ORIE 4741: Learning with Big Messy Data

Introduction

Professor Udell
Operations Research and Information Engineering
Cornell

September 3, 2020
Outline

Logistics

Stories

Definitions

Kinds of learning

Syllabus
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 (provides access campuswire Q&A)
  ▶ sign up for iClicker REEF

▶ Thursday 9/10/2020: homework 0

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

- Prof. Madeleine Udell
- TA: Chengrun Yang (ECE PhD)
- TA: Yuxuan Chen (Statistics PhD)
- TA: Anusha Avyukt (Statistics MPS)
- TA: Juliet Zhong (ORIE MEng + Undergraduate)
- TA: Allison Grimsted (ORIE Undergraduate)
- TA: Carrie Rucker (ORIE Undergraduate)
Tech stack

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

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

Zoom contingencies

▶ If I get logged off (eg, due to connectivity issues), your TAs will stay on and provide further instructions
▶ If the Zoom platform fails (eg, Zoom-bombing or Zoom outage), look on Campuswire for further instructions
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 async in any timezone
▶ yes, you can take section async, or not take the section
Questions

during lecture:
  - ask out loud
  - zoom chat to TA Carrie Rucker

outside of lecture:
  - ask at office hours
  - ask on campuswire
  - 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

applied work

▶ finance: Goldman Sachs, BlackRock, Capital One, Schonfeld, Two Sigma, . . .
▶ cybersecurity: DARPA, Expanse (formerly Qadium)
▶ healthcare: Apixio, Ontario
▶ clean energy: Aurora
▶ commerce: Retina.ai, Marketing Attribution
▶ politics: Obama 2012
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, . . .
▶ cybersecurity: DARPA, Expanse (formerly Qadium)
▶ healthcare: Apixio, Ontario
▶ clean energy: Aurora
▶ commerce: Retina.ai, Marketing Attribution
▶ 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>
<th></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>
<td>...</td>
</tr>
<tr>
<td>57</td>
<td>?</td>
<td>NY</td>
<td>$19,000</td>
<td>high school</td>
<td>yes</td>
<td>?</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>
<td>...</td>
</tr>
<tr>
<td>41</td>
<td>F</td>
<td>NV</td>
<td>$23,000</td>
<td>?</td>
<td>yes</td>
<td>Trump</td>
<td>...</td>
</tr>
</tbody>
</table>
## 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?
COVID projections

United States
Current deaths: 183,594 | Projected total deaths: 219,110 (by Nov 1)

Deaths per day

Total deaths

Reproduction number - $R_t$ (estimate)

Population: 331,814,684

https://covid19-projections.com/
COVID projections: Cornell data

https://covid.cornell.edu/testing/dashboard/
(click “participants” and use Zoom reactions to respond to poll)

So far .2% of Cornell population has had COVID-19. I think X% of Cornell population will get COVID this semester, where X is

- (yes) < .5%
- (no) .5 – 1%
- (go slower) 1 – 5%
- (go faster) 5 – 10%
- (coffee) > 10%
Poll

(click “participants” and use Zoom reactions to respond to poll)

I think

▶ (yes) I will get COVID
▶ (no) I will not get COVID
▶ (coffee) I’ve already had COVID
Poll

(click “participants” and use Zoom reactions to respond to poll)

I think a vaccine will be widely available by

▶ (yes) November
▶ (no) January
▶ (slower) Spring 2021
▶ (faster) Summer 2021
▶ (coffee) later
▶ (down) never
Data for projections: simulation

- simulate model given parameter values
- learn parameter values to match data

https://github.com/ORIE4741/demos/blob/master/SIR.ipynb
Simulation: results

by varying parameters, we see

- with infrequent testing (weekly+), pandemic is nearly impossible to control
- people who go to big parties get COVID
  - if too many people go to big parties, Cornell shuts down
  - if moderate (e.g. 10%) of people go to big parties, classes shut down
  - if < 1% go to big parties, Cornell stays open
- if PPE is effective,
  - people who just go to classes are nearly always ok
  - people who go to small parties might get COVID
- if PPE is not effective,
  - people who just go to classes might get COVID
  - people who go to small parties have even odds of getting COVID
Simulation: assumptions

simulation is evocative, not realistic. assumes

- fluid limit: large population, no randomness
- only 3 groups that mix internally
- only 3 modes of contact: parties, classes, external (eg grocery shopping)
- no latency period
- actions don’t depend on infection rates
- quarantine is effective
- ...
Application areas

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

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

---

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”

\[\text{image courtesy of Kim Minor @ IBM}\]
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

---

1 image courtesy of Kim Minor @ IBM
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

---

1Image 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.
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
**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

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

$n$ examples (patients, respondents, households, assets)
$d$ 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
- $a_{i:j}$ 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 $X \in \mathbb{R}^d$ correspond to fields in credit application
  - e.g., salary, years in residence, outstanding debt, number of credit lines, ...
- **output space:** $Y = \{+1, -1\}$
  - +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
**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?
Example: supervised learning for credit decisioning

- **Goal**: decide which credit card applicants should be approved
- **Input space**: entries of $X \in \mathbb{R}^d$ correspond to fields in credit application
  - e.g., salary, years in residence, outstanding debt, number of credit lines, ...
- **Output space**: $Y = \{+1, -1\}$
  - $+1$ 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; ...
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?
Kinds of learning

- **Unsupervised learning:** given $x_1, \ldots, x_n$, learn patterns or structure
- **Supervised learning:** given $(x_1, y_1), \ldots, (x_n, y_n)$, learn $f(x) = y$
- **Online learning:** for $i = 1, \ldots, n$, given $x_i$, predict and observe $y_i$, learn $f(x) = y$
- **Active learning:** for $i = 1, \ldots, n$, choose $x_i$, predict and observe $y_i$, learn $f(x) = y$
- **Reinforcement learning:** for $i = 1, \ldots, n$, choose action $a_i$, move to new state $x_i$, observe reward $r_i$, learn to maximize rewards

**Exercise:**

- what kinds of learning might be effective for autonomous driving? for deciding policy response to pandemic?
- what data is needed?

this class: mostly supervised and unsupervised learning
Kinds of learning

- **Unsupervised learning:**
  given $x_1, \ldots, x_n$, learn patterns or structure

- **Supervised learning:** given $(x_1, y_1), \ldots, (x_n, y_n)$, learn $f(x) = y$

- **Online learning:** for $i = 1, \ldots, n$, given $x_i$, predict and observe $y_i$, learn $f(x) = y$

- **Active learning:** for $i = 1, \ldots, n$, choose $x_i$, predict and observe $y_i$, learn $f(x) = y$

- **Reinforcement learning:** for $i = 1, \ldots, n$, choose action $a_i$, move to new state $x_i$, observe reward $r_i$, learn to maximize rewards

Exercise:
- what kinds of learning might be effective for autonomous driving?
- for deciding policy response to pandemic?
- what data is needed?

This class: mostly supervised and unsupervised learning
Kinds of learning

▶ **Unsupervised learning:**
  given $x_1, \ldots, x_n$, learn patterns or structure

▶ **Supervised learning:**
  given $(x_1, y_1), \ldots, (x_n, y_n)$, learn $f(x) = y$
Kinds of learning

- **Unsupervised learning:**
  given $x_1, \ldots, x_n$, learn patterns or structure

- **Supervised learning:**
  given $(x_1, y_1), \ldots, (x_n, y_n)$, learn $f(x) = y$

- **Online learning:** for $i = 1, \ldots, n$,
  given $x_i$, predict and observe $y_i$, learn $f(x) = y$

- **Active learning:**
  for $i = 1, \ldots, n$,
  choose $x_i$, predict and observe $y_i$, learn $f(x) = y$

- **Reinforcement learning:**
  for $i = 1, \ldots, n$,
  choose action $a_i$, move to new state $x_i$, observe reward $r_i$,
  learn to maximize rewards

Exercise:

- what kinds of learning might be effective for autonomous driving?
- for deciding policy response to pandemic?
- what data is needed?

this class: mostly supervised and unsupervised learning
Kinds of learning

- **Unsupervised learning:**
  given $x_1, \ldots, x_n$, learn patterns or structure

- **Supervised learning:**
  given $(x_1, y_1), \ldots, (x_n, y_n)$, learn $f(x) = y$

- **Online learning:** for $i = 1, \ldots, n$,
  given $x_i$, predict and observe $y_i$, learn $f(x) = y$

- **Active learning:** for $i = 1, \ldots, n$,
  choose $x_i$, predict and observe $y_i$, learn $f(x) = y$
Kinds of learning

- **Unsupervised learning:**
  given $x_1, \ldots, x_n$, learn patterns or structure

- **Supervised learning:**
  given $(x_1, y_1), \ldots, (x_n, y_n)$, learn $f(x) = y$

- **Online learning:** for $i = 1, \ldots, n$,
  given $x_i$, predict and observe $y_i$, learn $f(x) = y$

- **Active learning:** for $i = 1, \ldots, n$,
  choose $x_i$, predict and observe $y_i$, learn $f(x) = y$

- **Reinforcement learning:** for $i = 1, \ldots, n$,
  choose action $a_i$, move to new state $x_i$, observe reward $r_i$,
  learn to maximize rewards

Exercise:
- what kinds of learning might be effective for autonomous driving?
- for deciding policy response to pandemic?

This class: mostly supervised and unsupervised learning
Kinds of learning

▶ **Unsupervised learning:**
given $x_1, \ldots, x_n$, learn patterns or structure

▶ **Supervised learning:**
given $(x_1, y_1), \ldots, (x_n, y_n)$, learn $f(x) = y$

▶ **Online learning:** for $i = 1, \ldots, n$,
given $x_i$, predict and observe $y_i$, learn $f(x) = y$

▶ **Active learning:** for $i = 1, \ldots, n$,
choose $x_i$, predict and observe $y_i$, learn $f(x) = y$

▶ **Reinforcement learning:** for $i = 1, \ldots, n$,
choose action $a_i$, move to new state $x_i$, observe reward $r_i$,
learn to maximize rewards

**Exercise:**

▶ what kinds of learning might be effective for autonomous driving? for deciding policy response to pandemic?
▶ what data is needed?
Kinds of learning

- **Unsupervised learning:**
  given $x_1, \ldots, x_n$, learn patterns or structure

- **Supervised learning:**
  given $(x_1, y_1), \ldots, (x_n, y_n)$, learn $f(x) = y$

- **Online learning:** for $i = 1, \ldots, n$,
  given $x_i$, predict and observe $y_i$, learn $f(x) = y$

- **Active learning:** for $i = 1, \ldots, n$,
  choose $x_i$, predict and observe $y_i$, learn $f(x) = y$

- **Reinforcement learning:** for $i = 1, \ldots, n$,
  choose action $a_i$, move to new state $x_i$, observe reward $r_i$,
  learn to maximize rewards

**Exercise:**

- what kinds of learning might be effective for autonomous driving? for deciding policy response to pandemic?
- what data is needed?

this class: mostly supervised and unsupervised learning
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...
Next steps

▶ **ASAP:**
  ▶ enroll (or drop) (or get on wait list)
  ▶ fill out course survey (provides access to campuswire Q&A)
  ▶ sign up for iClicker REEF

▶ **Thursday 9/10/2020 9:30am:** homework 0

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

https://docs.google.com/spreadsheets/d/1vLbwi0WCOn0wU6cU_r0RHaNY7C0fDZ1F8Yq09pqYYuk