1: Introduction

Instructor: Yizhou Sun
yzsun@ccs.neu.edu

September 8, 2014
Course Information

• Course homepage:
  http://www.ccs.neu.edu/home/yzsun/classes/2014Fall_CS6220/index.htm
  • Class schedule
  • Slides
  • Announcement
  • Assignments
  • …
• 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
  • Monday, 6-9pm

• Where
  • Shillman Hall 335
Instructor and TA Information

- **Instructor:** Yizhou Sun
  - **Homepage:** [http://www.ccs.neu.edu/home/yzsun/](http://www.ccs.neu.edu/home/yzsun/)
  - **Email:** yzsun@ccs.neu.edu
  - **Office:** 320 WVH
  - **Office hour:** Wednesdays 1-3pm

- **TA:**
  - **Yupeng Gu**
    - **Email:** ypgu@ccs.neu.edu
    - **Office hours:** Tuesdays 2:30-4:30pm at 472 WVH
  - **Kosha Shah**
    - **Email:** shah.ko@husky.neu.edu
    - **Office hours:** Thursdays 10:00am-12:00pm at 102 Main Lab WVH
Grading

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

- Homework: 40%
  - Four assignments are expected
  - Deadline: 11:59pm of the indicated due date via Blackboard or class system
    - No Late Submission!
  - 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: Compete on the assigned course project
  • You are expected to submit a project report and your code at the end of the semester
Grading: Participation

• Participation (5%)
  • In-class participation
  • quizzes
  • Online participation (piazza)
    • piazza.com/northeastern/fall2014/cs6220
Textbook

- Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*, 3rd edition, Morgan Kaufmann, 2011
- References
  - "Machine Learning" by Tom Mitchell ([http://www.cs.cmu.edu/~tom/mlbook.html](http://www.cs.cmu.edu/~tom/mlbook.html))
  - "Introduction to Machine Learning" by Ethem ALPAYDIN ([http://www.cmpe.boun.edu.tr/~ethem/i2ml/](http://www.cmpe.boun.edu.tr/~ethem/i2ml/))
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?
- 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
- 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
Increasing potential to support business decisions

Data Sources

Data Preprocessing/Integration, Data Warehouses

Data Exploration

Statistical Summary, Querying, and Reporting

Data Mining

Information Discovery

Data Presentation

Visualization Techniques

Decision Making

End User

Business Analyst

Data Analyst

DBA

Paper, Files, Web documents, Scientific experiments, Database Systems
KDD Process: A Typical View from ML and Statistics

Input Data → Data Pre-Processing → Data Mining → Post-Processing

- Data integration
- Normalization
- Feature selection
- Dimension reduction

- Pattern discovery
- Association & correlation
- Classification
- Clustering
- Outlier analysis
- … … … …

- Pattern evaluation
- Pattern selection
- Pattern interpretation
- Pattern visualization

- 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
## 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>
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?
- 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 Walmart?
- Association, correlation vs. causality
- A typical association rule
  - Diaper $\rightarrow$ 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 Functions: Others

• Prediction
• Similarity search
• Ranking
• Outlier detection
• ...

32
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?

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

Machine Learning
Pattern Recognition
Statistics
Applications
Visualization
Algorithm
Database Technology
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
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

• By data types:
  • matrix data
  • set data
  • sequence data
  • time series
  • graph and network

• By functions:
  • Classification
  • Clustering
  • Frequent pattern mining
  • Prediction
  • Similarity search
  • Ranking
<table>
<thead>
<tr>
<th>Methods to Learn</th>
<th>Matrix Data</th>
<th>Set Data</th>
<th>Sequence Data</th>
<th>Time Series</th>
<th>Graph &amp; Network</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classification</strong></td>
<td>Decision Tree; Naïve Bayes; Logistic Regression SVM; kNN</td>
<td></td>
<td>HMM</td>
<td></td>
<td>Label Propagation</td>
</tr>
<tr>
<td><strong>Clustering</strong></td>
<td>K-means; hierarchical clustering; DBSCAN; Mixture Models; kernel k-means</td>
<td></td>
<td></td>
<td></td>
<td>SCAN; Spectral Clustering</td>
</tr>
<tr>
<td><strong>Frequent Pattern Mining</strong></td>
<td></td>
<td>Apriori; FP-growth</td>
<td>GSP; PrefixSpan</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Prediction</strong></td>
<td>Linear Regression</td>
<td></td>
<td></td>
<td>Autoregression</td>
<td></td>
</tr>
<tr>
<td><strong>Similarity Search</strong></td>
<td></td>
<td></td>
<td>DTW</td>
<td>P-PageRank</td>
<td></td>
</tr>
<tr>
<td><strong>Ranking</strong></td>
<td></td>
<td></td>
<td></td>
<td>PageRank</td>
<td></td>
</tr>
</tbody>
</table>
Evaluation

• How to determine whether a method is good or not?
  • Effectiveness
  • Efficiency
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