# CS583 – Data Mining and Text Mining

Course Web Page

http://www.cs.uic.edu/~liub/teach/cs583-fall-18/cs583.html

# General Information

- Instructor: Bing Liu
  - Email: liub@uic.edu, Tel: 312-355-1318, Office: SEO 931
- Section 1: 12:30pm-1:45pm Tue and Thu, BSB 369
- Section 2:: 3:30pm-4:45pm Tue and Thu, SES 138
- My office hours:
  - 2:00pm-3:15pm, Tuesday & Thursday (or by appointment)
- TA:
  - Section 1: Sahisnu Mazumder, sahisnumazumder@gmail.com
  - Section 2: Lichao Sun, lsun29@uic.edu
  - TA office hours: by appointment

# Course structure

- The course has two parts:
  - Lectures - Introduction to the main topics
  - Two projects (done in groups)
    - 1 programming project.
    - 1 research project.
- Lecture slides are available on the course web page.

# Grading

- Final Exam: 40%
- Midterm: 30%
  - 1 midterm
- Quiz: 10%
  - Multiple quizzes
- Projects: 20%
  - 1 programming (10%).
  - 1 research assignment (10%)

# Prerequisites

- Knowledge of
  - basic probability theory
  - algorithms

# Teaching materials

- **Required Text**
  - **Web Data Mining**: Exploring Hyperlinks, Contents and Usage data. By Bing Liu, Second Edition, Springer, ISBN 978-3-642-19459-7.
- References:
  - Data mining: Concepts and Techniques, by Jiawei Han and Micheline Kamber, Morgan Kaufmann, ISBN 1-55860-489-8.
  - Introduction to Data Mining, by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Pearson/Addison Wesley, ISBN 0-321-32136-7.
  - Principles of Data Mining, by David Hand, Heikki Mannila, Padhraic Smyth, The MIT Press, ISBN 0-262-08290-X.
  - Machine Learning, by Tom M. Mitchell, McGraw-Hill, ISBN 0-07-042807-7

# Topics

- Introduction
- Data pre-processing
- Association rules and sequential patterns
- Supervised learning (classification)
- Unsupervised learning (clustering)
- Semi-supervised learning
- Lifelong machine learning
- Information retrieval and Web search
- Social network analysis
- Opinion mining and sentiment analysis
- Recommender systems and collaborative filtering
- Web data extraction

# Feedback and suggestions

- Your feedback and suggestions are most welcome!
  - I need it to adapt the course to your needs.
  - Let me know if you find any errors in the textbook.
- Share your questions and concerns with the class – very likely others may have the same.
- No pain no gain
  - The more you put in, the more you get
  - Your grades are proportional to your efforts.

# Rules and Policies

- **Statute of limitations**: No grading questions or complaints, no matter how justified, will be listened to one week after the item in question has been returned.
- **Cheating**: Cheating will not be tolerated. All work you submitted must be entirely your own. Any suspicious similarities between students' work will be recorded and brought to the attention of the Dean. The MINIMUM penalty for any student found cheating will be to receive a 0 for the item in question, and dropping your final course grade one letter. The MAXIMUM penalty will be expulsion from the University.
- **Late assignments**: Late assignments will not, in general, be accepted. They will never be accepted if the student has not made special arrangements with me at least one day before the assignment is due. If a late assignment is accepted it is subject to a reduction in score as a late penalty.

# Introduction to the course

# What is data mining?

- Data mining is also called *knowledge discovery and data mining* (KDD)
- Data mining is
  - extraction of useful patterns from data sources, e.g., databases, texts, web, images, etc.
- Patterns must be:
  - valid, novel, potentially useful, understandable

# Classic data mining tasks

- Classification:

  mining patterns that can classify future (new) data into known classes.

- Association rule mining

  mining any rule of the form $X \rightarrow Y$, where $X$ and $Y$ are sets of data items. E.g.,

  Cheese, Milk→ Bread [sup =5%, confid=80%]

- Clustering

  identifying a set of similarity groups in the data

# Classic data mining tasks (contd)

- Sequential pattern mining:

  A sequential rule: $A \rightarrow B$, says that event $A$ will be immediately followed by event $B$ with a certain confidence

- Deviation detection:

  discovering the most significant changes in data

- Data visualization: using graphical methods to show patterns in data.

# Why is data mining important?

- Computerization of businesses produce huge amount of data
  - How to make best use of data?
  - Knowledge discovered from data can be used for competitive advantage.
- Online e-businesses are generate even larger data sets
  - Online retailers (e.g., amazon.com) are largely driving by data mining.
  - Web search engines are information retrieval (text mining) and data mining companies

# Why is data mining necessary?

- Make use of your data assets
- There is a big gap from stored data to knowledge; and the transition won't occur automatically.
- Many interesting things that one wants to find cannot be found using database queries

  "find people likely to buy my products"

  "Who are likely to respond to my promotion"

  "Which movies should be recommended to each customer?"

# Why data mining?

- The data is abundant (big data).
- The computing power is not an issue.
- Data mining tools are available
- The competitive pressure is very strong.
  - Almost every company is doing (or has to do) it

# Related fields

- Data mining is an multi-disciplinary field:
  - Machine learning
  - Statistics
  - Databases
  - Information retrieval
  - Visualization
  - Natural language processing
  - etc.

# Data mining (KDD) process

- Understand the application domain
- Identify data sources and select target data
- Pre-processing: cleaning, attribute selection, etc
- Data mining to extract patterns or models
- Post-processing: identifying interesting or useful patterns/knowledge
- Incorporate patterns/knowledge in real world tasks

# Data mining applications

- Marketing, customer profiling and retention, identifying potential customers, market segmentation.
- Engineering: identify causes of problems in products.
- Scientific data analysis, e.g., bioinformatics
- Fraud detection: identifying credit card fraud, intrusion detection.
- Text and web: a huge number of applications …
- Any application that involves a large amount of data …

# Text mining

- Data mining on text
  - Due to online texts on the Web and other sources
  - Text contains a huge amount of information of almost any imaginable type!
  - A major direction and tremendous opportunity!
- Main topics
  - Text classification and clustering
  - Information retrieval
  - Information extraction
  - Opinion mining

# Resources

- ACM SIGKDD (ACM Special Interest Group on Knowledge Discovery and Data Mining)
- Data mining related conferences
  - Data mining: KDD, ICDM, SDM, …
  - AI: ICML, NIPS, AAAI, IJCAI, ACL, …
  - Databases: SIGMOD, VLDB, ICDE, …
  - Web: WWW, WSDM, …
  - Information retrieval: SIGIR, CIKM, …
- Kdnuggets: http://www.kdnuggets.com/
  - News and resources. You can sign-up!

# Project assignments

- Done in groups:
  - Number of students per group: 2
- Project 1: Implementation
  - TBD
- Project 2: Research
  - TBD