RUTGERS - THE STATE UNIVERSITY OF NEW JERSEY

Data Mining

Instructor: **Dr. Hui Xiong**
E-mail: **hxiong@rutgers.edu**
WEB : http://datamining.rutgers.edu

- **Text Book:** "Introduction to Data Mining", by Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Addison Wesley, ISBN: 0-321-32136-7, 2005.
- **Course Description and Objectives**: Recent advances in database technology along with the phenomenal growth of the Internet have resulted in an explosion of data collected, stored, and disseminated by various organizations. Because of its massive size, it is difficult for analysts to sift through the data even though it may contain useful information. Data mining holds great promise to address this problem by providing efficient techniques to uncover useful information hidden in the large data repositories.

  The key objectives of this course are two-fold: (1) to teach the fundamental concepts of data mining and (2) to provide extensive hands-on experience in applying the concepts to real-world applications. The core topics to be covered in this course include classification, clustering, association analysis, and anomaly/novelty detection. This course consists of about 13 weeks of lecture, followed by 2 weeks of project presentations by students who will be responsible for developing and/or applying data mining techniques to applications such as intrusion detection, Web usage analysis, financial data analysis, text mining, bioinformatics, systems management, Earth Science, and other scientific and engineering areas. At the end of this course, students are expected to possess the fundamental skills needed to conduct their own research in data mining or to apply data mining techniques to their own research fields.

- **Course Web Site:**
  The Blackboard site for this course will contain lecture notes, reading materials, assignments, and late breaking news. It is accessible via: https://blackboard.newark.rutgers.edu.You should check it frequently to remain updated. You are responsible for keeping aware of the announcements on the course web site.

- **Grading Policy:**

Attendance (including in-class work)	10%
Assignments	25%
Project/Presentation/Paper	25%
Exam I	20%
Exam II	20%

  Note that the final letter grade is based on a curve.

- **Course Outline**

1. Introduction
   - What is data mining?

- Introduction to Data Mining Tasks (Classification, Clustering, Association Rules, Sequential Patterns, Regression, Deviation Detection)

2. Data and Preprocessing
   - Data Cleaning
   - Feature Selection
   - Dimensionality Reduction
3. Classification
   - Decision-Tree Based Approach (e.g. C4.5)
   - Rule-based Approach (e.g. Ripper)
   - Instance-based classifiers (e.g. k-Nearest Neighbor).
   - Bayesian Approach : Naive and Bayesian Networks
   - Classification Model Evaluation
4. Clustering
   - Partitional and Hierarchical Clustering Methods
   - Graph-based Methods
   - Density-based Methods
   - Cluster Validation
5. Association Analysis
   - Apriori Algorithm and its Extensions
   - Association Pattern Evaluation
   - Sequential Patterns and Frequent Subgraph Mining
6. Anomaly Detection
   - Statistical-based and Density-based Methods

- **Reading Material:** A lot of reading material from top conferences/journals will be made available online or in class as required. In addition, lecture notes will be available on line

- **Attendance:** Regular attendance is compulsory. You are **not** allowed to check your emails, access Web sites not related to the course or work on something that is beyond the scope of this course during the class time.

- **Assignments:** You may have discussions with your class members, but you have to submit your own work. Please be sure to keep a copy of the assignment by yourself in case that there is any problem with your hand-in or you have to use it later this semester. Assignments have to be submitted **before** the beginning of the class on the specified due day. **No late submissions will be accepted.** For assignments and project reports, you are encouraged to **type your work.**

- **Exams:** There will be **no make-up exams.** You are required to present a written proof for situations such as going on to an emergency room due to unexpected and serious illness. Chatting during the exam is **not** allowed. **No** collaboration between class members will be allowed during any exam. There will be **no** extra-credit project.

- Students are responsible for reviewing the specified chapters covered by the lecture. Please note that you are responsible for the ENTIRE contents of each chapter plus any additional

handouts, unless otherwise notified. You are not allowed to possess, look at, use, or in any other way derive advantage from the solutions prepared in prior years, whether these solutions are former students' work or copies of solutions that were made available by instructors.

- **Scholastic Dishonesty Policy:** The University defines academic dishonesty as cheating, plagiarism, unauthorized collaboration, falsifying academic records, and any act designed to avoid participating honestly in the learning process. Scholastic dishonesty also includes, but not limited to, providing false or misleading information to receive a postponement or an extension on assignments, and submission of essentially the same written assignment for two different courses without the permission of faculty members. The purpose of assignments is to provide individual feedback as well to get you thinking. Interaction for the purpose of understanding a problem is not considered cheating and will be encouraged. However, the actual solution to problems must be one's own.
- **Helpful Comments:** To get full benefit out of the class you have to work regularly. Read the textbook regularly and start working on the assignments soon after they are handed out. Plan to spend at least 12 hrs a week on this class doing assignments or reading.

Good Luck, and Welcome to the course *Data Mining*!

Dr. Hui Xiong