Data Mining and Data Science

Dr. Laura E. Brown
Rekhi 307
CS 1000
10/6/15
My Background
My Background
My Background

Engineering

Pre-Med
My Background

B.S. in Engineering with Concentration in Computer Science
My Background
My Background

M.S.E in Computer Science

• Starting to get into Artificial Intelligence

• Studying AI:
  • Planning
  • Machine Learning
My Background
My Background

Ph.D in Biomedical Informatics

Research Focus:

• Algorithm Development for
  • Learning Bayesian Network
  • Causal Discovery
  • Feature Selection
My Research

• Machine Learning and Artificial Intelligence

• Applications to multiple domains
  • Healthcare

Predict Length of Survival in Oncology Patients
My Research

Model Cross-Architecture Co-tenancy Performance Interference
My Research

• Machine Learning and Artificial Intelligence
• Applications to multiple domains
  • Healthcare
  • Computer Systems
  • Energy Systems

Distributed Management Of Microgrids
My Research

• Machine Learning and Artificial Intelligence
• Applications to multiple domains
  • Healthcare
  • Computer Systems
  • Energy Systems

• Other Projects
  • Data Mining in Mining
  • Recommendation Systems
  • Pediatric Decision Support Tool

Work with ~8 undergraduate and graduate students
My Classes

- Discrete Structures, CS2311
- Artificial Intelligence, CS4811/CS5811
- Data Mining, CS4821
About Me
Data Mining and Data Science

Why? What?
Data Explosion

Growing by a factor of 44

2009
0.8 Zb

2020
35.2 Zettabytes

Real-time data and new kinds of data, coupled with unprecedented processing power, present new and unique challenges

Data Explosion

• Flood of Data
• All types of data:
  • Scientific data: astronomy, biology, medicine, ...
• Examples:
  • Remote sensors on a satellite
  • Telescopes scanning the skies
  • High-throughput gene expression data
  • Scientific simulations


Slide adapted from Tan, Steinbach, Kumar
Data Explosion

• Flood of Data

• All types of data:
  • Scientific data: astronomy, biology, medicine, ...
  • Business transactions, phone calls, texts, ...

• Examples:
  • Web data, click-through, e-commerce
  • Purchases at brick and mortar stores
  • Bank / credit transactions
  • Directed advertisement
Data Explosion

• Flood of Data

• All types of data:
  • Scientific data: astronomy, biology, medicine, ...
  • Business transactions, phone calls, texts, ...
  • Web, text, tweets, images, video, ...

• Examples
  • Emails
  • Tweets
  • Images
  • Videos
What is Data Mining?

• “Data-driven discovery of models and patterns from massive observational data sets”
• “Non-trivial extraction of implicit, previously unknown and potentially useful information from data”
• “Exploration and analysis, by automatic or semi-automatics means, of large quantities of data in order to discover meaningful patterns”
Alternative and Related Names

- Knowledge discovery in databases (KDD)
- Knowledge extraction
- Data / pattern analysis
- Data archeology
- Data dredging
- Information harvesting
- Business intelligence
- Predictive analytics
- Data Science
- ...

What is Data Science?

Domain Expertise

Math and Statistics

Data Science

Hacking Skills
What is Data Science?
What is Data Science?

Data Scientist Skillset

- Statistics
- Machine Learning
- Optimization
- Programming
- CS Fundamentals
- Visualization
- Of The Shelf Toolboxes
- Business and Domain Knowledge
- Communication
- Storytelling
- Big Data
- Cloud Computing
Data Everywhere

• Drowning in data, but **starving for knowledge**
• Data may contain hidden information and patterns
• Human analysis may takes days/weeks/months/never find useful information

• **CAUTION!** Throwing more data at a problem does not always lead to better results
  • Need a good problem/question definition
  • Want data to assist in answering the question
Small World Experiment
Data Science Example
Six Degrees of Separation
Small World Experiment
Data Science Example
Six Degrees of Separation
Small World Experiment
Data Science Example
Small World Experiment

- Problem reported by Stanley Milgram ‘67
  - Selected 300 people in Omaha, Nebraska and Wichita, Kansas
  - Asked them to get a letter to a stock-broker in Boston
    - Passing it through friends
  - 20% of the paths reached the target
  - Mean number of intermediaries = 5.2

- Six degrees of separation
Small World Experiment, 2003

• Redo small world experiment with e-mail [Dodds, Muhamed, Watts, ’03]
  • 18 different targets in 13 different countries
  • Over 60,000 participants, with 24,000 paths
  • Ave. chain length = 4.01
  • Only 384 path completed (1.5%)
  • Correcting for this get typical path length of 7
Small World Experiment, 2008

• Microsoft Messenger instant messages [Leskovec and Horvitz, ‘08]
  • 30 billion conversations between 240 million people
  • Communication graph:
    • 180 million nodes
    • 1.3 billion undirected edges

• Ave. path length is 6.6
Small World Experiment, 2011

• Redo small-world experiment with Facebook [Backstrom, et al. ‘11]
  • Experiment with entire Facebook network of active users
  • ~720 million users, ~69 billion friendship links
• Ave. path length = 4.74
  • 3.74 intermediaries -> “degrees of separation”
6 Degrees of Kevin Bacon

• “Oracle of Bacon”
  • ~1.2 million actors and ~200,000 nicknames
  • Google searchable item:
    • “Elvis Presley bacon number”

• Bacon number – number of edges (on the shortest path) to Kevin Bacon
Academic Collaboration Graph

- Erdös Number – number of edges (on the shortest path to Paul Erdös)
  - Paul Erdös was a mathematician published at least 1.525 publications
  - Authored papers with 504 direct collaborators
An induced subgraph of the collaboration graph with authors of Erdős number ≤ 2.
Erdős numbers are small!
Academic Collaboration Graph

- Famous People Erdös Numbers
  - Albert Einstein – 2
  - Erwin Schrödinger – 8
  - Noam Chomsky – 4
  - John Nash – 4
  - Alan Turing – 5
  - Stephen Hawking – 4
  - Bill Gates – 4
  - Sergey Brin – 3
Other Collaboration Graphs

http://exposedata.com/marvel/
http://exposedata.com/marvel/

Created with the Marvel Universe Social Graph dataset from Infochimps in Gephi for the Data Insight SF competition.
Other Collaboration Graphs

http://i.imgur.com/bBTxU.png
Online Dating
Data Science Examples
The Problem

• Scientific approach to love and marriage through online dating website

• The website does not have user profiles to browse
  • eHarmony computes a compatibility score between two people
  • Uses optimization algorithms to determine users’ best matches
The Company

- Successful at matchmaking
  - Nearly 4% of US marriages in 2012 are result of eHarmony
- Successful business
  - Generated over $1 billion in cumulative revenue
- Started by clinical psychologist who counseled divorcing couples
  - In 1997, conducted a research project interviewing 5000+ couples
- Company went live in 2000
Compatibility Score

- Based on 29 different “dimensions of personality”
  - Character, emotions, values, etc.
- Each user takes a 436 question survey
- Matches must meet >25/29 compatibility areas

Slide adapted from Bertsimas & O’Hair
Matching Problem

- Integer Optimization
- Consider a simple problem 3 people with 3 people with compatibility scores between 1 and 5

Slide adapted from Bertsimas & O’Hair
Matching Problem

- How should the pairs be made to maximize compatibility?
Matching Problem

• Forms an Optimization Problem

\[
\begin{align*}
\text{max} & \quad w_{11}x_{11} + w_{12}x_{12} + w_{13}x_{13} + w_{21}x_{21} + \ldots + w_{33}x_{33} \\
\text{subject to:} & \quad x_{11} + x_{12} + x_{13} = 1 \\
& \quad x_{21} + x_{22} + x_{23} = 1 \\
& \quad x_{31} + x_{32} + x_{33} = 1 \\
& \quad x_{11} + x_{21} + x_{31} = 1 \\
& \quad x_{12} + x_{22} + x_{32} = 1 \\
& \quad x_{13} + x_{23} + x_{33} = 1 \\
\end{align*}
\]

\(x_{11}, x_{21}, x_{31}, x_{12}, x_{22}, x_{32}, x_{13}, x_{23}, x_{33}\) are binary

Make exactly one match
Make exactly one match

Slide adapted from Bertsimas & O’Hair
Successful Approach

• Company continued to grow throughout the 2000s
  • In 2005, 90 members married every day
  • In 2007, 236 members married
  • In 2009, 542 members married
• Maintains 14% of US online dating market
  • Only competitor with larger portion is Match.com, 24%
  • New competitors rising: OkCupid also using a mathematical matching system
    • http://www.wired.com/2014/01/how-to-hack-okcupid/all/
Successful Approach

• Company continued to grow throughout the 2000s

• In 2005, 90 members married every day

• In 2007, 236 members married

• In 2009, 542 members married

• Maintains 14% of US online dating market

• Only competitor with larger portion is Match.com, 24%

• New competitors rising: OkCupid also using a mathematical matching system

http://www.wired.com/2014/01/how-to-hack-okcupid/
AUTOMATING TINDER WITH EIGENFACES

While my friends were getting sucked into "swiping" all day on their phones with Tinder, I eventually got fed up and designed a piece of software that automates everything on Tinder.

An update on Tinderbox: Tinderbox has always been a strictly fun project and I've been glad to share it with the community. From the beginning, I've stated that my support would be very limited. Moving forward, I'd like to openly communicate that I will be dropping any further support for the software. I'm happy that there has been so much interest and support, and moving forward I am putting my energy towards other very promising projects. The code will continue living on Github and feel free to fork it and edit it yourself. Happy Tindering!

What you can do!

Getting Started with Data Mining / Data Science
Future Steps for You

• Find Some Data, or find a problem you want to work on
  • Lots of publicly available data available

## Data Analysis Challenges

### Active Competitions

<table>
<thead>
<tr>
<th>Competitions</th>
<th>Flight Quest 2: Flight Optimization</th>
<th>Belkin Energy Disaggregation Competition</th>
<th>Personalize Expedia Hotel Searches - ICDM 2013</th>
<th>StumbleUpon Evergreen Classification Challenge</th>
<th>See Click Predict Fix</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Competitions</td>
<td>Optimize flight routes based on current weather and traffic.</td>
<td>Disaggregate household energy consumption into individual appliances</td>
<td>Learning to rank hotels to maximize purchases</td>
<td>Build a classifier to categorize webpages as evergreen or non-evergreen</td>
<td>Predict which 311 issues are most important to citizens</td>
</tr>
<tr>
<td></td>
<td>3 months Coming soon $220,000</td>
<td>30 days 105 teams $25,000</td>
<td>35 days 165 teams $25,000</td>
<td>31 days 393 teams $5,000</td>
<td>58 days 14 teams $4,000</td>
</tr>
<tr>
<td><a href="http://www.kaggle.com">http://www.kaggle.com</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Government Data

• National Data
  • US  http://www.data.gov
  • UK  http://data.gov.uk
  • France  http://www.data.gouv.fr
  • United Nations: http://data.un.org
  • Other countries

• Open States: http://openstates.org/

• Congress API: http://sunlightlabs.github.io/congress/

• Cities & States
  • NYC  http://data.cityofnewyork.us/
  • Other Cities and Resources
Data Collectors

• Data Mob datamob.org
• Infochimps Marketplace
  http://www.infochimps.com/marketplace
• AggData  http://www.aggdata.com/
Academic Collections

• UCI Machine Learning
• KDD Nuggets Datasets
• CMU Statlib
• ArXiv Data
• Public Data Sets on AWS
• Stanford Large Network Data

• Health-related:
  • Gene Expression Omnibus NCBI-GEO
  • GenBank NCBI-GenBank
APIs

- Twitter
- PLoS
- Facebook
- Google maps
- SoundCloud
- GitHub
- ...

- Design APIs - apiary
Learn more about Data Analysis

- Data Mining course – CS 4821, spring
- Geospatial Data Mining course – SU 5050, spring
- Michigan Tech’s Data Science Program
Thanks!

Questions?

Laura Brown

lebrown@mtu.edu