Chapter 1: Introduction

Instructor: Yizhou Sun
yzsun@ccs.neu.edu

January 8, 2013
Course Information

• Class homepage: http://www.ccs.neu.edu/home/yzsun/classes/2013Spring_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, 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**
  - Mondays, 6-9pm
  - *Exceptions:* two makeup classes for Monday holidays

- **Where**
  - Snell Library 246
  - *Exception:* classroom changes for one makeup class
Instructor and TA Information

• Instructor: Yizhou Sun
  • Homepage: http://www.ccs.neu.edu/home/yzsun/
  • Email: yzsun@ccs.neu.edu
  • Office: 476 WVH
  • Office hour: Wednesdays 3-5pm
    • Send me email to set up an appointment if you cannot make it during this time

• TA: Cheng Li
  • Email: chengli@ccs.neu.edu
  • Office: 102 Main Lab
  • Office hour: TBD

• Discussions via Piazza
Grading

- Homework: 25%
  - Three assignments are expected
  - Deadline: 11:59pm of the indicated due date via Blackboard
    - No late submissions are accepted
  - No copying or sharing of homework solutions allowed!
    - But you can discuss general challenges and ideas with others

- Course project: 20%
  - 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

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

- Final exam: 30%
  - Closed book exam, but you can take a “cheating sheet” of A4 size
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 Coverage

• Textbook Chapters

1. Introduction
2. Getting to Know Your Data
3. Data Preprocessing
4. Data Warehouse and OLAP Technology: An Introduction
5. Advanced Data Cube Technology
6. Mining Frequent Patterns & Association: Basic Concepts
7. Mining Frequent Patterns & Association: Advanced Methods
8. Classification: Basic Concepts
9. Classification: Advanced Methods
10. Cluster Analysis: Basic Concepts
11. Cluster Analysis: Advanced Methods
12. Outlier Analysis
Chapter 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
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
Before 1600: **Empirical science**

1600-1950s: **Theoretical science**
- Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.

1950s-1990s: **Computational science**
- Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
- Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.

1990-now: **Data science**
- The flood of data from new scientific instruments and simulations
- The ability to economically store and manage petabytes of data online
- The Internet and computing Grid that makes all these archives universally accessible
- Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes
- **Data mining** is a major new challenge!
Chapter 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
  - Data mining: a misnomer?

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

- Watch out: Is everything “data mining”?
  - Simple search and query processing
  - (Deductive) expert systems
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 Cleaning → Data Integration → Data Warehouse → Task-relevant Data → Data Mining → Pattern Evaluation
Example: A Web Mining Framework

- Web mining usually involves
  - Data cleaning
  - Data integration from multiple sources
  - Warehousing the data
  - Data cube construction
  - Data selection for data mining
  - Data mining
  - Presentation of the mining results
  - Patterns and knowledge to be used or stored into knowledge-base
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
Which View Do You Prefer?

• Which view do you prefer?
  • KDD vs. ML/Stat. vs. Business Intelligence
  • Depending on the data, applications, and your focus
Chapter 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
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.
Chapter 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
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
Chapter 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
Data Mining Function: (1) Generalization

- Information integration and data warehouse construction
  - Data cleaning, transformation, integration, and multidimensional data model
- Data cube technology
  - Scalable methods for computing (i.e., materializing) multidimensional aggregates
  - OLAP (online analytical processing)
- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region
Data Mining Function: (2) 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?

• How to mine such patterns and rules efficiently in large datasets?

• How to use such patterns for classification, clustering, and other applications?
Data Mining Function: (3) 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: (4) 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: (5) Outlier Analysis

• Outlier analysis
  • Outlier: A data object that does not comply with the general behavior of the data
  • Noise or exception? — One person’s garbage could be another person’s treasure
  • Methods: by product of clustering or regression analysis, ...
  • Useful in fraud detection, rare events analysis
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
  • ...
Chapter 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
Data Mining: Confluence of Multiple Disciplines

- Machine Learning
- Pattern Recognition
- Statistics
- Applications
- Visualization
- Algorithm
- Database Technology
- High-Performance Computing
Why Confluence of Multiple Disciplines?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications
Chapter 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
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)
- From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining
Chapter 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
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)

- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- Data mining and society
  - Social impacts of data mining
  - Privacy-preserving data mining
  - Invisible data mining
Chapter 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
A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - *Knowledge Discovery in Databases* (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - *Advances in Knowledge Discovery and Data Mining* (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD’95-98)
  - *Journal of Data Mining and Knowledge Discovery* (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
- ACM Transactions on KDD (2007)
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

- J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3rd ed., 2011
- 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
Chapter 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
Summary

• Data mining: Discovering interesting patterns and knowledge from massive amount of data
• A natural evolution of science and information technology, in great demand, with wide applications
• A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
• Mining can be performed in a variety of data
• Data mining functionalities: characterization, discrimination, association, classification, clustering, trend and outlier analysis, etc.
• Data mining technologies and applications
• Major issues in data mining