RUTGERS
Rutgers Business School
Newark and New Brunswick

26:198:644: Data Mining	**Professor Hui Xiong**
Fall 2017	1 Washington Park, Room 1066
Classroom: 1WP-120	Phone: 973-353-5261
Lecture: Friday 1:00PM – 3:50PM	Email: hxiong@rutgers.edu
Office Hours: Friday 10:00 AM – 11:00AM	WEB: http://datamining.rutgers.edu

TA: Mr. Ziiun Yao Office Hours: Friday 10:30AM – 11:30AM Cubicle: 1057B

COURSE DESCRIPTION

Recent advances in information 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.

Awareness of the importance of data mining for business is becoming wide spread. The industry has created more and more job opportunities for people who have interdisciplinary data analytic skills. Indeed, this course intends to bridge the gap between data mining techniques and business applications. The students have the opportunities to learn both domain and technical knowledge to face the big data challenges in the industry.

COURSE MATERIALS

- **Text Book:** "Introduction to Data Mining", by Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Addison Wesley, ISBN: 0-321-32136-7, 2005.

- **Additional Reading Materials**

  a) "Data Science for Business", by Foster Provost and Tom Fawcett, O'REILLY, ISBN: 978-1-449-36132-7, 2013.
  b) "Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Mining Systems) by Jiawei Han and Micheline Kamber, 2011.

LEARNING GOALS AND OBJECTIVES

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.

In particular, after taking this course, the students should be able to (1) approach business problems data-analytically. Think carefully & systematically about whether & how data can improve business performance, to make better-informed decision for management, finance, marketing and some other business activities; (2) interact competently on the topic of data mining for business intelligence. Know the basics of data mining processes, algorithms, & systems well enough to interact with CTOs, data mining experts, consultants, etc.

Focus: This course will explain the fundamental principles, uses, and some technical details of data mining techniques by lectures and real-world case studies. The emphasis is on understanding the business applications of data mining techniques. We will discuss the mechanics of how data analytics techniques work as is necessary to understand the fundamental concepts and business applications.

GRADING POLICY

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

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

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.

COURSE OUTLINE

1. Introduction
   - Why Big Data? What is data mining? Why data mining? Data Mining Process, relation to Business Intelligence techniques.
   - Introduction to Data Mining Tasks (Classification, Clustering, Association Analysis, Anomaly Detection). What is a model? Basic terminologies, predictive modeling.
   - Real-world data mining applications

2. Data and Preprocessing
   - Understanding of Data, what is data? Types of attributes, properties of attribute values, types of data, data quality

- Sampling, Data Normalization, Data Cleaning, Similarity Measures
- Feature Selection/Instance Selection, the importance of feature selection/instance selection in various big data scenarios

3. Classification
   - Decision-Tree Based Approach (e.g. C4.5)
   - Rule-based Approach (e.g. Ripper)
   - Instance-based classifiers (e.g. k-Nearest Neighbor).
   - Support Vector Machines (SVMs)
   - Ensemble Learning
   - Classification Model Selection and Evaluation
   - Applications: B2B customer buying stage prediction, Recommender Systems

4. Clustering
   - Partitional and Hierarchical Clustering Methods
   - Graph-based Methods
   - Density-based Methods
   - Cluster Validation
   - Applications: Customer Profiling, Market Segmentation

5. Association Analysis
   - Apriori Algorithm and its Extensions
   - Association Pattern Evaluation
   - Sequential Patterns and Frequent Subgraph Mining
   - Applications: B2B Customer Buying Path Analysis, Medical Informatics, Telecommunication alarm diagnosis

6. Anomaly Detection
   - Statistical-based and Density-based Methods
   - Ethics of data mining, privacy, what can/do firms know?

7. Data Mining Case Studies
   - Big Data Analytics in Mobile Environments
   - Fraud Detection and Prevention with Data Mining Techniques
   - Big Data Analytics in Business Environments

- **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.

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

Professor Hui Xiong