CSE 4334/5334 Data Mining - Section 001

Spring 2021, Modality: Online

Instructor(s): Prof. Sharma Chakravarthy

Office Number: 632 ERB

Office Telephone Number: None

Email Address: sharmac@cse.uta.edu

Office Hours: Tu/Th: 11:00 am to Noon + by appointment
(Note: On TEAMS 2212M CSE 4334 5334 001-office-hrs channel)
Channel link:

TA: Mr. Enamul Karim, enamul.karim@mavs.uts.edu

Office Hours: Mondays: 4:30 pm to 5:30 Pm and Wednesdays: 2 pm to 3 pm + by appointment
(Note: On TEAMS 2212M CSE 4334 5334 001-office-hrs channel)

Section Information: CSE 5334 – 001

Time and Place of Class Meetings: Tu/Th 9:30 am to 10:50 am
(On TEAMS 2212M CSE 4334 5334 001-lectures channel)
Channel:

CANVAS URL: uta.instructure.com/
(For lecture notes, course projects, announcements, uploads, discussions, and grades)

Research URL: itlab.uta.edu/sharma
(For those interested in MS/PhD thesis or research projects)

Course Description: This is an introductory course on data mining. Data Mining refers to the process of automatic discovery of patterns and knowledge from large data repositories, including databases, data warehouses, Web, document collections, and data streams. We will study the basic topics of data mining, including data preprocessing, overview of data warehousing and OLAP, frequent pattern and association rule mining, correlation analysis, classification and prediction, and clustering, as well as advanced topics covering techniques and applications of data mining as part of big data analysis.

Student Learning Outcomes: A solid understanding of the basic concepts, principles, and techniques in data mining; an ability to analyze real-world data sets, to model data mining problems, and to assess different solutions; an ability to design, implement, and evaluate data mining software.

Prerequisites

- For CSE 4334: CSE 3330 Database Systems I and IE 3301 Engineering Probability (or MATH 3313 Introduction to Probability) or consent of instructor.

- For CSE 5334: There is no official prerequisites. You should have sound CSE background from your Bachelor's program (e.g., programming, data structures and algorithms, discrete mathematics, basics of probabilities and statistics). If you don't have database course from anywhere, you are allowed to take the course, but please get the consent of the instructor. You also must get the consent of the instructor if you have CSE deficiency courses to take.

Course Outline: This course will cover various topics under the umbrella of data mining. Includes projects either for developing algorithms or for analyzing given data sets using specified mining techniques. Topics will be covered under the following modules:

Module I: Intro to mining; what is not mining, overview of probability, supervised and unsupervised learning, predictive modeling (decision trees, Naïve Bayes, and may be SVM)

Module II: Cluster analysis (k-means, DBSCAN)

Module III: Association rules (Apriori and FP tree)

Module IV: Outlier detection, Graph-based approaches to mining, Multilayer Network analysis

Required Textbooks and Other Course Materials:

- (Required) Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. Introduction to Data Mining, 2nd Edition, Pearson, 2019. ISBN-13: 978-0-13-312890-1. (Sample chapters at http://www-users.cs.umn.edu/~kumar/dmbook/index.php). Also, available as eBook.
- (Required for relevant chapters) [MRS] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. Introduction to Information Retrieval, Cambridge University Press. 2008. (Free book at http://nlp.stanford.edu/IR-book/)
- (Reference) Jure Leskovec, Anand Rajaraman and Jeff Ullman. Mining of Massive Datasets, 2nd ed., Cambridge University Press. (Free book at http://www.mmds.org/#ver21)

Course Schedule is posted on Canvas as a separate document by the 1st day of classes which clearly indicates the test dates and project assignments and due dates in addition to the details of material covered in each lecture.

"As the instructor for this course, I reserve the right to adjust this schedule in any way that serves the educational needs of the students enrolled in this course. – Sharma Chakravarthy"

Descriptions of major assignments and examinations: Project: Since the emphasis of this course is on learning mining techniques for data analysis and acquiring a good understanding of *its application* to different types of data sets (that you are likely to encounter in real world), there will be a number of hands-on implementation/analysis projects as part of this course. I plan on using R and R studio (Open source license at https://www.rstudio.com/products/rstudio/download/). A wikibook on R is also available at https://www.webpages.uidaho.edu/~stevel/517/Data%20Mining%20Algorithms%20In%20R.pdf. **I may also use pandas as well to introduce alternative systems and packages that can be used for mining and analysis.**

Students may be asked to present the details of their implementation and/or analysis in the class. Demo and evaluation of the project will have an optional discussion session with the students to discuss the approach taken. Demos are expected for all projects to gauge learning outcomes.

Attendance: At The University of Texas at Arlington, taking attendance is not required but attendance is a critical indicator of student success. Each faculty member is free to develop his or her own methods of evaluating students' academic performance, which includes establishing course-specific policies on

attendance. As the instructor of this section, I encourage every student to not only attend the lectures but ask questions and make the class interactive. Based on my experience, this provides the best approach to learn and do well in the course! However, while UT Arlington does not require instructors to take attendance in their courses, the U.S. Department of Education requires that the University have a mechanism in place to mark when Federal Student Aid recipients "begin attendance in a course." UT Arlington instructors will report when students begin attendance in a course as part of the final grading process. Specifically, when assigning a student a grade of F, faculty report must the last date a student attended their class based on evidence such as a test, participation in a class project or presentation, or an engagement online via Canvas. This date is reported to the Department of Education for federal financial aid recipients.

This is an online course. All lectures will be online using TEAMS in real-time and the video will be available for later viewing. Tests and presentations will be conducted online and will be communicated to the students on Canvas and during the lecture. All tests will use lock down browser with respondus. Students should familiarize themselves with taking online tests/exams using lockdown browser and respondus. You may have to answer on separate sheets, scan them, and upload them on to Canvas.

Grading: There will be 3 hands-on projects that will constitute approximately 45% of the grade. There will be 3 in-class tests that will constitute approximately 45% of the grade. This is an initial proposal. Using office hours to do well in the course will carry 5% of the grade. Asking meaningful questions on the discussion board will constitute 5% of the grade. The instructor reserves the right to re-distribute the percentages if deemed necessary. The students may be asked to make an in-class presentation on the project experiences. Attendance and class participation is important for doing well on the course. **Grading for undergraduates will be done separately from graduates in a combined offering.**

For projects, as part of the document, what has been designed and implemented by each partner (division of labor for a team project) should be clearly stated and documented. All team members will get the same grade (including same penalty in case of plagiarism, collusion, etc.) on the project. Hence, choose your project partner wisely. Change of partners during the semester will not be allowed.

No pre-defined cut-offs. Based on past observations, typically, class average corresponds to a B grade. One standard deviation above the class average is guaranteed to earn an A; one standard deviation below class average is likely to be a C. passing grade is 50%. Note that you have to do consistently well on all projects and exams to earn an A grade.

Expectations for Out-of-Class Study: Beyond the time required to prepare for and attend each class meeting, students enrolled in this course should expect to spend at least an additional 6 hours per week in course-related activities, including reading required materials, completing assignments, preparing for exams, etc. Meet the instructor or the TA for any doubts on projects or class lecture material.

Assignments and Deadlines:

- All the assignments must be submitted through Canvas. We will NOT take hardcopy or email submission, unless the university verifies that Canvas was malfunctioning or unavailable. If you are not able to submit through Canvas due to its technical failure, you can email your assignment to us, together with a screenshot showing the technical failure. We will verify with the university.
- Everything is due by 11:59pm on the due date. The deadline is automatically managed by Canvas. If you are **allowed to submit assignments/home works late**, you automatically lose designated points as specified in the project assignment after the due time, till you reach 0. (Each individual assignment is 100 points.) We cannot waive the penalty, unless there was a case of illness or other substantial impediment beyond your control, with proof in documents. Please come and discuss with me or the TA. Penalty will be levied on a day basis. No partial-day penalties.

- **Regrading/Grade grievances:** Once the grade of a quiz/exam/project is distributed, you will have 5 business days to dispute it and get it re-evaluated. No re-evaluation will be entertained after the 5 day period. For projects, as part of the document, what has been designed and implemented by each partner (if it is done as a team) should be clearly stated and documented. All team members will get the same grade/penalty on the project.

Make-up Exams: No makeup tests or exams will be given unless there is a justifiable, documented reason.

How to Do Well in This Course: Based on the feedback I have received over the years, Students who get the most out of this course will be the ones who keep up with the material covered in the class on a regular basis. If you want to do well, attend all the lectures, read the assigned sections of the book/papers, and start early on your projects. Working out the assigned practice questions and questions from book chapters will immensely help in doing well on quizzes/tests/exams. If you are having difficulty, you owe it to yourself to get help. We will be more than happy to help you. Don't be lazy or afraid to come and see us. We will hold extensive office hours. If you can't make it to office hours but really need help, contact one of us for an appointment. I sincerely want all of you to do well**. It is your responsibility to check the web site at least twice a week.**

Institution Information

UTA students are encouraged to review the below institutional policies and informational sections and reach out to the specific office with any questions. To view this institutional information, please visit the Institutional Information page (https://resources.uta.edu/provost/course-related-info/institutional-policies.php) which includes the following policies among others:

- Drop Policy
- Disability Accommodations
- Title IX Policy
- Academic Integrity
- Student Feedback Survey
- Final Exam Schedule

Academic Integrity: Students enrolled all UT Arlington courses are expected to adhere to the UT Arlington Honor Code:

> *I pledge, on my honor, to uphold UT Arlington's tradition of academic integrity, a tradition that values hard work and honest effort in the pursuit of academic excellence.*
>
> *I promise that I will submit only work that I personally create or contribute to group collaborations, and I will appropriately reference any work from other sources. I will follow the highest standards of integrity and uphold the spirit of the Honor Code.*

Per UT System *Regents' Rule* 50101, §2.2, suspected violations of university's standards for academic integrity (including the Honor Code) will be referred to the Office of Student Conduct. Violators will be disciplined in accordance with University policy, which may result in the student's suspension or expulsion from the University.

The outcome of any cheating, plagiarism or collusion either on the project or on the quiz/exam/hw will be an automatic Fail grade (F) in the course. For projects, all members of the team will receive the same grade/penalty; so choose your partner wisely and make sure you are aware of what your partner is doing!

Electronic Communication: UT Arlington has adopted MavMail as its official means to communicate with students about important deadlines and events, as well as to transact university-related business regarding financial aid, tuition, grades, graduation, etc. All students are assigned a MavMail account and are responsible for checking the inbox regularly. There is no additional charge to students for using this account, which remains active even after graduation. Information about activating and using MavMail is available at http://www.uta.edu/oit/cs/email/mavmail.php.

Student Feedback Survey: At the end of each term, students enrolled in classes categorized as "lecture," "seminar," or "laboratory" shall be directed to complete an online Student Feedback Survey (SFS). Instructions on how to access the SFS for this course will be sent directly to each student through MavMail approximately 10 days before the end of the term. Each student's feedback enters the SFS database anonymously and is aggregated with that of other students enrolled in the course. UT Arlington's effort to solicit, gather, tabulate, and publish student feedback is required by state law; students are strongly urged to participate. For more information, visit http://www.uta.edu/sfs.

Final Review Week: A period of five class days prior to the first day of final examinations in the long sessions shall be designated as Final Review Week. The purpose of this week is to allow students sufficient time to prepare for final examinations. During this week, there shall be no scheduled activities such as required field trips or performances; and no instructor shall assign any themes, research problems or exercises of similar scope that have a completion date during or following this week *unless specified in the class syllabus*. During Final Review Week, an instructor shall not give any examinations constituting 10% or more of the final grade, except makeup tests and laboratory examinations. In addition, no instructor shall give any portion of the final examination during Final Review Week. During this week, classes are held as scheduled. In addition, instructors are not required to limit content to topics that have been previously covered; they may introduce new concepts as appropriate.

Emergency Exit Procedures: Should we experience an emergency event that requires us to vacate the building, students should exit the room and move toward the nearest exit. When exiting the building during an emergency, one should never take an elevator but should use the stairwells. Faculty members and instructional staff will assist students in selecting the safest route for evacuation and will make arrangements to assist individuals with disabilities.

Student Success Programs

UT Arlington provides a variety of resources and programs designed to help students develop academic skills, deal with personal situations, and better understand concepts and information related to their courses. Resources include tutoring by appointment, drop-in tutoring, etutoring, supplemental instruction, mentoring (time management, study skills, etc.), success coaching, TRIO Student Support Services, and student success workshops. For additional information, please email resources@uta.edu, or view the Maverick Resources website.

The IDEAS Center (https://www.uta.edu/ideas/) **(**2nd Floor of Central Library) offers **FREE** tutoring and mentoring to all students with a focus on transfer students, sophomores, veterans and others undergoing a transition to UT Arlington. Students can drop in or check the schedule of available peer tutors at www.uta.edu/IDEAS, or call (817) 272-6593.

The English Writing Center (411LIBR)

The Writing Center offers **FREE** tutoring in 15-, 30-, 45-, and 60-minute face-to-face and online sessions to all UTA students on any phase of their UTA coursework. Register and make appointments online at the Writing Center (https://uta.mywconline.com). Classroom visits, workshops, and specialized services for graduate students and faculty are also available. Please see Writing Center: OWL for detailed information on all our programs and services.

The Library's 2nd floor Academic Plaza (http://library.uta.edu/academic-plaza) offers students a central hub of support services, including IDEAS Center, University Advising Services, Transfer UTA and various college/school advising hours. Services are available during the library's hours of operation.

Librarian to Contact

Each academic unit has access to Librarians by Academic Subject that can assist students with research projects, tutorials on plagiarism and citation references as well as support with databases and course reserves.

Mandatory Face Covering Policy

All students and instructional staff are required to wear facial coverings while they are on campus, inside buildings and classrooms. Students that fail to comply with the facial covering requirement will be asked to leave the class session. If students need masks, they may obtain them at the Central Library, the E.H. Hereford University Center's front desk or in their department. Students who refuse to wear a facial covering in class will be asked to leave the session by the instructor, and, if the student refuses to leave, they may be reported to UTA's Office of Student Conduct.

Emergency Phone Numbers

In case of an on-campus emergency, call the UT Arlington Police Department at **817-272-3003** (non-campus phone), **2-3003** (campus phone). You may also dial 911. Non-emergency number 817-272-3381

Library Information

This final section is not part of the syllabus template, but a message from the UT Arlington Library.

Research or General Library Help

Ask for Help

- Academic Plaza Consultation Services (library.uta.edu/academic-plaza)
- Ask Us (ask.uta.edu/)
- Research Coaches (http://libguides.uta.edu/researchcoach)

Resources

- Library Tutorials (library.uta.edu/how-to)
- Subject and Course Research Guides (libguides.uta.edu)
- Librarians by Subject (library.uta.edu/subject-librarians)
- A to Z List of Library Databases (libguides.uta.edu/az.php)
- Course Reserves (https://uta.summon.serialssolutions.com/#!/course_reserves)
- Study Room Reservations (openroom.uta.edu/)

Ask A Librarian ..... http://ask.uta.edu

The following URL houses a page where we have gathered many commonly used resources needed by students in online courses: http://www.uta.edu/library/services/distance.php.

The subject librarian for your area can work with you to build a customized course page to support your class if you wish. For examples, visit http://libguides.uta.edu/os and http://libguides.uta.edu/pols2311fm . If you have any questions, please feel free to contact Suzanne Beckett, at sbeckett@uta.edu or at 817.272.0923.