ORIE 4741: Learning with Big Messy Data

Introduction

Professor Udell
Operations Research and Information Engineering
Cornell

November 11, 2016
Outline

Stories

Definitions

Kinds of learning

Syllabus

Logistics
Oh, you work with big messy data? Maybe you could help us out...?
# Demography

<table>
<thead>
<tr>
<th>age</th>
<th>gender</th>
<th>state</th>
<th>income</th>
<th>education</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>F</td>
<td>CT</td>
<td>$53,000</td>
<td>college</td>
</tr>
<tr>
<td>57</td>
<td>?</td>
<td>NY</td>
<td>$19,000</td>
<td>high school</td>
</tr>
<tr>
<td>?</td>
<td>M</td>
<td>CA</td>
<td>$102,000</td>
<td>masters</td>
</tr>
<tr>
<td>41</td>
<td>F</td>
<td>NV</td>
<td>$23,000</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>gender</td>
<td>heart disease</td>
<td>statins?</td>
<td>...</td>
</tr>
<tr>
<td>-----</td>
<td>--------</td>
<td>---------------</td>
<td>----------</td>
<td>-----</td>
</tr>
<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>
Medicine

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.

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 University School of Medicine. 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.
Pollution

[Snow, 1854]
## Pollution

<table>
<thead>
<tr>
<th>location</th>
<th>time</th>
<th>CO2</th>
<th>O2</th>
<th>O3</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>.7</td>
<td>.9</td>
<td>?</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>.5</td>
<td>.7</td>
<td>?</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>.4</td>
<td>.5</td>
<td>1.4</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Marketing
## 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>(\ldots)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>.05</td>
<td>-.21</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>GOOG</td>
<td>-.11</td>
<td>.24</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>FB</td>
<td>.07</td>
<td>-.18</td>
<td>(\ldots)</td>
</tr>
<tr>
<td></td>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
</tr>
</tbody>
</table>
Hello!

We write emails for all sorts of reasons. For some of those emails, like requesting information or asking a coworker to take action, it's important to write messages that are easy to read, process, and respond to. Until now, there's never been a way to know if your messages are optimized to get a response before sending them.

Today, there is! We are delighted to announce Boomerang Respondable - the first AI assistant that helps you craft perfect emails.
## Data by Volume

<table>
<thead>
<tr>
<th>SECTOR</th>
<th>DATA STORED IN THE U.S., IN PETABYTES (2009)</th>
<th>PETABYTES PER FIRM*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete manufacturing</td>
<td><img src="image" alt="Bar Graph" /></td>
<td>0.94</td>
</tr>
<tr>
<td>Government</td>
<td><img src="image" alt="Bar Graph" /></td>
<td>1.28</td>
</tr>
<tr>
<td>Communications/Media</td>
<td><img src="image" alt="Bar Graph" /></td>
<td>1.75</td>
</tr>
<tr>
<td>Process manufacturing</td>
<td><img src="image" alt="Bar Graph" /></td>
<td>0.81</td>
</tr>
<tr>
<td>Banking</td>
<td><img src="image" alt="Bar Graph" /></td>
<td>1.89</td>
</tr>
<tr>
<td>Health-care providers</td>
<td><img src="image" alt="Bar Graph" /></td>
<td>0.36</td>
</tr>
<tr>
<td>Securities/Investment services</td>
<td><img src="image" alt="Bar Graph" /></td>
<td>3.78</td>
</tr>
<tr>
<td>Professional services</td>
<td><img src="image" alt="Bar Graph" /></td>
<td>0.27</td>
</tr>
<tr>
<td>Retail</td>
<td><img src="image" alt="Bar Graph" /></td>
<td>0.68</td>
</tr>
</tbody>
</table>

*For firms with more than 1,000 employees

Source: McKinsey Global Institute analysis of data from IDC (data stored) and U.S. Dept. of Labor
Outline

Stories

Definitions

Kinds of learning

Syllabus

Logistics
Big

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

\(^1\text{image courtesy of Kim Minor @ IBM}\)
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”

\(^1\)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>

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

1image courtesy of Kim Minor @ IBM
An algorithm for **big data** is one with computational and memory requirements that scale **linearly** (or nearly linearly) in the size of the data.
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
**Big: our definition**

<table>
<thead>
<tr>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>An algorithm for <strong>big data</strong> is one with computational and memory requirements that scale <strong>linearly</strong> (or nearly linearly) in the size of the data.</td>
</tr>
</tbody>
</table>

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!
Data table

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

$$
\begin{bmatrix}
A
\end{bmatrix} =
\begin{bmatrix}
\begin{array}{ccc}
a_{11} & \cdots & a_{1d} \\
\vdots & \ddots & \vdots \\
a_{n1} & \cdots & a_{nd}
\end{array}
\end{bmatrix}
$$

- $a_i$ is $i$th row of $A$: feature vector for $i$th example
- $a_{ij}$ 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

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

- \( x_i \in \mathcal{X} \) 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 card applications

- goal: decide which applicants should be approved for a credit card
- 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: $D = (x_1, y_1), \ldots, (x_n, y_n)$ give credit applications of previous customers, and correct decisions in hindsight
Example: supervised learning for credit card applications

- goal: decide which applicants should be approved for a credit card
- 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: $\mathcal{D} = (x_1, y_1), \ldots, (x_n, y_n)$ give credit applications of previous customers, and correct decisions in hindsight
- noise?
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

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

▶ Unsupervised learning: given \(x_1, \ldots, x_n\), learn patterns or structure

▶ 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 \(x_i\), predict \(y_i\), observe reward \(r_i\), learn \(f(x) = y\)

this class: supervised and unsupervised learning (and possibly some online learning)
Kinds of learning

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

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

- **Unsupervised learning:**
  given \(x_1, \ldots, x_n\), learn patterns or structure
Kinds of learning

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

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

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

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

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

- **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

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

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

- **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 \(x_i\), predict \(y_i\), observe reward \(r_i\), learn \(f(x) = y\)
Kinds of learning

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

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

- **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 \(x_i\), predict \(y_i\), observe reward \(r_i\), learn \(f(x) = y\)

this class: supervised and unsupervised learning (and possibly some online learning)
Outline

Stories

Definitions

Kinds of learning

Syllabus

Logistics
Goals

- plot
- predict
- cluster
- impute
- denoise
- recommend
- understand
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

- ASAP:
  - enroll (contact )
  - tell us who you are at
    http://tinyurl.com/joprtkr
- before next lecture: post a question or comment to piazza about this lecture
- due 8/30/16: homework 0
Questions?