CS 583 - Fall 2020

Data Mining and Text Mining

Taught Online Asynchronously
Using Blackboard and Zoom

Course Objective

This course has three objectives. First, to provide students with a sound basis in data mining tasks and techniques. Second, to ensure that students are able to read, and critically evaluate data mining research papers. Third, to ensue that students are able to implement and to use some of the important data mining and text mining algorithms.

Think and Ask! If you have questions about any topic or assignment, DO ASK me, TA or your classmates for help, I am here to make the course understood. DO NOT delay your questions. There is no such thing as a stupid question. The only obstacle to learning is laziness.

Online Platforms

We will mainly use Blackboard and Zoom for teaching and learning. Please get yourself familarized with the systems.

General Information

- Instructor: Bing Liu
  - Email: Bing Liu
- Teaching Assistant (TA)
  - Sections 1 and 2: Sepideh Esmaeilpour
  - Email: sesmae2@uic.edu

Section 1

- Course Call Number: 30286
- Lecture time slots: 12:30 - 1:45pm Tue & Thu
- Instructor office hours: 10:00am-11:30am Tue
- TA office hour: 4:00pm-5:00pm Thu
- Office hours will be conducted on Zoom

Section 2

- Course Call Number: 45283
- Lecture time slots: 2:00pm - 3:15pm Tue & Thu
- Instructor office hours: 10:00am-11:30am Tue
- TA office hour: 4:00pm-5:00pm Thu
- Office hours will be conducted on Zoom

Grading

- Quizzes: 30%
- No Midterm. It is replaced by
  - (40%) Assignments
- No Final Exam: It is replaced by
  - (15%) Programming assignments
  - (15%) A small text mining research project
- Assignments and the research project are done in groups of 2. Discussions with other students are allowed, but each group has to write your own code.
  - Grading: live demo + code submission
  - **MOSS**: Sharing code with your classmates is not acceptable!!! All programs will be screened using the Moss (Measure of Software Similarity.) system.

Prerequisites

- Knowledge of probability and algorithms
- Any program language for projects

Teaching materials

- Required Textbook:
  - **Web data Mining** - Exploring Hyperlinks, Contents and Usage Data, By Bing Liu, Second Edition, Springer, July 2011, ISBN 978-3-642-19459-7
- References
  - Data mining: Concepts and Techniques, by Jiawei Han and Micheline Kamber, Morgan Kaufmann Publishers, 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.
    - Data Miining. by Charu Aggarwal, Springer, 2015. ISBN 978-3-319-14142-8
    - Machine Learning, by Tom M. Mitchell, McGraw-Hill, ISBN 0-07-042807-7
    - Principles of Data Mining, by David Hand, Heikki Mannila, Padhraic Smyth, The MIT Press, ISBN 0-262-08290-X.
    - Lifelong machine learning, by Zhiyuan Chen and Bing Liu, Morgan & Claypool Publishers, November 2016.
    - Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, by Bing Liu, Cambridge University Press, 2015.
- Data mining resource site: KDnuggets Directory

Topics (subject to change; the reading list follows each chapter title)

1. Introduction
2. Data pre-processing
    - Data cleaning
    - Data transformation
    - Data reduction
    - Discretization
3. Association rules and sequential patterns (Sections 2.1 - 2.7)
    - Basic concepts
    - Apriori Algorithm
    - Mining association rules with multiple minimum supports
    - Mining class association rules
    - Sequetial pattern mining
    - Summary
4. Supervised learning (Classification) (Chapter 3)
    - Basic concepts
    - Decision trees
    - Classifier evaluation
    - Rule induction
    - Classification based on association rules
    - Naive-Bayesian learning
    - Naive-Bayesian learning for text classification
    - Support vector machines
    - K-nearest neighbor
    - Bagging and boosting
    - Summary
5. Unsupervised learning (Clustering) (Chapter 4)
    - Basic concepts
    - K-means algorithm
    - Representation of clusters
    - Hierarchical clustering
    - Distance functions
    - Data standardization
    - Handling mixed attributes
    - Which clustering algorithm to use?
    - Cluster evaluation
    - Discovering holes and data regions
    - Summary
6. Information retrieval and Web search (Sections 6.1 - 6.6, and 6.8)
    - Basic text processing and representation
    - Cosine similarity
    - Relevance feedback and Rocchio algorithm
7. Semi-supervised learning (Sections 5.1.1, 5.1.2, 5.2.1 - 5.2.4)
    - LU learning: Learning from labeled and unlabeled examples
        - Learning from labeled and unlabeled examples using EM
        - Learning from labeled and unlabeled examples using co-training
    - PU learning: Learning from positive and unlabeled examples
8. Social network analysis (Sections 7.1 - 7.4)
    - Centrality and prestige
    - Citation analysis: co-citation and bibliographic coupling
    - The PageRank algoithm (of Google)
    - The HITS algorithm: authorities and hubs
    - Mining communities on the Web
9. Sentiment analysis and opinion mining (Sections 11.1 - 11.6; check out my two books)
    - Opinion mining problem

   - Document-level Sentiment classification
   - Sentence-level subjectivity and sentiment classification
   - Aspect-level sentiment analysis
   - Mining comparative opinions
   - Opinion lexicon generation

10. Recommender systems and collaborative filtering (Section 12.4)
    - Content-based recommendation
    - Collaborative filtering based recommendation
        - K-nearest neighbor
        - Association rules
        - Matrix factorization
11. Lifelong and continual learning
    - Introduction to lifelong learning?
    - Open-world learning
    - ON-the-job learning
    - Lifelong and continual learning chatbots

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 (this includes, exams and program) 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 submission**: Late submission of assignment or quiz will not be accepted unless it is due to some extraordinary circumstances.

UIC Conunseling Center

We value your mental health and emotional wellness as part of the UIC student experience. The UIC Counseling Center offers an array of services to provide additional support throughout your time at UIC, including workshops, peer support groups, counseling, self-help tools, and initial consultations to speak to a mental health counselor about your concerns. Please visit the Counseling Center website for more information (https://counseling.uic.edu/). Further, if you think emotional concerns may be impacting your academic success, please contact your faculty and academic advisers to create a plan to stay on track.

My Home Page

By Bing Liu, Aug 15, 2020