- governance
- advertising
- retail
- ecommerce
- finance
- ...
Outline

Logistics

Stories

Definitions

Kinds of learning

Syllabus
Big

- NASA, 1997: “taxing the capacities of main memory, local disk, and even remote disk”
Big

- NASA, 1997: “taxing the capacities of main memory, local disk, and even remote disk”
- OED, 2015: “data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges”

[1] Image courtesy of Kim Minor @ IBM
Big

- NASA, 1997: “taxing the capacities of main memory, local disk, and even remote disk”
- OED, 2015: “data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges”
- 4 Vs:

1. **Volume**: Data at rest
   - Terabytes to exabytes of existing data to process
2. **Velocity**: Data in motion
   - Streaming data, milliseconds to seconds to respond
3. **Variety**: Data in many forms
   - Structured, unstructured, text and multimedia
4. **Veracity**: Data in doubt
   - Uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception and model approximations

¹image courtesy of Kim Minor @ IBM
Big

- NASA, 1997: “taxing the capacities of main memory, local disk, and even remote disk”
- OED, 2015: “data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges”

4 Vs:

- **Volume**: Data at rest
  - Terabytes to exabytes of existing data to process
- **Velocity**: Data in motion
  - Streaming data, milliseconds to seconds to respond
- **Variety**: Data in many forms
  - Structured, unstructured, text and multimedia
- **Veracity**: Data in doubt
  - Uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception and model approximations

5th V: value

\(^1\)image courtesy of Kim Minor @ IBM
Big: our definition

Definition

An algorithm for **big data** is one with computational and memory requirements that scale **linearly** (or nearly linearly) in the size of the data.
Big: our definition

Definition

An algorithm for **big data** is one with computational and memory requirements that scale **linearly** (or nearly linearly) in the size of the data.

why this definition? independent of

- hardware
- business
An algorithm for **big data** is one with computational and memory requirements that scale **linearly** (or nearly linearly) in the size of the data.

why this definition? independent of

- hardware
- business

if you use only algorithms for **big data**, then you’re working with **big data**
Messy

- noisy: some (or all) values suffer errors, inaccuracies, or malicious corruption
Messy

- noisy: some (or all) values suffer errors, inaccuracies, or malicious corruption
- missing: some values are missing, inconsistent, not recorded, or lost
Messy

- noisy: some (or all) values suffer errors, inaccuracies, or malicious corruption
- missing: some values are missing, inconsistent, not recorded, or lost
- heterogeneous: values of many different types
  - continuous values (e.g., 4.2, $\pi$)
  - discrete values (e.g., 0, 4, 994)
  - nominal values (e.g., apple, banana, pear)
  - ordinal values (e.g., rarely, sometimes, often)
  - graphs or networks (e.g., person 1 is friends with person 2)
  - text (e.g., doctor’s note describing symptoms)
  - sets (e.g., items purchased)
Learning
Learning

► machine learning?
Learning

- machine learning?
- human learning?
Learning

- machine learning?
- human learning?
- when data is **big** and **messy**, machine help is essential for human learning!
Outline

Logistics

Stories

Definitions

Kinds of learning

Syllabus
Data table

\[ n \text { examples (patients, respondents, households, assets)} \]
\[ d \text { features (tests, questions, sensors, times)} \]

\[
A = \begin{bmatrix}
a_{11} & \cdots & a_{1d} \\
\vdots & \ddots & \vdots \\
a_{n1} & \cdots & a_{nd}
\end{bmatrix}
\]

- \( a_i \) is \( i \)th row of \( A \): feature vector for \( i \)th example
- \( a_{.j} \) is \( j \)th column of \( A \): values for \( j \)th feature across all examples
- \( a_{ij} \) is \( j \)th feature of \( i \)th example
**Supervised learning**

- identify one column of data that we want to predict

\[
\begin{bmatrix}
A
\end{bmatrix}
= \begin{bmatrix}
x_{11} & \cdots & x_{1\,d-1} & y_1 \\
\vdots & \ddots & \vdots & \vdots \\
x_{n1} & \cdots & x_{n\,d-1} & y_n
\end{bmatrix}
= \begin{bmatrix}
X & y
\end{bmatrix}
\]

- \( x_i \in \mathcal{X} \) for \( i = 1, \ldots, n \) are rows of \( X \)
- \( y_i \in \mathcal{Y} \) for \( i = 1, \ldots, n \) are entries of \( y \)
Supervised learning

- identify one column of data that we want to predict

\[
\begin{bmatrix}
A
\end{bmatrix} = \begin{bmatrix}
x_{11} & \cdots & x_{1d-1} & y_1 \\
\vdots & \ddots & \vdots & \vdots \\
x_{n1} & \cdots & x_{nd-1} & y_n
\end{bmatrix} = \begin{bmatrix}
X & y
\end{bmatrix}
\]

- \( x_i \in \mathcal{X} \) for \( i = 1, \ldots, n \) are rows of \( X \)
- \( y_i \in \mathcal{Y} \) for \( i = 1, \ldots, n \) are entries of \( y \)
- we believe there is a mapping \( f : \mathcal{X} \rightarrow \mathcal{Y} \)

\[ y_i \approx f(x_i) \]

- our goal is to learn \( f \)
Example: supervised learning for credit decisioning

- goal: decide which credit card applicants should be approved
- input space: entries of $\mathcal{X} \in \mathbb{R}^d$ correspond to fields in credit application
  - e.g., salary, years in residence, outstanding debt, number of credit lines, . . .
- output space: $\mathcal{Y} = \{+1, -1\}$
  - +1 means approve
  - -1 means reject
- data: $D = (x_1, y_1), \ldots, (x_n, y_n)$
  applications of previous customers, and credit approval decisions made by humans

Q: what are potential problems with using a model built with this data?
A: wrong objective: human decision may not be correct decision; covariate shift: future data may look unlike past data; . . .
Example: supervised learning for credit decisioning

- goal: decide which credit card applicants should be approved
- input space: entries of $\mathcal{X} \in \mathbb{R}^d$ correspond to fields in credit application
  - e.g., salary, years in residence, outstanding debt, number of credit lines, ...
- output space: $\mathcal{Y} = \{+1, -1\}$
  - $+1$ means approve
  - $-1$ means reject
- data: $D = (x_1, y_1), \ldots, (x_n, y_n)$ applications of previous customers, and credit approval decisions made by humans

Q: what are potential problems with using a model built with this data?
Example: supervised learning for credit decisioning

► goal: decide which credit card applicants should be approved
► input space: entries of $\mathcal{X} \in \mathbb{R}^d$ correspond to fields in credit application
  ▶ e.g., salary, years in residence, outstanding debt, number of credit lines, . . .
► output space: $\mathcal{Y} = \{+1, -1\}$
  ▶ $+1$ means approve
  ▶ $-1$ means reject
► data: $\mathcal{D} = (x_1, y_1), \ldots, (x_n, y_n)$
  applications of previous customers, and credit approval decisions made by humans

Q: what are potential problems with using a model built with this data?
A: wrong objective: human decision may not be correct decision;
covariate shift: future data may look unlike past data; . . .
Exercise: formalizing real problems

- identify a prediction goal
- identify the input space $\mathcal{X}$
- identify the output space $\mathcal{Y}$
- identify the data $\mathcal{D} = (x_1, y_1), \ldots, (x_n, y_n)$ you’d like to use
- what kinds of noise do you expect in the data?
Outline

Logistics

Stories

Definitions

Kinds of learning

Syllabus
Course objectives (I)

► plot
► predict
► cluster
► impute
► denoise
► recommend
► understand
Course objectives (II)

this course is about

▶ algorithms for big messy data
▶ learning to ask the right questions

at the end of the course, you should have learned

▶ at least one method to solve any problem
▶ machine learning is not magic; it’s math
▶ when not to trust your solution
Course objectives (II)

this course is about

▶ algorithms for big messy data
▶ learning to ask the right questions

at the end of the course, you should have learned

▶ at least one method to solve any problem
▶ machine learning is not magic; it’s math
▶ when not to trust your solution

the rest you can learn online...
ORIE 4741: Learning with Big Messy Data

want to take this class?

▶ ASAP:
  ▶ enroll (or drop) (or get on wait list)
  ▶ fill out course survey
  ▶ sign up for discussion forum
  ▶ sign up for iClicker

▶ Thursday 9/2/2021: homework 0

links on course website:
https://people.orie.cornell.edu/mru8/orie4741/
Questions?

https://docs.google.com/spreadsheets/d/1vLbwi0WCOn0wU6cU_r0RHAiY7C0fDZ1F8Yq09pqYYuk