ORIE 4741: Learning with Big Messy Data

Introduction

Professor Udell
Operations Research and Information Engineering
Cornell

September 3, 2019
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

▶ **before next lecture:** post a question or comment to piazza about this lecture

▶ **Thursday 9/5/19:** homework 0

▶ no section on Monday 9/2/19 (Labor Day)

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

- Prof. Madeleine Udell
- TA: Chengrun Yang (ECE PhD)
- TA: Artem Bolshakov (Physics PhD)
- TA: Anthonia Carter (InfoSci PhD)
- TA: Jamie Wong (ORIE MEng + Undergraduate)
- TA: Sijia Fan (ORIE MEng)
Outline

Stories

Definitions

Kinds of learning

Syllabus

Logistics
Oh, you work with big messy data? Maybe you could help us out...?
## 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>Clinton</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>Clinton</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?
Medicine
## 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>
Computers trounce pathologists in predicting lung cancer type, severity

Automating the analysis of slides of lung cancer tissue samples increases the accuracy of tumor classification and patient prognoses, according to a new study.

Computers can be trained to be more accurate than pathologists in assessing slides of lung cancer tissues, according to a new study by researchers at the Stanford University School of Medicine.

The researchers found that a machine-learning approach to identifying critical disease-related features accurately differentiated between two types of lung cancers and predicted patient survival times better than the standard approach of pathologists classifying tumors by grade and stage.

“Pathology as it is practiced now is very subjective,” said Michael Snyder, PhD, professor and chair of the Department of Genetics at Stanford. "Assessing a biopsied slice of tissue to determine the grade and severity of a tumor can be highly subjective, but Stanford researchers found that computers could be trained to make accurate assessments of lung cancer tissue."
**Medicine: radiology**

*Radiology Business*

**AI can help radiologists learn from past, provide better care in future**

Machine learning models can be trained to learn from how radiologists make decisions when interpreting screening mammograms, according ...  
2 weeks ago

*Health Imaging*

**AI boosts accuracy of DBT, slashes radiologists' reading times**

Utilizing an AI system for digital breast tomosynthesis (DBT) can improve radiologists' accuracy while dramatically reducing reading times, ...  
1 month ago

*Radiology Business*

**AI can triage screening mammograms, save radiologists time**

Using deep learning models to triage screening mammograms can improve radiologist specificity without hurting sensitivity, according to new ...  
3 weeks ago

*The Independent*

**AI may be the future of radiology as clinicians struggle to meet demand**
Marketing
<table>
<thead>
<tr>
<th>customer</th>
<th>product 1</th>
<th>product 2</th>
<th>product 3</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>yes</td>
<td>?</td>
<td>yes</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>yes</td>
<td>yes</td>
<td>?</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>?</td>
<td>?</td>
<td>yes</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Finance
<table>
<thead>
<tr>
<th>ticker</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>⋮</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>.05</td>
<td>-.21</td>
<td>⋮</td>
</tr>
<tr>
<td>GOOG</td>
<td>-.11</td>
<td>.24</td>
<td>⋮</td>
</tr>
<tr>
<td>FB</td>
<td>.07</td>
<td>-.18</td>
<td>⋮</td>
</tr>
</tbody>
</table>
Environmental, social and governance (ESG) data

▶ one row for each asset at each time
▶ one column per key performance indicator (KPI)
  ▶ carbon emissions
  ▶ e-waste management
  ▶ climate change risk
  ▶ worker safety
  ▶ ...
▶ values: numerical ratings 1, ..., 10 or boolean \{0, 1\}
▶ triangular missing pattern: KPI/asset coverage increases with time

goals for ESG analysis:
▶ impute missing items?
▶ audit/improve on vendor data quality?
▶ predict long term returns?
Big data in every industry

Data stored per firm by industry

Securities and Investment Services
Communications and Media
Manufacturing
Healthcare
Education
Construction
Consumer and Recreational Services

Firms with >1,000 employees, data in terabytes, 2009

Data volume is growing fast
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
- solar: Aurora
- commerce: Retina.ai
- 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
▶ solar: Aurora
▶ commerce: Retina.ai
▶ politics: Obama 2012
Outline

Stories

Definitions

Kinds of learning

Syllabus

Logistics
Big

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

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”

\[1\text{image courtesy of Kim Minor @ IBM}\]
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:

![Diagram showing 4 Vs: Volume, Velocity, Variety, and Veracity]

$^{1}$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:

<table>
<thead>
<tr>
<th>Volume</th>
<th>Velocity</th>
<th>Variety</th>
<th>Veracity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data at rest</td>
<td>Data in motion</td>
<td>Data in many forms</td>
<td>Data in doubt</td>
</tr>
<tr>
<td>Terabytes to exabytes of existing data to process</td>
<td>Streaming data, milliseconds to seconds to respond</td>
<td>Structured, unstructured, text and multimedia</td>
<td>Uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception and model approximations</td>
</tr>
</tbody>
</table>

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.
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
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, π)
  - 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

(machine learning?)

(human learning?)

(when data is big and messy, machine help is essential for human learning!)

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!
Data table

$n$ examples (patients, respondents, households, assets)
$d$ features (tests, questions, sensors, times)

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

- $a_i$ 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
Outline

Stories

Definitions

Kinds of learning

Syllabus

Logistics
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

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

- \( x_i \in \mathcal{X} \) 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: \( \mathcal{D} = (x_1, y_1), \ldots, (x_n, y_n) \)
  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; ...
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?
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?
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?
- 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?
- 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?
- 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
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 \( f(x) = y \)

**Exercise:**
- what kinds of learning might be effective for autonomous driving?
- 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?
- 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?
▶ 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?
- what data is needed?

this class: mostly supervised and unsupervised learning
Outline

Stories

Definitions

Kinds of learning

Syllabus

Logistics
Course objectives (I)

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

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)

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...
Outline

Stories

Definitions

Kinds of learning

Syllabus

Logistics
This class

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

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

want to take this class?

▶ **ASAP:**
  ▶ enroll (or drop) (or get on wait list)
  ▶ fill out course survey

▶ **before next lecture:** post a question or comment to piazza about this lecture

▶ **due Thursday 9/5/19:** homework 0

▶ no section on Monday 9/2/19 (Labor Day)

...links on course website

https://people.orie.cornell.edu/mru8/orie4741/
Questions?