partic

CSC380: Principles of Data Science

MW 3:30-4:45pm, WSEL200SW

Course Description

This course introduces students to principles of data science that are necessary for computer scientists to make effective decisions in their professional careers. A number of computer science sub-disciplines now rely on data collection and analysis. For example, computer systems are now complicated enough that comparing the execution performance of two different programs becomes a statistical estimation problem rather than a deterministic computation. This course teaches students the basic principles of how to properly collect and process data sources in order to derive appropriate conclusions from them. The course has main components of: basic probability, basic statistics and data wrangling, and basic data analysis using programming libraries.

Instructor and Contact Information

Chicheng Zhang, Gould-Simpson 720 chichengz at cs dot arizona dot edu

Teaching assistants: Yichen Li, Jianwei Shen, Tian Tan; Emails: {yichenl, sjwjames, tiantan} at arizona dot edu

Office Hours: See D2L

Course homepage

D2L

Course Format and Teaching Methods

The course is taught in a lecture format, with lectures given by the instructor. Lectures will involve discussions making use of lecture slides and input from students. Lectures will augment, enhance, and reiterate material from the relevant textbooks. Assignments and exams all involve individual work.

Course Objectives

An introduction to basic concepts in data science and machine learning. Topics include: descriptive statistics, basic data analysis, basic data visualization, predictive models and training, basic supervised and unsupervised learning models, evaluation measures.

Expected Learning Outcomes

A student who successfully completes this course will be able to:
• Explain the difference between different measures of centrality and variability (means vs medians, variance vs interquartile range, etc.)
• Articulate the meaning of statistical hypothesis tests
• Learn how to use probability and non-probability sampling to collect data from a population and how to identify potential sampling bias
• Convert a "raw" data source into a version appropriate for downstream analysis using Python
•Write appropriate visualizations for different sources and types of data

• Explain why we seek to build machine learning models that generalize rather than memorize their inputs
• Explain the different uses for training, validation, and testing datasets
• Select appropriate evaluation measure for the dataset and task being solved
• Articulate the difference between supervised and unsupervised machine learning, as well as select the appropriate methodology for a given problem

Transferable Career Skills

National Association of Colleges and Employers (NACE) Career Readiness:

Career readiness is a foundation from which to demonstrate requisite core competencies that broadly prepare the college-educated for success in the workplace and lifelong career management. For new college graduates, career readiness is key to ensuring successful entrance into the workforce.

There are eight career readiness competencies, each of which can be demonstrated in a variety of ways." (NACE, 2025)

- Career & Self Development
- Communication
- Critical Thinking
- Equity & Inclusion
- Leadership
- Professionalism
- Teamwork
- Technology

In this course, we will focus on the following competencies:

- Technology: Students will learn data analysis and machine learning software libraries such as Pandas and scikit-learn and apply them to real-world datasets in a hands-on manner.
- Teamwork: The course will include lectures, in-class discussions, and activities, with a strong emphasis on teamwork through paired assignments and group projects that require collaboration with other students both during and outside of class.
- Communication: Students are encouraged to emphasize communication by interacting with teaching assistants and using various channels, such as D2L, email, class discussions, and Piazza, to stay updated on course materials.
- Critical Thinking: Students will learn best practices and common pitfalls in data collection and analysis pipelines and be able to use such knowledge to do better data analysis in their own applications.

Makeup Policy for Students Who Register Late

Makeup can only be made for HW1 with the submission deadline being 8 days after the enrollment.

Course Communications

The primary path for outside-lecture communications will be Piazza. The instructor and TAs can also be reached through email.

Required Texts or Readings

There is no single designated textbook for this course. Much of the course materials and assigned readings will be based on the following books:

Watkins, J., "An Introduction to the Science of Statistics: From Theory to Implementation" (Available at https://www.math.arizona.edu/~jwatkins/statbook.pdf)

Wasserman, L. "All of Statistics: A Concise Course in Statistical Inference." Springer, 2004 (Accessible online via UA library)

James, G., Witten, D., Hastie, T., & Tibshirani, R. An Introduction to Statistical Learning with Applications in Python. New York: Springer (Available at https://www.statlearning.com/) (Download ISL with Ptyhon).

Steven S. Skiena, "The Data Science Design Manual", Springer, 2017 (Available at https://link.springer.com/book/10.1007/978-3-319-55444-0)

Sam Lau, Joey Gonzalez, Deb Nolan, "Learning Data Science", O'Reilly, 2023 (Available at https://learningds.org/intro.html)

Assignments and Examinations: Schedule/Due Dates

Assignments

There will be 7 assignments which will be given approximately once in two weeks (the exact dates are below). Each assignment will have the same overall weight toward the final grade and the assignment with the lowest score will be dropped. Assignments must be done individually or in pairs.

Project

There will be one project. The submissions are due on Thursday, May 8. The project can be done in groups of at most 4.

Quizzes

There will be 12 quizzes. Each quiz will be announced one day before the quiz. Each quiz can be done individually or in pairs. Each quiz will have the same overall weight toward the final grade and the two quizzes with the lowest scores will be dropped.

Exams

There will be one midterm exam and one final exam. The midterm will tentatively be held during the regular lecture time on Wednesday, Mar 5. The final is scheduled for Tuesday, May 13, 3:30 - 5:30pm.

	Description	*Assigned Date*
HW1	*Data Processing*	*1/22*
HW2	*Probability 1*	*1/31*
HW3	*Probability 2*	*2/13*
HW4	*Statistics 1*	*2/27*
Midterm		*3/5*
HW5	*Statistics 2, exploratory data analysis*	*3/17*
HW6	*Linear Supervised Learning Models*	*4/3*
HW7	*Nonlinear & Unsupervised Learning Models*	*4/13*
Project		*3/16*
Final Exam		*5/13*

Homeworks will generally be due in 7-12 days. No late submissions will be accepted.

Final Examination

The final is scheduled for Tuesday, May 13, 3:30 - 5:30pm.

Link to the Final Exam Regulations and Final Exam Schedule: https://registrar.arizona.edu/faculty-staff-resources/room-class-scheduling/schedule-classes/final-exams

Grading Scale and Policies

Course grades use the regular scale of A, B, C, D, and E. Grades are determined from a weighted average of the percentage of points earned on assignments and exams, according to the following weighting:

30% Assignment (5% each, lowest assignment grade is dropped),

15% Project

10% Quiz (1% each, two quizzes with the lowest scores are dropped),

20% midterm exam

25% final exam

Attaining a weighted score of at least 90% guarantees an A in the course, at least 80% guarantees a B, at least 70% guarantees a C, and at least 60% guarantees a D. These thresholds may be lowered, but will not be raised.

Additionally there is a 5 point bonus opportunity based on participation. There are four conditions of these bonus points: A "participation instances" in lectures, B participations in piazza (A + B >= 8), 4 appearances in instructor's office hours, and ~~4 discussion session participations~~. All ~~four~~ conditions must be satisfied to guarantee the all-or-nothing 5 bonus points. The participation instances in lectures are in the form of answering an in-class question. The piazza participations are in the form of correctly answering a technical question posed by a peer. Note that at most one participation per month is counted for the piazza participations.

On homework, very-high-level ideas can be discussed with friends outside the 2-person group, but solutions must represent the group's work and must be written up separately for each group. Use of solutions from previous offerings of the course or from any other external source is not permitted.

Assignment that is late (submitted after the due date and time) is not accepted. Any project report that is late is not accepted. There will be no makeups for missed quizzes. Each graded assignment will be returned before the next assignment is due and the exams will be graded at most 15 days after they are given. Grading delays beyond these promised return-by dates will be announced as soon as possible with an explanation for the delay.

Incomplete (I) or Withdrawal (W):

Requests for incomplete (I) or withdrawal (W) must be made in accordance with University policies, which are available at https://catalog.arizona.edu/policy/courses-credit/grading/grading-system.

Dispute of Grade Policy
Requests for regrading assignments or exams must be submitted within one week of receiving the graded homework or exam.

Honors Credit

Students wishing to contract this course for Honors Credit should e-mail me to set up an appointment to discuss the terms of the contact and to sign the Honors Course Contract Request Form. The form is available at http://www.honors.arizona.edu/honors-contracts

Scheduled Topic and Activities

This is a list of the topics we will cover. Please note that the schedule may be adjusted based on instructor discretion.

Jan 15: Introduction to Data Science

Jan 22: Basic data manipulation using Pandas; descriptive statistics; basic data visualization

Jan 27, 29: Probability introduction

Feb 3, 5: Conditional Probability, Random Variables and Distributions

Feb 10, 12: Expectation, Special Distributions

Feb 17, 19: Special Distributions, Law of Large Numbers, Central Limit Theorem

Feb 24: Data visualization

Feb 26, Mar 3: Introduction to Estimation, Unbiased Estimation

Mar 5: Midterm

Mar 17, 19: Maximum Likelihood Estimation, Method of Moments

Mar 24, 26: Interval Estimation, Hypothesis Testing

Mar 31, Apr 2: Hypothesis Testing, Data Collection

Apr 7, 9: Introduction to Modeling, Basic Machine Learning Pipeline

Apr 14, 16: Linear Regression, Logistic Regression, Nearest Neighbor Methods

Apr 21, 23: Nonlinear prediction models

Apr 28, 30: Unsupervised Learning: Clustering, Linear algebra preliminaries

May 5: SVD and PCA

May 7: Course Wrap-up

Classroom Behavior Policy

To foster a positive learning environment, students and instructors have a shared responsibility. We want a safe, welcoming, and inclusive environment where all of us feel comfortable with each other and where we can challenge ourselves to succeed. To that end, our focus is on the tasks at hand and not on extraneous activities (e.g., texting, chatting, reading a newspaper, making phone calls, web surfing, etc.).

Students are asked to refrain from disruptive conversations with people sitting around them during lecture. Students observed engaging in disruptive activity will be asked to cease this behavior. Those who continue to disrupt the class will be asked to leave lecture or discussion and may be reported to the Dean of Students.

Some learning styles are best served by using personal electronics, such as laptops and iPads. These devices can be distracting to other learners. Therefore, students who prefer to use electronic devices for note-taking during lecture should use one side of the classroom.

Notification of Objectionable Materials

This course will contain material of a mature nature, which may include explicit language, depictions of nudity, sexual situations, and/or violence. The instructor will provide advance notice when such materials will be used. Students are not automatically excused from interacting with such materials, but they are encouraged to speak with the instructor to voice concerns and to provide feedback.

Safety on Campus and in the Classroom

For a list of emergency procedures for all types of incidents, please visit the website of the Critical Incident Response Team (CIRT): https://cirt.arizona.edu/case-emergency/overview

Also watch the video available at

https://arizona.sabacloud.com/Saba/Web_spf/NA7P1PRD161/app/me/ledetail;spf-url=common%2Flearningeventdetail%2Fcrtfy000000000003841

University-wide Policies link

Links to the following UA policies are provided here: https://catalog.arizona.edu/syllabus-policies

- Absence and Class Participation Policies
- Threatening Behavior Policy
- Accessibility and Accommodations Policy
- Code of Academic Integrity
- Nondiscrimination and Anti-Harassment Policy

Academic Integrity policies

For homeworks: If you discuss high-level ideas with a friend outside group, please mention their name in your solutions.

• Verbatim copying of solutions from any external source (web search, ChatGPT, etc.) is not acceptable. If you make use of any external source, mention the source in your solution, and make sure that your solution is your own work.

• If the instructors suspect that there is academic integrity violation, we will meet and you may plead your case. If the instructors are not convinced, the case will be reported to the Dean of Students office.

Department-wide Syllabus Policies and Resources link

Links to the following departmental syllabus policies and resources are provided here, https://www.cs.arizona.edu/cs-course-syllabus-policies :

- Department Code of Conduct
- Class Recordings
- Illnesses and Emergencies
- Obtaining Help
- Preferred Names and Pronouns
- Confidentiality of Student Records
- Additional Resources

- Land Acknowledgement Statement

Subject to Change Statement

Information contained in the course syllabus, other than the grade and absence policy, may be subject to change with advance notice, as deemed appropriate by the instructor.