The University of Texas at El Paso
Department of Computer Science
Course Title: Data Mining
Course No.: CS 4390 (CRN 24434) and CS 5390 (CRN 24309)
Spring 2017 Syllabus

Course description: Data mining refers to exploration of data to discover knowledge. The knowledge discovered goes beyond the general pattern search where queries are known. Data mining is an analytic process to discover the unknowns about the data. The topics covered in this course are useful to gain insight and expertise on mining large-scale datasets. Along with regular lectures and discussions in this course, there will be a semester-long project and hands-on activities, especially on algorithm design, tool development, and data analysis. The course will also cover recent state-of-the-art algorithms used to discover useful information from massive amount of data. This course is beneficial for students both from industry and research perspectives.

Class meetings: Mondays and Wednesdays (3:00 pm - 4:20 pm)
Meeting place: CCSB 1.0202

Instructor: Mahmud Shahriar Hossain
Email: mhossain [at] utep [dot] edu
Web: http://www.cs.utep.edu/mhossain/
Phone: (915) 747 6340
Office: CCSB 3.0504
Office Hours: Mondays and Wednesdays (12:30 pm - 1:30 pm) and by appointment

Prerequisite for the course: Data structures, and at least one statistics course
Reference books:

- Data Mining: Concepts and Techniques, 3rd Edition, Jiawei Han, Micheline Kamber, Jian Pei
- Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Ian Witten, Eibe Frank, Mark Hall
- Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, Vipin Kumar
- The Elements of Statistical Learning, 2nd Edition, Trevor Hastie, Robert Tibshirani, Jerome Friedman, http://statweb.stanford.edu/~tibs/ElemStatLearn/
- Mining of Massive Datasets, Anand Rajaraman, Jure Leskovec, Jeffrey D. Ullman, http://infolab.stanford.edu/~ullman/mmds/book.pdf
- Introduction to Information Retrieval, Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, http://nlp.stanford.edu/IR-book/

OUTCOMES:
On successful completion of this course, students will gain the following expertise.

1. Students will be able to connect data analytic problems with one of the following groups of data mining topics, or with a subset of these groups: pattern mining and association analysis, cluster analysis, classification, link mining, graph mining, anomaly detection, recommendation systems, and dimensionality reduction.
2. Students will gain expertise in open-form data exploration to gain insight in absence of a given analytic problem.
3. Students will become familiar with data mining application development process, which includes harnessing knowledge about a target domain.

4. Students will be able to design objective functions with certain number of free variables and parallelize the optimization using multiple processors.

GRADUATE LEVEL VS. UNDERGRADUATE LEVEL EXPECTATIONS:
Graduate students are given additional and more advanced assignment and exam questions when graduate and undergraduate levels of this course are cross-listed. This course has a group project where graduate and undergraduate students work together in a team. Graduate students are given advanced study-materials and tasks for the project. The contributions of the graduate-student members of each team are focused toward critical research while the undergraduate students focus on the development of an analytic tool.

EVALUATION:
Grading components:

Midterm	15%
Final	15%
Homework, quiz, in-class exercise	25%
Project	40%
Attendance	5%

The grading scale is A: 90-100, B: 80-89, C: 70-79, D: 60-69, F: below 60.

Grade appeals: All exam/homework/quiz grades must be appealed within 7 days of the grade being posted.

Project: There will be a semester long group project. Students are encouraged to form groups of three to four members. Each group can decide to work on a topic that closely relates to the interest of the members. The project must have major data mining components in it. The target of the project is to deliver a publishable report along with the description of the methodology and full experimental results. It is expected that the final report of the project will be submitted to a major data mining conference or a relevant workshop.

Exams: A midterm and final exam will be given. Make-up exams will not be permitted except under unusual circumstances with satisfactory written justification. Any student who misses an exam due to an unexcused absence will receive a grade of zero for that exam with no opportunity for make-up or substitution. University excused absences will be excused; the exam related arrangements should be made in advance in those cases.

Homework: Regular homework will be assigned which will require significant effort outside of class. The assignments are designed to challenge you by requiring that you apply learned concepts to new situations. You should start your homework immediately after you receive it.

Quizzes and exercises: There will be regular quizzes and exercises in the class. The quizzes are not scheduled rather may appear suddenly in any day. There will be individual exercises in the class as well.

Attendance: The instructor's policy is to penalize those students who are absent. Students are expected to actively participate in classes, and show the courtesy by not arriving late or leaving early. Although attendance has a weight of 5% of the total score, the instructor reserves the right to penalize the final grade for low attendance based on the fact that collecting information regarding technology and active participation in the classroom environment is the core of this course.

CLASS POLICIES:

Electronic devices: If you bring cell phones, or similar electronic equipment into the classroom, please turn them off or put them in a "quiet" mode. Please do not read text messages or send text messages during the class. It is recommended that you bring a laptop in the classroom but you cannot chat, check messages, or surf on the internet unless the instructor advises you to do so.

WEARING HEADPHONES OF ANY TYPE IS STRICTLY PROHIBITED IN THE CLASSROOM. Headphones can be permitted only if the student has appropriate documentation approved by the Center for Accommodations and Support Services (CASS).

Standards of Conduct: In the classroom and in all academic activities, students are expected to uphold the highest standards of academic integrity. Any form of scholastic dishonesty is an affront to the pursuit of knowledge and jeopardizes the quality of the degree awarded to all graduates of UTEP.

Any student who commits an act of scholastic dishonesty is subject to discipline. Scholastic dishonesty includes, but is not limited to, cheating, plagiarism, the submission for credit of any work or any materials that are attributable in whole or in part to another person, taking an examination for another person, an act designed to give unfair advantage to a student or the attempt to commit such acts. Proven violations of the detailed regulations, as printed in the Handbook of Operating Procedures may result in sanctions ranging from disciplinary probation to a failing grade in the course, to suspension or dismissal, among others. The Handbook of Operating Procedures: Student Conduct and Discipline can be accessed at the following link: http://admin.utep.edu/Default.aspx?tabid=73922 .

DISABILITIES:

If you have a disability and need classroom accommodations, please contact The Center for Accommodations and Support Services (CASS) at 747-5148, or by email to cass@utep.edu, or visit their office located in UTEP Union East, Room 106. For additional information, please visit the CASS website at www.sa.utep.edu/cass.

TOPICS:

This course will cover the basic techniques in data mining including the preparation and manipulation of data for analysis and the creation of data from multiple dissimilar sources. Students will get hands-on experience on the data mining and knowledge discovery process.

1. Data exploration
2. Classification
3. Feature analysis
4. Association analysis
5. Link analysis
6. Cluster analysis
7. Graph mining
8. Anomaly detection
9. Recommendation systems
10. Visual analytics
11. Distributed and parallel computing

The instructor will include topics that he finds relevant as the semester progresses.

OTHER REQUIREMENTS:
Students are encouraged to bring a laptop in the classroom for in-class exercises and hands-on activities. We will be using Matlab's Statistical Toolbox, Optimization Toolbox, Global Optimization Toolbox, and Parallel Computing Toolbox for some of the exercises. It is recommended that students have Matlab Student version installed in their machines.

The instructor reserves the right to make necessary changes to this syllabus and to the delivery of the course.