1: Introduction

Instructor: Yizhou Sun
yzsun@cs.ucla.edu

January 6, 2019
Course Information


• Class Schedule
  • Slides
  • Announcement
  • Assignments
  • ...

• Prerequisites

• You are expected to have background knowledge in data structures, algorithms, basic linear algebra, and basic statistics.

• You will also need to be familiar with at least one programming language, and have programming experiences.
Meeting Time and Location

• When
  • M&W, 10:00pm-11:50pm

• Where
  • BROAD 2100A
Instructor and TA Information

• Instructor: Yizhou Sun
  • Homepage: http://web.cs.ucla.edu/~yzsun/
  • Email: yzsun@cs.ucla.edu
  • Office: 3531E
  • Office hour: Tuesdays 3-5pm
• TAs:
  • Yunsheng Bai (yba@cs.ucla.edu)
    • office hours: Tuesday 12:30-1:30 and Wednesday 2:30-3:30 @BH 3256S
  • Shengming Zhang (michaelzhang@cs.ucla.edu)
    • office hours: 2-4pm Thursdays @BH 3256S
Grading

• Homework: 25%
• Midterm exam: 25%
• Final exam: 20%
• Course project: 25%
• Participation: 5%
Homework: 25%

- 6 assignments are expected
- Deadline: 11:59pm of the indicated due date via ccle system
  - Late submission policy: get original score $1(t \leq 24)e^{-\frac{\ln(2)}{12}t}$ if you are $t$ hours late.

- No copying or sharing of homework!
  - But you can discuss general challenges and ideas with others
  - Suspicious cases will be reported to The Office of the Dean of Students
Grading: Midterm and Final Exams

- Midterm exam: 25%
- Final exam: 20%
  - Closed book exams, but you can take a “reference sheet” of A4 size
Grading: Course Project

- Course project: 25%
  - Group project (4-5 people for one group)
  - Goal: Solve a given data mining problem
    - Choose among several tasks
    - Crawl data + mine data + present results
  - You are expected to submit a project report and your code at the end of the quarter
Grading: Participation

• Participation (5%)
  • In-class participation
  • Quizzes
  • Online participation (piazza)
    • https://piazza.com/class/jqls8uec97014o
• Recommended: Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*, 3rd edition, Morgan Kaufmann, 2011

• References
  • "Data Mining: The Textbook" by Charu Aggarwal (http://www.charuaggarwal.net/Data-Mining.htm)
  • "Data Mining" by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (http://www-users.cs.umn.edu/~kumar/dmbook/index.php)
  • "Machine Learning" by Tom Mitchell (http://www.cs.cmu.edu/~tom/mlbook.html)
  • "Introduction to Machine Learning" by Ethem ALPAYDIN (http://www.cmpe.boun.edu.tr/~ethem/i2ml/)
  • "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" by Trevor Hastie, Robert Tibshirani, and Jerome Friedman (http://www-stat.stanford.edu/~tibs/ElemStatLearn/)
  • "Pattern Recognition and Machine Learning" by Christopher M. Bishop (http://research.microsoft.com/en-us/um/people/cmbishop/prml/)
Goals of the Course

• Know what data mining is and learn the basic algorithms
• Know how to apply algorithms to real-world applications
• Provide a starting course for research in data mining
1. Introduction

• Why Data Mining?

• What Is Data Mining?

• A Multi-Dimensional View of Data Mining
  • What Kinds of Data Can Be Mined?
  • What Kinds of Patterns Can Be Mined?
  • What Kinds of Technologies Are Used?
  • What Kinds of Applications Are Targeted?

• Content covered by this course
Why Data Mining?

• The Explosive Growth of Data: from terabytes to petabytes
  • Data collection and data availability
    • Automated data collection tools, database systems, Web, computerized society
  • Major sources of abundant data
    • Business: Web, e-commerce, transactions, stocks, ...
    • Science: Remote sensing, bioinformatics, scientific simulation, ...
    • Society and everyone: news, digital cameras, YouTube, social media, mobile devices, ...
  • We are drowning in data, but starving for knowledge!
  • “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets
1. Introduction

• Why Data Mining?
• What Is Data Mining?
• A Multi-Dimensional View of Data Mining
  • What Kinds of Data Can Be Mined?
  • What Kinds of Patterns Can Be Mined?
  • What Kinds of Technologies Are Used?
  • What Kinds of Applications Are Targeted?
• Content covered by this course
What Is Data Mining?

• Data mining (knowledge discovery from data)
  • Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data

• Alternative names
  • Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process
Data Mining in Business Intelligence

Increasing potential to support business decisions

Decision Making

Data Presentation
Visualization Techniques

Data Mining
Information Discovery

Data Exploration
Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems

End User
Business Analyst
Data Analyst
DBA
KDD Process: A Typical View from ML and Statistics

- This is a view from typical machine learning and statistics communities
1. Introduction

• Why Data Mining?

• What Is Data Mining?

• A Multi-Dimensional View of Data Mining
  • What Kinds of Data Can Be Mined?
  • What Kinds of Patterns Can Be Mined?
  • What Kinds of Technologies Are Used?
  • What Kinds of Applications Are Targeted?

• Content covered by this course
Multi-Dimensional View of Data Mining

• **Data to be mined**
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks

• **Knowledge to be mined (or: Data mining functions)**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels

• **Techniques utilized**
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.

• **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.
1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?
- Content covered by this course
## Vector Data

<table>
<thead>
<tr>
<th></th>
<th>Sex</th>
<th>Race</th>
<th>Height</th>
<th>Income</th>
<th>Marital Status</th>
<th>Years of Educ.</th>
<th>Liberalness</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1001</td>
<td>M</td>
<td>1</td>
<td>70</td>
<td>50</td>
<td>1</td>
<td>12</td>
<td>1.73</td>
</tr>
<tr>
<td>R1002</td>
<td>M</td>
<td>2</td>
<td>72</td>
<td>100</td>
<td>2</td>
<td>20</td>
<td>4.53</td>
</tr>
<tr>
<td>R1003</td>
<td>F</td>
<td>1</td>
<td>55</td>
<td>250</td>
<td>1</td>
<td>16</td>
<td>2.99</td>
</tr>
<tr>
<td>R1004</td>
<td>M</td>
<td>2</td>
<td>65</td>
<td>20</td>
<td>2</td>
<td>16</td>
<td>1.13</td>
</tr>
<tr>
<td>R1005</td>
<td>F</td>
<td>1</td>
<td>60</td>
<td>10</td>
<td>3</td>
<td>12</td>
<td>3.81</td>
</tr>
<tr>
<td>R1006</td>
<td>M</td>
<td>1</td>
<td>68</td>
<td>30</td>
<td>1</td>
<td>9</td>
<td>4.76</td>
</tr>
<tr>
<td>R1007</td>
<td>F</td>
<td>5</td>
<td>66</td>
<td>25</td>
<td>2</td>
<td>21</td>
<td>2.01</td>
</tr>
<tr>
<td>R1008</td>
<td>F</td>
<td>4</td>
<td>61</td>
<td>43</td>
<td>1</td>
<td>18</td>
<td>1.27</td>
</tr>
<tr>
<td>R1009</td>
<td>M</td>
<td>1</td>
<td>69</td>
<td>67</td>
<td>1</td>
<td>12</td>
<td>3.25</td>
</tr>
</tbody>
</table>
## Set Data

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diaper, Milk</td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diaper, Milk</td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diaper, Milk</td>
</tr>
</tbody>
</table>
Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities)." –from wiki
Text Data – Topic Modeling

Topics

- gene 0.04
- dna 0.02
- genetic 0.01

- life 0.02
- evolve 0.01
- organism 0.01

- brain 0.04
- neuron 0.02
- nerve 0.01

- data 0.02
- number 0.02
- computer 0.01

Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions...

Topic proportions and assignments


Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
Text Data – Word Embedding

king - man + woman = queen
## Sequence Data

### Syntenic Assemblies for CG15386

<table>
<thead>
<tr>
<th>Synteny</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD106</td>
<td>ATGCTTAGTAATTCCCTACTTTAAGTCCCCTTTTGGCTATTGGCGCTTCCGAGGAAATGGG</td>
</tr>
<tr>
<td>NEWC</td>
<td>ATGCTTAGTAATTCCCTACTTTAAGTCCCCTTTTGGCTATTGGCGCTTCCGAGGAAATGGG</td>
</tr>
<tr>
<td>W501</td>
<td>ATGCTTAGTAATTCCCTACTTTAAGTCCCCTTTTGGCTATTGGCGCTTCCGAGGAAATGGG</td>
</tr>
<tr>
<td>MD199</td>
<td>ATGCTTAGTAATTCCCTACTTTAAGTCCCCTTTTGGCTATTGGCGCTTCCGAGGAAATGGG</td>
</tr>
<tr>
<td>C1674</td>
<td>ATGCTTAGTAATTCCCTACTTTAAGTCCCCTTTTGGCTATTGGCGCTTCCGAGGAAATGGG</td>
</tr>
<tr>
<td>SIM4</td>
<td>ATGCTTAGTAATTCCCTACTTTAAGTCCCCTTTTGGCTATTGGCGCTTCCGAGGAAATGGG</td>
</tr>
<tr>
<td>MD106</td>
<td>CTACGGCCCTAATGGTGCTAAAGAGCCGAACGCAGCTCGACAAAAATAGACGCATCAAAAGCTT</td>
</tr>
<tr>
<td>NEWC</td>
<td>CTACGGCCCTAATGGTGCTAAAGAGCCGAACGCAGCTCGACAAAAATAGACGCATCAAAAGCTT</td>
</tr>
<tr>
<td>W501</td>
<td>CTACGGCCCTAATGGTGCTAAAGAGCCGAACGCAGCTCGACAAAAATAGACGCATCAAAAGCTT</td>
</tr>
<tr>
<td>MD199</td>
<td>CTACGGCCCTAATGGTGCTAAAGAGCCGAACGCAGCTCGACAAAAATAGACGCATCAAAAGCTT</td>
</tr>
<tr>
<td>C1674</td>
<td>CTACGGCCCTAATGGTGCTAAAGAGCCGAACGCAGCTCGACAAAAATAGACGCATCAAAAGCTT</td>
</tr>
<tr>
<td>SIM4</td>
<td>CTACGGCCCTAATGGTGCTAAAGAGCCGAACGCAGCTCGACAAAAATAGACGCATCAAAAGCTT</td>
</tr>
<tr>
<td>MD106</td>
<td>CCGTTTTCAAGTACCAAAATGATGCGGATGACGAGCGAAAGGCCTGCTTGTATAAGGAAG</td>
</tr>
<tr>
<td>NEWC</td>
<td>CCGTTTTCAAGTACCAAAATGATGCGGATGACGAGCGAAAGGCCTGCTTGTATAAGGAAG</td>
</tr>
<tr>
<td>W501</td>
<td>CCGTTTTCAAGTACCAAAATGATGCGGATGACGAGCGAAAGGCCTGCTTGTATAAGGAAG</td>
</tr>
<tr>
<td>MD199</td>
<td>CCGTTTTCAAGTACCAAAATGATGCGGATGACGAGCGAAAGGCCTGCTTGTATAAGGAAG</td>
</tr>
<tr>
<td>C1674</td>
<td>CCGTTTTCAAGTACCAAAATGATGCGGATGACGAGCGAAAGGCCTGCTTGTATAAGGAAG</td>
</tr>
<tr>
<td>SIM4</td>
<td>CCGTTTTCAAGTACCAAAATGATGCGGATGACGAGCGAAAGGCCTGCTTGTATAAGGAAG</td>
</tr>
<tr>
<td>MD106</td>
<td>CTGCAGGAGGCCTGACCCACCAAGGTGCCCACATTCTACAGGTCAGCGCCGAGAAATAG</td>
</tr>
<tr>
<td>NEWC</td>
<td>CTGCAGGAGGCCTGACCCACCAAGGTGCCCACATTCTACAGGTCAGCGCCGAGAAATAG</td>
</tr>
<tr>
<td>W501</td>
<td>CTGCAGGAGGCCTGACCCACCAAGGTGCCCACATTCTACAGGTCAGCGCCGAGAAATAG</td>
</tr>
<tr>
<td>MD199</td>
<td>CTGCAGGAGGCCTGACCCACCAAGGTGCCCACATTCTACAGGTCAGCGCCGAGAAATAG</td>
</tr>
<tr>
<td>C1674</td>
<td>CTGCAGGAGGCCTGACCCACCAAGGTGCCCACATTCTACAGGTCAGCGCCGAGAAATAG</td>
</tr>
<tr>
<td>SIM4</td>
<td>CTGCAGGAGGCCTGACCCACCAAGGTGCCCACATTCTACAGGTCAGCGCCGAGAAATAG</td>
</tr>
</tbody>
</table>
Sequence Data – Seq2Seq

Encoder

Embed

He loved to eat

Decoder

NULL Er liebte zu essen

Er liebte zu essen
Time Series

Weekly U.S. Retail Gasoline Prices, Regular Grade

Source: Energy Information Administration
Graph / Network
Graph / Network
– Community Detection
<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
Image Data – Neural Style Transfer
Image Data – Image Captioning

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"girl in pink dress is jumping in air."

"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."
1. Introduction

• Why Data Mining?

• What Is Data Mining?

• A Multi-Dimensional View of Data Mining
  • What Kinds of Data Can Be Mined?
  • What Kinds of Patterns Can Be Mined?
  • What Kinds of Technologies Are Used?
  • What Kinds of Applications Are Targeted?

• Content covered by this course
Data Mining Function: Association and Correlation Analysis

• Frequent patterns (or frequent itemsets)
  • What items are frequently purchased together in your Amazon transactions?

Frequently bought together

• Association, correlation vs. causality
  • A typical association rule
    • Diaper $\rightarrow$ Beer [0.5%, 75%] (support, confidence)
Data Mining Function: Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, ...
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, ...
# Image Classification Example

<table>
<thead>
<tr>
<th>mite</th>
<th>container ship</th>
<th>motor scooter</th>
<th>leopard</th>
</tr>
</thead>
<tbody>
<tr>
<td>mite</td>
<td>container ship</td>
<td>motor scooter</td>
<td>leopard</td>
</tr>
<tr>
<td>black widow</td>
<td>lifeboat</td>
<td>go-kart</td>
<td>jaguar</td>
</tr>
<tr>
<td>cockroach</td>
<td>amphibian</td>
<td>moped</td>
<td>cheetah</td>
</tr>
<tr>
<td>tick</td>
<td>fireboat</td>
<td>bumper car</td>
<td>snow leopard</td>
</tr>
<tr>
<td>starfish</td>
<td>drilling platform</td>
<td>golfcart</td>
<td>Egyptian cat</td>
</tr>
<tr>
<td>grille</td>
<td>mushroom</td>
<td>cherry</td>
<td></td>
</tr>
<tr>
<td>convertible</td>
<td>agaric</td>
<td>dalmation</td>
<td>squirrel monkey</td>
</tr>
<tr>
<td>grille</td>
<td>mushroom</td>
<td>grape</td>
<td>spider monkey</td>
</tr>
<tr>
<td>pickup</td>
<td>jelly fungus</td>
<td>elderberry</td>
<td>titi</td>
</tr>
<tr>
<td>beach wagon</td>
<td>gill fungus</td>
<td>currant</td>
<td>indri</td>
</tr>
<tr>
<td>fire engine</td>
<td>dead-man’s-fingers</td>
<td></td>
<td>howler monkey</td>
</tr>
</tbody>
</table>
Data Mining Function: Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications
Clustering Example
Data Mining Functions: Others

- Prediction
- Similarity search
- Ranking
- Outlier detection
- ...

1. Introduction

• Why Data Mining?

• What Is Data Mining?

• A Multi-Dimensional View of Data Mining
  • What Kinds of Data Can Be Mined?
  • What Kinds of Patterns Can Be Mined?
  • What Kinds of Technologies Are Used?
  • What Kinds of Applications Are Targeted?

• Content covered by this course
Data Mining: Confluence of Multiple Disciplines

- Machine Learning
- Pattern Recognition
- Statistics
- Applications
- Algorithm
- Database Technology
- Visualization
- High-Performance Computing
1. Introduction

• Why Data Mining?

• What Is Data Mining?

• A Multi-Dimensional View of Data Mining
  • What Kinds of Data Can Be Mined?
  • What Kinds of Patterns Can Be Mined?
  • What Kinds of Technologies Are Used?
  • What Kinds of Applications Are Targeted?

• Content covered by this course
Applications of Data Mining

• Web page analysis: from web page classification, clustering to PageRank & HITS algorithms
• Collaborative analysis & recommender systems
• Basket data analysis to targeted marketing
• Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
• Data mining and software engineering (e.g., IEEE Computer, Aug. 2009 issue)
• Social media
• Game
Google Flu Trends

- https://www.youtube.com/watch?v=6111nS66Dpk
NetFlix Prize

- [https://www.youtube.com/watch?v=4_e2sNYYfxA](https://www.youtube.com/watch?v=4_e2sNYYfxA)

## Leaderboard

Showing Test Score. [Click here to show quiz score]

Display top ▼ leaders.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team Name</th>
<th>Best Test Score</th>
<th>% Improvement</th>
<th>Best Submit Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BellKor's Pragmatic Chaos</td>
<td>0.8567</td>
<td>10.06</td>
<td>2009-07-26 18:18:28</td>
</tr>
<tr>
<td>2</td>
<td>The Ensemble</td>
<td>0.8567</td>
<td>10.06</td>
<td>2009-07-26 18:38:22</td>
</tr>
<tr>
<td>3</td>
<td>Grand Prize Team</td>
<td>0.8582</td>
<td>9.90</td>
<td>2009-07-10 21:24:40</td>
</tr>
<tr>
<td>4</td>
<td>Opera Solutions and Vandelay United</td>
<td>0.8588</td>
<td>9.84</td>
<td>2009-07-10 01:12:31</td>
</tr>
<tr>
<td>5</td>
<td>Vandelay Industries</td>
<td>0.8591</td>
<td>9.81</td>
<td>2009-07-10 00:32:20</td>
</tr>
<tr>
<td>6</td>
<td>PragmaticTheory</td>
<td>0.8594</td>
<td>9.77</td>
<td>2009-06-24 12:06:56</td>
</tr>
<tr>
<td>7</td>
<td>BellKor in BigChaos</td>
<td>0.8601</td>
<td>9.70</td>
<td>2009-05-13 08:14:09</td>
</tr>
<tr>
<td>8</td>
<td>Dace</td>
<td>0.8612</td>
<td>9.59</td>
<td>2009-07-24 17:18:43</td>
</tr>
<tr>
<td>9</td>
<td>Feeds2</td>
<td>0.8622</td>
<td>9.48</td>
<td>2009-07-12 13:11:51</td>
</tr>
<tr>
<td>10</td>
<td>BigChaos</td>
<td>0.8623</td>
<td>9.47</td>
<td>2009-04-07 12:33:59</td>
</tr>
<tr>
<td>11</td>
<td>Opera Solutions</td>
<td>0.8623</td>
<td>9.47</td>
<td>2009-07-24 00:34:07</td>
</tr>
<tr>
<td>12</td>
<td>BellKor</td>
<td>0.8624</td>
<td>9.46</td>
<td>2009-07-26 17:19:11</td>
</tr>
</tbody>
</table>
Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski\textsuperscript{a,1}, David Stillwell\textsuperscript{a}, and Thore Graepel\textsuperscript{b}

\textsuperscript{a}Free School Lane, The Psychometrics Centre, University of Cambridge, Cambridge CB2 3RQ United Kingdom; and \textsuperscript{b}Microsoft Research, Cambridge CB1 2FB, United Kingdom

Edited by Kenneth Wachter, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 29, 2012)

We show that easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. The analysis presented is based on a dataset of over 58,000 volunteers who provided their Facebook Likes, detailed demographic profiles, and the results of several psychometric tests. The proposed model uses dimensionality reduction for preprocessing the Likes data, which are then entered into logistic/linear regression to predict individual psychodemographic profiles from Likes. The model correctly discriminates between homosexual and heterosexual men in 88% of cases, African Americans and Caucasian Americans in 95% of cases, and between Democrat and Republican in 85% of cases. For the personality trait “Openness,” prediction accuracy is close to the test–retest accuracy of a standard personality test. We give examples of associations between attributes and Likes and discuss implications for online personalization browsing logs (11–15). Similarly, it has been shown that personality can be predicted based on the contents of personal Web sites (16), music collections (17), properties of Facebook or Twitter profiles such as the number of friends or the density of friendship networks (18–21), or language used by their users (22). Furthermore, location within a friendship network at Facebook was shown to be predictive of sexual orientation (23).

This study demonstrates the degree to which relatively basic digital records of human behavior can be used to automatically and accurately estimate a wide range of personal attributes that people would typically assume to be private. The study is based on Facebook Likes, a mechanism used by Facebook users to express their positive association with (or “Like”) online content, such as photos, friends’ status updates, Facebook pages of products, sports, musicians, books, restaurants, or popular Web sites. Likes represent a very generic class of digital records, similar to Web search queries, Web browsing histories, and credit card purchases. For example, observing users’ Likes related to music
1. Introduction

• Why Data Mining?

• What Is Data Mining?

• A Multi-Dimensional View of Data Mining
  • What Kinds of Data Can Be Mined?
  • What Kinds of Patterns Can Be Mined?
  • What Kinds of Technologies Are Used?
  • What Kinds of Applications Are Targeted?

• Content covered by this course
Course Content

• Functions to be covered
  • Prediction and classification
  • Clustering
  • Frequent pattern mining and association rules
  • Similarity search

• Data types to be covered
  • Vector data
  • Set data
  • Sequential data
  • Text data
# Methods to Learn

<table>
<thead>
<tr>
<th></th>
<th>Vector Data</th>
<th>Set Data</th>
<th>Sequence Data</th>
<th>Text Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classification</strong></td>
<td>Logistic Regression; Decision Tree; KNN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SVM; NN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Clustering</strong></td>
<td>K-means; hierarchical clustering; DBSCAN;</td>
<td></td>
<td></td>
<td>PLSA</td>
</tr>
<tr>
<td></td>
<td>DBSCAN; Mixture Models</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Prediction</strong></td>
<td>Linear Regression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GLM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Frequent Pattern</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mining</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Apriori; FP growth</td>
<td></td>
<td>GSP; PrefixSpan</td>
<td></td>
</tr>
<tr>
<td><strong>Similarity Search</strong></td>
<td></td>
<td></td>
<td></td>
<td>DTW</td>
</tr>
</tbody>
</table>
Where to Find References? DBLP, CiteSeer, Google

• Data mining and KDD (SIGKDD: CDROM)
  • Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  • Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD

• Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
  • Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA

• AI & Machine Learning
  • Conferences: ICML, AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  • Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.

• Web and IR
  • Conferences: SIGIR, WWW, WSDM, CIKM, etc.
  • Journals: WWW: Internet and Web Information Systems,

• Statistics
  • Conferences: Joint Stat. Meeting, etc.
  • Journals: Annals of statistics, etc.

• Visualization
  • Conference proceedings: CHI, ACM-SIGGraph, etc.
  • Journals: IEEE Trans. visualization and computer graphics, etc.
Recommended Reference Books

- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011
- B. Liu, Web Data Mining, Springer 2006
- Y. Sun and J. Han, Mining Heterogeneous Information Networks, Morgan & Claypool, 2012
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005
- S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998
Major Concepts Related to Probability and Statistics

- Elements of Probability
  - Sample space, event space, probability measure
  - Conditional probability
  - Independence, conditional independence

- Random variables
  - Cumulative distribution function, Probability mass function (for discrete random variable), Probability density function (for continuous random variable)
  - Expectation, variance
  - Some frequently used distributions
    - Discrete: Bernoulli, binomial, geometric, passion
    - Continuous: uniform, exponential, normal

- More random variables
  - Joint distribution, marginal distribution, joint and marginal probability mass function, joint and marginal density function
  - Chain rule
  - Bayes’ rule
  - Independence
  - Expectation, conditional expectation, and covariance
Major Concepts in Linear Algebra

• Vectors
  • Addition, scalar multiplication, norm, dot product (inner product), projection, cosine similarity

• Matrices
  • Addition, scalar multiplication, matrix-matrix multiplication, trace, eigenvalues and eigenvectors
Optimization Related

- MLE and MAP Principle
- Gradient descent / stochastic gradient descent
- Newton’s method
- Expectation-Maximum algorithm (EM)
Other Courses

- **CS247: Advanced Data Mining**
  - Focus on Text, Recommender Systems, and Networks/Graphs
  - Will be offered in Spring 2019

- **CS249: Probabilistic Models for Structured Data**
  - Focus on Probabilistic Models on text and graph data
  - Are offered in Winter 2019