

Fall 2021 INFO 523 – Data Mining and Discovery

Instructor: Dr. Cristian Roman-Palacios, Professor, Information Technology

Office: 445D, School of Information in the Harvill Building

Email: cromanpa94@email.arizona.edu

Office Hours: Office hours will be on zoom. Feel free to email me!

Course Description:

This course will introduce students to the concepts and techniques of data mining for further use in scientific discovery. Data mining includes methods developed in the fields of statistics, large-scale data analytics, machine learning, pattern recognition, database technology and artificial intelligence to automatically or semi-automatically analyze large quantities of data to extract previously unknown, interesting patterns. Topics in this course include: understanding types of data; preprocessing data; classifying data performing association and correlation rule analyses, cluster analysis and outlier detection; data mining trends; and research frontiers. We will use software packages for data mining with a focus on explaining the underlying algorithms of the software and understanding their use and limitations. The course includes laboratory exercises with data mining case studies using data from many different resources.

Course Prerequisites:

Students are expected to know the basics in computer programming (e.g., variables, arrays, loops, if-then conditions) and statistics (e.g., normal distribution, significance tests).

Course Objectives:

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

By the end of this course, students will:

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

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 take advantage of office hours. **The instructor welcomes students' input on needed workload. Feel free to get in touch with the instructor.**

Course Materials:

Required textbook (freely available through UA library!):

- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
- Torgo, L. (2011). *Data mining with R: learning with case studies*. Chapman and Hall/CRC.
- Kabacoff, R. I. (2015). *R in action: data analysis and graphics with R*. Simon and Schuster.

Technical Prerequisites:

All students taking this course will need to satisfy the standard School of Information technical requirements. Having a computer with a solid internet connection will be essential.

Course Requirements:

The five 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 (30%): These are selected exercises from different sources. Please use R when writing code is required to solve the homework. If you feel more comfortable writing code in another language, please get in touch with the instructor. Homework submissions should be fully reproducible and should include the relevant script (e.g. *.R or *.Rmd) and data (e.g. *.csv, *.RData; if necessary). Please annotate your scripts as much as possible.
- R Exercises (30%): R code provided for you to review and practice different topics that are discussed in class. If you are new to R, type and run the R code in R Studio and turn in your *.R file ***with your own annotations***. If you are more comfortable using R, feel free to make any changes to the code (or data), annotate them, improve it and turn in the resulting script as an *.R file.
- Final Project (30%). Please find more details on both assignments on D2L. In short, we will have a one-week conference session with students in the course presenting to and attending talks by their classmates. There will be a final project, which could be related to the conference session, in which students will use their data mining skills to answer a research question.

- **Class Participation (10%):** The final percentage earned for this component will be calculated as 10% weighted by the total number of weeks the student either (i) participated in D2L discussions, (ii) attended office hours, (iii) helped improve materials, or (vi) pointed out new resources to classmates.

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

Grading:

A=90+ (Superior Work)

B=80-89 (Very Good)

C=70-79 (Marginally Satisfactory)

D=60-69 (Failed to meet requirements)

E=50-59 (Failed to meet requirements)

F=0-49 (Failed to meet requirements)

Group Work Policy:

- Students are encouraged to work in groups of 2–3 students on any assigned work. In an online class, group work creates interaction and prevents students from feeling isolated. Students can use the 'Form Group' forum in Discussions to find teammates.
- Students are expected to contribute actively and responsively to all group work.
- Group work is seen as the result of a group collectively. 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.
- 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 or participation marks.
- When submitting homework, please note the team leader on duty and all the members in the submission in the header of the R file.

Assignment Policy:

- Do not subject yourself to the late penalty: please do not make the instructor 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. 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 malfunction, email your assignment to the instructor.
- Be sure to check your submissions are successful.
- 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 a 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 her the opportunity. Similarly, acknowledge help received from classmates or others. These acknowledgements will not hurt your grade, instead they reveal the academic integrity in you as a young scholar/researcher.

COURSE, SCHOOL, AND UNIVERSITY 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 Code of Integrity

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.

Policies Against Disruptive and Threatening Behavior

Students are also bound by the University's policies related to disruptive behavior and threatening behavior.

Accessibility and Accommodations

At the University of Arizona we strive to make learning experiences as accessible as possible. If you anticipate or experience physical or academic barriers based on disability or pregnancy, you are welcome to let me know so that we can discuss options. You are also encouraged to contact Disability Resources (520-621-3268) to explore reasonable accommodation.

"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.

Additional Policies:

The Arizona Board of Regents' Student Code of Conduct, ABOR Policy 5-308, prohibits threats of physical harm to any member of the University community, including to one's self. See: <http://policy.web.arizona.edu/threatening-behavior-students>

All student records will be managed and held confidentially. See: <http://www.registrar.arizona.edu/ferpa/default.htm>

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

Course Schedule:

The instructor will also post a link to a google calendar D2L.

Week	Date	Main topics	Homework due	Final project – virtual conference	Additional readings	Need to turn something in?
0.5	11-Oct	Introduction to data mining.			Han & Kamber, Ch 1	No!
		Know your data: data types, distributions, basic tests.				
1	18-Oct	Introduction to R: basic math, data types, loops, if-else statements, working with data structures, loading/writing data, writing basic functions.	Homework 1: data mining and data types.	Identify a partner! Feel free to start developing ideas for your project.	Kabacoff, Ch 1–4	Yes! (Homework)
		Introduction to R and reproducibility: GitHub, RStudio, Markdown and Shiny.				
2	25-Oct	Data pre-processing: base R.	Homework 2: Basic R and shiny.	Select a topic and start working on the presentation!	Kabacoff, Ch 6	Yes! (Homework)
		Data pre-processing: tidyverse.				

3	1-Nov	Exploratory data analysis in R: Descriptive stats, variance, sd, se, covariation, correlation, skewness, and kurtosis.	Homework 3: tidyverse and base R	Submit your availability for week 7	Kabacoff, Ch 7	Yes! (Homework and availability)
		Plotting in base and ggplot: density plots, boxplots, dot plot, bar charts, pie charts, statistical modeling.			Kabacoff, Ch 11 and 16	
4	8-Nov	Unsupervised techniques: dimensionality reduction (e.g. PCA).	Homework 4: Basic plotting and basic shiny app for plotting data	Checkpoint Abstract and small R package with your code (or shiny app!).		Yes! (Homework and progress on your final project)
		Unsupervised techniques: clustering (e.g. k-means, k-medoids, hierarchical cluster analysis).				
5	15-Nov	Supervised techniques, regression I: linear and logistic models.	Homework 5: unsupervised techniques		Kabacoff, Ch 8	Yes! (Homework)
		Supervised techniques, regression II: decision trees, GLS, GLM, LMM, GAM,			Kabacoff, Ch 13	

		random forest; accounting for temporal/spatial correlation structures.				
6	22-Nov	Making code available: R packages		Submit a small R package with your code (or shiny app!). Turn in the conference abstract.		Yes! (Final project! Part 1)
7	29-Nov	Mini conference		Turn in recordings and feedback		Yes! (Final project! Part 2)
7.5	6-Dec	Final review and discussion	Homework 6: supervised methods			Homework!

