1: Introduction

Instructor: Yizhou Sun
yzsun@ccs.neu.edu

September 9, 2013
Course Information

- Course homepage: [http://www.ccs.neu.edu/home/yzsun/classes/2013Fall_CS6220/index.htm](http://www.ccs.neu.edu/home/yzsun/classes/2013Fall_CS6220/index.htm)
  - Class schedule
  - Slides
  - Announcement
  - Assignments
  - …

- Piazza: [https://piazza.com/northeastern/fall2013/cs6220/home](https://piazza.com/northeastern/fall2013/cs6220/home)
• Prerequisites
  • **CS 5800 or CS 7800, or consent of instructor**
  • **More generally**
    • 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
  • Tuesdays, 6-9pm
• Where
  • Behrakis Health Sciences Center 310
Instructor and TA Information

• Instructor: Yizhou Sun
  • Homepage: http://www.ccs.neu.edu/home/yzsun/
  • Email: yzsun@ccs.neu.edu
  • Office: 320 WVH
  • Office hour: Wednesdays 3-5pm

• TA:
  • Moonyoung (Moon) Kang
    • Email: yerihyo@gmail.com
    • Office hours: Tuesdays 2-4pm at 472 WVH
  • Qizhen Ruan
    • Email: ruan.qi@husky.neu.edu
    • Office hours: Mondays 4:30-6:30pm at 102 Main Lab WVH
Grading

• Homework: 40%
• Midterm exam: 25%
• Course project: 30%
• Participation: 5%
Grading: Homework

• Homework: 40%
  • Four assignments are expected
    • 2 paper-based assignments
    • 2 program-based assignments
  • Deadline: 11:59pm of the indicated due date
    via Blackboard or class system
      • within 1 hour late: 90% max; within 8 hours late: 60% max;
        otherwise: 0%
  • No copying or sharing of homework!
    • But you can discuss general challenges and ideas with others
Grading: Midterm Exam

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

- Course project: 30%
  - Group project (3-4 people for one group)
  - Goal: Choose one interesting problem, formalize it as a data mining task, collect data, provide solutions, and evaluate and compare your solutions.
  - You are expected to submit one project proposal early this semester, and your datasets, code, and a project report at the end of the semester
  - You are expected to present your project at the end of the semester.
Grading: Participation

• Participation (5%)
  • In-class participation
  • Online participation (piazza)
Textbook

- Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*, 3rd edition, Morgan Kaufmann, 2011
- References
  - "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/)
Course Content

• By data types:
  • matrix data
  • set data
  • sequence data
  • time series
  • graph and network (Next Semester: Advanced Topics)

• By functions:
  • Classification
  • Clustering
  • Frequent pattern mining
  • Prediction
  • Similarity search
  • Ranking
Goal of the Course

- Know what is data mining and 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?

• Major Issues in Data Mining
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
  • We are drowning in data, but starving for knowledge!
  • “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets
Big Data Challenges

• Video 1: Big Data Challenges (Ads by DataStax)
  • [http://www.youtube.com/watch?v=or6Pse8fxD4](http://www.youtube.com/watch?v=or6Pse8fxD4)

• Video 2: Explaining Big Data
  • [http://www.youtube.com/watch?v=7D1CQ_LOizA](http://www.youtube.com/watch?v=7D1CQ_LOizA)
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?

• Major Issues in Data Mining

• A Brief History of Data Mining and Data Mining Society

• Summary
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

End User

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

DBA

Data Preprocessing/Integration, Data Warehouses

Data Exploration
- Statistical Summary, Querying, and Reporting

Data Analyst

Data Mining
- Information Discovery

Business Analyst

Data Presentation
- Visualization Techniques

Decision Making
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?
- Major Issues in Data Mining
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?

• Major Issues in Data Mining
Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web
# Matrix 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>
## Sequence Data

### Syntenic Assemblies for CG15386

<table>
<thead>
<tr>
<th>Syndrome</th>
<th>Sequence 1</th>
<th>Sequence 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD106</td>
<td>ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAAATGGG</td>
<td>CTACGGCCCTAATTGGGTCTAACAAGACCGCAACCTCGACAATATTAGACGCATCAAAGCCT</td>
</tr>
<tr>
<td>NEWC</td>
<td>ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAAATGGG</td>
<td>CTACGGCCCTAATTGGGTCTAACAAGACCGCAACCTCGACAATATTAGACGCATCAAAGCCT</td>
</tr>
<tr>
<td>W501</td>
<td>ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAAATGGG</td>
<td>CTACGGCCCTAATTGGGTCTAACAAGACCGCAACCTCGACAATATTAGACGCATCAAAGCCT</td>
</tr>
<tr>
<td>MD199</td>
<td>ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAAATGGG</td>
<td>CTACGGCCCTAATTGGGTCTAACAAGACCGCAACCTCGACAATATTAGACGCATCAAAGCCT</td>
</tr>
<tr>
<td>C1674</td>
<td>ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAAATGGG</td>
<td>CTACGGCCCTAATTGGGTCTAACAAGACCGCAACCTCGACAATATTAGACGCATCAAAGCCT</td>
</tr>
<tr>
<td>SIM4</td>
<td>ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAAATGGG</td>
<td>CTACGGCCCTAATTGGGTCTAACAAGACCGCAACCTCGACAATATTAGACGCATCAAAGCCT</td>
</tr>
<tr>
<td>MD106</td>
<td>CCGTTTAAGTACCAAACTGAAGCGGATGAGCAGCAAAAAGGTCTCTGTTTATGAAAGAG</td>
<td>CCGTTTCAAGTACCAAACTGAAGCGGATGAGCAGCAAAAAGGTCTCTGTTTATGAAAGAG</td>
</tr>
<tr>
<td>NEWC</td>
<td>CCGTTTCAAGTACCAAACTGAAGCGGATGAGCAGCAAAAAGGTCTCTGTTTATGAAAGAG</td>
<td>CCGTTTCAAGTACCAAACTGAAGCGGATGAGCAGCAAAAAGGTCTCTGTTTATGAAAGAG</td>
</tr>
<tr>
<td>W501</td>
<td>CCGTTTCAAGTACCAAACTGAAGCGGATGAGCAGCAAAAAGGTCTCTGTTTATGAAAGAG</td>
<td>CCGTTTCAAGTACCAAACTGAAGCGGATGAGCAGCAAAAAGGTCTCTGTTTATGAAAGAG</td>
</tr>
<tr>
<td>MD199</td>
<td>CCGTTTCAAGTACCAAACTGAAGCGGATGAGCAGCAAAAAGGTCTCTGTTTATGAAAGAG</td>
<td>CCGTTTCAAGTACCAAACTGAAGCGGATGAGCAGCAAAAAGGTCTCTGTTTATGAAAGAG</td>
</tr>
<tr>
<td>C1674</td>
<td>CCGTTTCAAGTACCAAACTGAAGCGGATGAGCAGCAAAAAGGTCTCTGTTTATGAAAGAG</td>
<td>CCGTTTCAAGTACCAAACTGAAGCGGATGAGCAGCAAAAAGGTCTCTGTTTATGAAAGAG</td>
</tr>
<tr>
<td>SIM4</td>
<td>CCGTTTCAAGTACCAAACTGAAGCGGATGAGCAGCAAAAAGGTCTCTGTTTATGAAAGAG</td>
<td>CCGTTTCAAGTACCAAACTGAAGCGGATGAGCAGCAAAAAGGTCTCTGTTTATGAAAGAG</td>
</tr>
<tr>
<td>MD106</td>
<td>CTGCAAGGAGGCGTCCACACCAGGCGGCTGCCCCAACTGACTACAGGTGATCCGGCGAGAATAG</td>
<td>CTGCAAGGAGGCGTCCACACCAGGCGGCTGCCCCAACTGACTACAGGTGATCCGGCGAGAATAG</td>
</tr>
<tr>
<td>NEWC</td>
<td>CTGCAAGGAGGCGTCCACACCAGGCGGCTGCCCCAACTGACTACAGGTGATCCGGCGAGAATAG</td>
<td>CTGCAAGGAGGCGTCCACACCAGGCGGCTGCCCCAACTGACTACAGGTGATCCGGCGAGAATAG</td>
</tr>
<tr>
<td>W501</td>
<td>CTGCAAGGAGGCGTCCACACCAGGCGGCTGCCCCAACTGACTACAGGTGATCCGGCGAGAATAG</td>
<td>CTGCAAGGAGGCGTCCACACCAGGCGGCTGCCCCAACTGACTACAGGTGATCCGGCGAGAATAG</td>
</tr>
<tr>
<td>MD199</td>
<td>CTGCAAGGAGGCGTCCACACCAGGCGGCTGCCCCAACTGACTACAGGTGATCCGGCGAGAATAG</td>
<td>CTGCAAGGAGGCGTCCACACCAGGCGGCTGCCCCAACTGACTACAGGTGATCCGGCGAGAATAG</td>
</tr>
<tr>
<td>C1674</td>
<td>CTGCAAGGAGGCGTCCACACCAGGCGGCTGCCCCAACTGACTACAGGTGATCCGGCGAGAATAG</td>
<td>CTGCAAGGAGGCGTCCACACCAGGCGGCTGCCCCAACTGACTACAGGTGATCCGGCGAGAATAG</td>
</tr>
<tr>
<td>SIM4</td>
<td>CTGCAAGGAGGCGTCCACACCAGGCGGCTGCCCCAACTGACTACAGGTGATCCGGCGAGAATAG</td>
<td>CTGCAAGGAGGCGTCCACACCAGGCGGCTGCCCCAACTGACTACAGGTGATCCGGCGAGAATAG</td>
</tr>
</tbody>
</table>
Time Series

Weekly U.S. Retail Gasoline Prices, Regular Grade

Source: Energy Information Administration
Graph / Network
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?
• Major Issues in Data Mining
Data Mining Function: Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper → Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?
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, ...
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
Data Mining Function: Others

- Prediction
- Similarity search
- Ranking
- Outlier detection
- ...
Evaluation of Knowledge

• Are all mined knowledge interesting?
  • One can mine tremendous amount of “patterns” and knowledge
  • Some may fit only certain dimension space (time, location, ...)
  • Some may not be representative, may be transient, ...

• Evaluation of mined knowledge → directly mine only interesting knowledge?
  • Descriptive vs. predictive
  • Coverage
  • Typicality vs. novelty
  • Accuracy
  • Timeliness
  • ...
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?
- Major Issues in Data Mining
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?
- Major Issues in Data Mining
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
Example

- Street Bump Boston Project
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?

• Major Issues in Data Mining
Major Issues in Data Mining (1)

- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Data mining: An interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern- or constraint-guided mining
- User Interaction
  - Interactive mining
  - Incorporation of background knowledge
  - Presentation and visualization of data mining results
Major Issues in Data Mining (2)

• Diversity of data types
  • Handling complex types of data
  • Mining dynamic, networked, and global data repositories
• Efficiency and Scalability
  • Efficiency and scalability of data mining algorithms
  • Parallel, distributed, stream, and incremental mining methods
• Data mining and society
  • Social impacts of data mining
  • Privacy-preserving data mining
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: Machine learning (ML), 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, 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