INFO 523 Syllabus Fall 2023

Course description

INFO 523 Data Mining and Discovery- This course will introduce students to the concepts and techniques of data mining for knowledge discovery. It includes methods developed in the fields of statistics, large-scale data analytics, machine learning, pattern recognition, database technology and artificial intelligence for automatic or semi-automatic analysis of large quantities of data to extract previously unknown patterns. Topics include understanding varieties of data, data preprocessing, classification, association and correlation rule analysis, cluster analysis, outlier detection, and data mining trends and research frontiers. We will use software packages for data mining, explaining the underlying algorithms and their use and limitations. The course include laboratory exercises, with data mining case studies using data from many different resources.

Course Offering

909-2234-1 INFO 523 002 - Data Mining and Discovery

Instructor Information

Dr. Greg Chism,

Assistant Professor of Practice,

School of Information

gchism@arizona.edu

Office: Harvill 420

Office Hours: Mondays, 2-3pm, Harvill 420

Prerequisites

Students are assumed to know the basics in computer programming (e.g., variables, arrays, loops, if-then conditions), statistics (e.g., normal distribution, significance tests), and relational database (e.g., ER-diagram, SQL statements).

Course Format

Live in-person lectures, W 1:00-3:30pm, Modern Languages, RM 413. Attendance is mandatory.

Course Objective

INFO 523 is an elective in the iSchool's M.S. program. As a multidisciplinary field, the course introduce concepts and work from many areas critical to information studies including statistics, machine learning, pattern recognition, database technology, and data visualization.

Learning Outcomes

By the end of this course, students will:

- Understand a large set of concepts of data mining and knowledge discovery.
- Evaluate and use algorithms and software packages to perform data mining analyses.
- Explain and interpret results from data mining analyses.

Course Competencies

The course addresses the MS Competencies: C1 [A, B, C, D], C2, and C3

Course Workload

For each lecture hour, students are expected to spend 3-5 hours, completing the required reading and course work. Students finding themselves spending excessively more time should definitely take advantage of office hours. **The instructor welcomes students' inputs on needed workload adjustment. If students need individual workload adjustments, please also contact the instructor.**

Course Materials

Readings for the course will come from the following textbooks. All of them are **freely available online** and you do not need to purchase a physical copy of either book to succeed in this class.

Required textbooks:

- `[Data mining conceptual]` Jiawei Han, Jian Pei, Hanghang Tong. Data Mining Concepts and Techniques. 4th edition. Morgan Kaufmann, 2023.
- `[Data mining algorithms]` Pawel Cichosz. Data Mining Algorithms: Explained Using R. Wiley, 2015.
- `[Data mining case studies]` Luis Torgo. Data Mining with R: Learning with Case Studies. Chapman and Hall/CRC, 2016.

Recommended textbooks:

- `[Intro to Data Mining in R]` Michael Hahsler. Fundamentals of Data Visualization. Online Book, 2021.
- `[ggplot2-book]` Hadley Wickham, Danielle Navarro, and Thomas Lin Pedersen. ggplot2: Elegant Graphics for Data Analysis. (in progress) 3rd edition. Springer, 2022.
- `[r4ds]` Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund. R for Data Science. 2nd edition. O'Reilly, 2022.

Course Schedule

An up-to-date schedule, assignments, and due dates can be found on the course website: datamineaz.org.

Course Requirements

The four components that go into the final course grade are described below. The percentage of the final grade is in parentheses next to each. All but the final exam should be completed in groups (see Group Work Policy below).

- Homework (35%): These are selected exercises from the required textbook. Try to complete as much as possible these homework problems using R, unless specifically asked. You can turn in the homework in .R file is preferred (using comments for narrative answers), as it will be easier for the instructor to run and check your code. Where a graphic or complex math formula is needed, you can hand-draw the graph/formula and turn in a picture of it.

- R Exercises (35%): These are R code provided for you to review and practice. If you are new to R, type and run the R code in R Studio and turn in your .R file. If you know R already, skim through the R code and then apply them to a different data set of your choice (you need to include at least half of the provided code to earn credit for an exercise) and return your .R file.
- Final project (25%): A final data mining project of your choice or using open datasets.
- Class Participation (attending class, post and answer questions) (5%)

The work and course requirements are subject to change at the discretion of the instructor with proper notice to the students.

Grading:

Letter Grade	Final Course Grade
A	>= 90
B	80 - 89.99
C	70 - 79.99
D	60 - 69.99
E	50-59.99
F	< 60

Group Work Policy

- Students are encouraged to work in groups, which consists of 2-3 students on all assigned work but the final exam. Group work for studying this course (involving understanding the concept and algorithms) is very important. It creates interaction and prevents students from being feel isolated. Students can use the Form Group forum in Discussions to find teammates. Students need to get permission from the instructor to opt-out of group work.
- Students are expected to contribute actively and responsively to all group work.
- Students in a group will take turns leading a homework. The leader is responsible for coordinating the work among group members and submit one final copy of the work. Individual members are encouraged to complete an assignment on his/her own first, and then the group compare their work and submit the best one. The group leader should not do excessive amount of work for the assignment.

- Group work is seen as the result of the group collectively. Please do not blame the leader on duty if the group submission didn't get a good grade. Every member shares the responsibility of presenting the best work possible.
- An individual student should not be the leader for two consecutive homeworks.
- All students are welcome to report non-contributing team members directly to the instructor. Members not contributing actively to group work may receive reduced assignment grades / participation marks.
- When submitting a homework, note the team leader on duty and all the members in the submission.

Assignment Policy

Do not subject yourself to the late penalty: please do not make the instructor to assign a B for an A work just because you are late.

- All work must be turned in on the date due by midnight (11:59pm) Tucson time. Late work without a prior notice (at least 2 days before the due date) to and approval by the Instructor will receive 5% deduction for each late day. For example, if your work is marked at 80% but you handed it in 1 min after the due time, your mark for that assignment will be 80%*0.95=76%. Assignments late for 5 days will not be marked without an approved extension.
- In case of group work, each member is expected to contribute equally to an assignment. "Free riders" will receive a zero grade for the assignment. Groups should report non-responsive member(s) to the instructor.
- In case of a D2L or GitHub malfunction, email your assignment (so your assignment will not be marked as late) and then resubmit your assignment to D2L dropbox. Only assignments in D2L dropbox will be graded. Email submissions only establish a timestamp on an assignment.
- Be sure to check your submissions are successful. "I am not sure what had happened, but I honestly thought I had submitted my assignment on time" is not an acceptable excuse for waiving the late penalties.
- All work may be checked by **Turnitin.com** or other tools made available to the Instructor. Students may find answers to homework questions on the Web. Yes, students are allowed to check out and learn from those answers, but to avoid an plagiarism charge, students must (1) cite the source URL and (2) present their work in their own words. **Please** do not impose the difficult and time consuming task of reporting plagiarism to your Instructor, but know that the Instructor **will** report any such case if you give the opportunity. Similarly, acknowledge help received from classmates or others. These

acknowledgement will not hurt your grade, instead they reveal the academic integrity in you as a young scholar/researcher.

Final Projects

Your task for this project is to showcase your knowledge of any topic related to data mining.

This is intentionally vague -- part of the challenge is to design a project that showcases best your interests and strengths.

One requirement is that your project should feature some element that you had to learn on your own. This could be a package you use that we didn't teach in class (e.g., a package for 3D visualizations) or a workflow (e.g., making a package) or anything else. If you're not sure if your "new" thing counts, just ask!

More information will be provided throughout the semester.

Final Project Date

Tuesday, December 14

Course policies

Inclusive Excellence

Inclusive Excellence is a fundamental part of the University of Arizona's strategic plan and culture. As part of this initiative, the institution embraces and practices diversity and inclusiveness. These values are expected, respected and welcomed in this course.

This course supports elective gender pronoun use and self-identification; rosters indicating such choices will be updated throughout the semester, upon student request. As the course includes group work and in-class discussion, it is vitally important for us to create an educational environment of inclusion and mutual respect.

Academic honesty

TL;DR: Don't cheat!

Students are expected to abide by The University of Arizona Code of Academic Integrity. 'The guiding principle of academic integrity is that a student's submitted work must be the student's own.' If you have any questions regarding what acceptable practice under this Code is, please ask an Instructor.

Please abide by the following as you work on assignments in this course:

- **Collaboration:** Only work that is clearly assigned as team work should be completed collaboratively.
- The homework assignments must also be completed individually and you are welcomed to discuss the assignment with classmates at a high level (e.g., discuss what's the best way for approaching a problem, what functions are useful for accomplishing a particular task, etc.). However you may not directly share answers to homework questions (including any code) with anyone other than myself and the teaching assistants.
- For the projects, collaboration within teams is not only allowed, but expected. Communication between teams at a high level is also allowed however you may not share code or components of the project across teams.
- **Sharing and reusing code:** I am well aware that a huge volume of code is available on the web to solve any number of problems. Unless I explicitly tell you not to use something, the course's policy is that you may make use of any online resources (e.g. RStudio Community, StackOverflow) but you must explicitly cite where you obtained any code you directly use (or use as inspiration). Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism. On individual assignments you may not directly share code with another student in this class, and on team assignments you may not directly share code with another team in this class.
- **Generative AI (e.g., ChatGPT):** I am additionally aware of the potential code AI for coding (I taught a workshop on it…). While these tools are amazing, learners should be aware of the impacts that using such tools can have on core competency. David Humphrey, a computer science professor, **wrote about ChatGPT and its potentially negative impacts on core learning.** It is a good read about the pitfalls of using generative AI in an educational context. By using a generative AI, learners may miss the opportunity to discover how something works and why things are done that way. It is also important to note that the iSchool generally bans utilizing ChatGPT and generative AI in our Academic Integrity Policy.

iSchool Academic Integrity Policy

This policy agreed upon by faculty in the UArizona iSchool applies in addition to the Dean of Students' Code of Academic Integrity.

Students in courses at the UArizona iSchool are expected to maintain rigor in their academic performance with intent to learn, practice, and overcome challenges toward personal growth and enrichment. As future professionals in digital environments, iSchool students are also expected to exercise transparency and integrity in collaborations and in the use of tools and resources that may aid completion in assignments for our courses.

Consider the following PROHIBITED practices in this course, unless the instructor has specifically written instructions or permission to do otherwise:

- Posting a question on an online site such as Chegg.com, and copying and pasting some or all of the response into an assessment
- Posting an assessment from the course on online sharing sites such as Course Hero. Aiding other students in violation of academic integrity is also a violation, and is potential copyright infringement.
- Generating and submitting, in whole or in part, text or code through Artificial Intelligence such as ChatGPT, QuillBot, and text summarizers
- Using, in whole or in part, computer code not written by the student (for example, from another student, a book, or the internet) in an assignment or project. This includes using such code in modified or unmodified form.
- Searching for solutions to projects or assignments on the internet or through other tools, when your instructor intended for you to learn the solution through exercises (e.g. Googling for the solution to a question on an assignment).
- Simultaneously submitting the same assignment as another student enrolled into the course without prior permission from the instructor

Exceptions: Clear Instructions will be Provided

In any cases in which this course requires or permits students to use practices in the list above, clear written instructions will specify the tools allowed or required, so students can be certain they are working as instructed. See the UArizona iSchool Academic Integrity Policy, the UArizona Code of Academic Integrity and Syllabus policy for more information.

LLMs and ChatGPT

Large language models (LLMs) like ChatGPT are a type of artificial intelligence (AI) engine that can look like it generates the code you need for R labs and short answer questions. You are encouraged to use ChatGPT to debug code and experiment. However, abuse of ChatGPT can be traced (e.g., failing to give credit or cite ChatGPT when it is used) which could result in your suspension or termination from the course and even your program of study. Keep in mind, too, that while the code may appear legitimate, early studies have shown ChatGPT is not all that accurate with sophisticated coding. Exercise your scholarly discretion and maintain a sense of integrity in your statistical learning journey.

See my policies on this subject above.

“Incomplete” grade

The grade of I may be awarded only at the end of a term, when all but a minor portion of the course work has been satisfactorily completed. The grade of I is not to be awarded in place of a failing grade or when the student is expected to repeat the course; in such a case, a grade other than I must be assigned. Students should make arrangements with the instructor to receive an incomplete grade before the end of the term. If the incomplete is not removed by the instructor within one year the I grade will revert to a failing grade.

Tutoring

Tutoring can be found through the UArizona Think Tank.

Additional university policies

Additional policies can be found at this link (please read through them): https://catalog.arizona.edu/syllabus-policies

Important Dates

Information contained in this course syllabus, may be subject to change, as deemed appropriate by the instructor.Important dates

- **Monday, August 21:** Classes begin, Monday schedule
- **Monday, August 28:** Drop/add ends
- **Monday, September 4:** Labor Day, no class
- **Sunday, September 17:** Last day to drop without a W (withdraw)
- **Sunday, October 29**: Last day to withdraw from a class online through UAccess
- **Friday, November 10**: Veteran’s Day, limited support available
- **Thursday - Sunday, November 23 - 26**: Thanksgiving recess, no support available
- **Thursday, December 14**: Project presentations

For more important dates, see the full UArizona Academic Calendar.

Graduate Student Resources

University of Arizona’s Basic Needs Resources page for graduate students: http://basicneeds.arizona.edu/index.html