ASTR 596: Fundamentals of Data Science

Prof. Gautham Narayan
✉: gsn@illinois.edu
☎: +1 (217) 300-7322

Lecture: Astronomy 134, Mon & Wed, 1400 – 1530
🌍: https://github.com/gnarayan/ast596_2023_Spring
Office Hours: Astronomy 129, TBD

COURSE DESCRIPTION & LEARNING GOALS

This course will cover a number of statistical techniques that are relevant to astrophysical studies. These include robust statistics, regression, model building and hypotheses testing, MCMC methods, parameter estimation, time series analysis, clustering and dimensionality reduction, and hierarchical modeling. We will also cover best practices for writing code and version control. These techniques are ubiquitous in science and industry. My goal is to provide a survey of these techniques, together with realistic problems, so that you see how they work and what their implicit assumptions are.

PREREQUISITES

Undergraduate calculus or analysis, undergraduate statistics, undergraduate linear algebra, and some familiarity with programming in Python. You will also need a computer with a working `conda` and `git` installation for much of the coursework. Ideally, this is your own laptop but you can use the UIUC campus cluster. Request access at https://campuscluster.illinois.edu/new_forms/user_form.php.

TEXTS & READINGS

- "Statistics, Data Mining, and Machine Learning in Astronomy", Ž. Ivezić, A. Connolly, J. T. VanderPlas & A. Gray
- "Python Data Science Handbook", J T. VanderPlas

Copies of both books are available online (the previous edition for ICVG).

ICVG: Through O'Reilly (free registration with your illinois.edu email required)
https://learning.oreilly.com/library/view/statistics-data-mining/9780691151687/
or through JSTOR: https://www.jstor.org/stable/j.ctt4cgbdj

VdP: On GitHub https://jakevdp.github.io/PythonDataScienceHandbook/

Other Resources:

"Modern Statistical Methods for Astronomy", E. Feigelson with J. Babu (**FB**) is detailed, though focuses on using R. You may also find "Bayesian Models for Astrophysical Data", J. M. Hilbe, R. S. de Souza, & E. E. O. Ishida helpful. Copies of both are in the library.

The LSST Data Science Fellowship Program has a huge collection of worked notebooks and video lectures: https://github.com/LSSTC-DSFP/LSSTC-DSFP-Sessions.

I recommend "Data Analysis: A Bayesian Tutorial", D. S. Sivia and J. Skilling if you need a quick refresher on prerequisite material.

GRADING

Your grade is determined from a combination of assignments, midterm and a final project. Policies for each are below. Attendance is at your own discretion, and there are no planned opportunities for extra credit. You are welcome to discuss your grades and your work in the course with me during office hours.

Points:
- Weekly Assignments: 50%
- Midterm: 20%
- Final: 30%

This course (like every other one at UIUC...) uses a plus (+) and minus (−) grading scale for course grades.

97-100=A+; 93-96=A; 90-92=A-; 87-89=B+; 83-86=B; 80-82=B-; 77-79=C+; 73-76=C; 70-72=C-; 67-69=D+; 63-66=D; 60-62=D-; 0-59=F

COURSE POLICIES

I've outlined standards for this course below. Times listed in this syllabus are US/Central throughout. If something is not covered by my policies, please discuss it with me. My contact information is at the beginning of this syllabus and in the "Reaching Me" section.

Assignment & Exam Policies: Assignments, as well as midterm and final examinations are open book and take home. You may work in groups, and may discuss the assignments and ways to tackle it, but you must write/code your solution independently. This means you get to talk with each other, discuss how you'd solve a problem, but come up with your own solution, but not share your solutions. Over the course of the last three semesters, a total of 5 students thought I'd not catch that level of cheating. They all failed.

Assignments/exams will be posted to the course GitHub repo on Thursdays. Make a fork of the repo, create a folder with your name for your work, write/code up your solution as directed in the assignment, commit, and open a pull request when you are satisfied with your work before Noon the following Wednesday. You are allowed to drop ONE assignment from your total, for whatever reason, no questions asked (and if you don't elect to, I'll drop your lowest).

The midterm and final examinations will be posted online on Mar. 2, 2023 and May 4, 2023 respectively, and will be due on Mar. 8, 2023 and May 10, 2023 respectively by Noon. If you have a conflict with these dates, please contact me as soon as possible. Make up examinations will have different questions. Exams include all material covered prior, and will require a more substantial time commitment that the weekly assignments.

While I am open to accommodating students who need to take these tests at different times for whatever reason, all grades for the course are due to the Provost by May 19, 2023, and I cannot provide extensions beyond that date, unless there are absolutely extenuating circumstances (see below).

Grades of Incomplete: Incomplete (I) grades are given only in situation where unexpected emergencies prevent you from completing the course and the remaining work can be completed the next semester. Documentation must be provided, and the instructor is the final authority on whether you qualify for an incomplete. Incomplete work must be finished by the

10th day of instruction in the Fall 2023 semester, else the "I" will automatically be recorded as a "F" on your transcript.

Late or Missed Assignments: All work is assigned on Thursday and due the following Thursday before class begins. If you know that you will be turning an assignment in late please notify me in advance. A full letter grade will be deducted for each day an assignment is late until a "F" grade is achieved, unless you have a documented medical excuse or you have notified me of other extenuating circumstances. Remember that you may drop ONE assignment from your total, for whatever reason, no questions asked.

Accessibility Accommodation: It is my goal that this class be an accessible and welcoming experience for all students, including those with disabilities that may impact learning in this class. If the design of this course poses barriers to you effectively participating and/or demonstrating learning in this course, please meet with me, with or without an Accessibility Services accommodation letter, to discuss reasonable options or adjustments. You are welcome to talk to me at any point in the semester about course design concerns, but it is always best if we can talk at least one week prior to the need for any modifications.

During our discussion, I may suggest the possibility/necessity of your contacting the Office of Disability Resources and Educational Services (1207 S. Oak St., Champaign, IL 61820; 217-333-1970) disability@illinois.edu; `http://disability.illinois.edu/`) to talk about academic accommodations.

Plagiarism: Don't. You are going to be using GitHub for assignments, so there's a record of your commits, and it is trivial to check if chunks of your work match someone else. You may work in groups together, and may discuss the assignments and ways to tackle it, but you must write/code your solution independently. Read the University of Illinois' policy on plagiarism.

Plagiarism and cheating of any kind on an assignment or examination will result at least in an "F" for that work, and may also lead to an "F" for the entire course. Plagiarism and cheating subjects a student to referral to the Senate Committee for Student Discipline for further action.

I am confident in each of your ability to tackle the course work. My group work policy is designed to encourage you to learn how to collaborate, but the assignments are designed to test YOUR grasp of the material. If you feel you need help with material, come see me during office hours or any time my door is open.

Classroom Behavior: I expect you to live up to your roles as student-scholars. Students must follow the University of Illinois' standards for personal and academic conduct. Proper conduct entails creating a **positive** learning experience for all students, regardless of sex, race, religion, sexual orientation, social class, or any other feature of personal identification; therefore, **sexist, racist, prejudicial, homophobic, or other derogatory remarks will not be tolerated.**

Syllabus Amendment: This syllabus may be amended or modified in any way upon notice, with the version on GitHub being authoritative. Most such changes will affect the tentatve schedule, but be sure that you know if any due dates change.

Important Dates:

- Jan. 18, 2023: First day of class
- Mar. 2, 2022: Midterm Exam Assigned (due Mar. 8 by Noon)
- May 3, 2023: Last day of classes
- May 4, 2023: Final Exam Assigned (due May 10 by Noon)
- May 20, 2023: Grades available for viewing on Student Self-service portal

CLASS SCHEDULE (subject to revision)

- **WEEK 0**
  First steps, crash course in python
- **WEEK 1**
  Probability distributions, descriptive statistics, the Central Limit theorem and when it doesn't hold, robust statistics, and hypothesis testing (ICVG Ch. 3, FB Ch. 2)
- **WEEK 2**
  Statistical inference, frequentist properties such as unbiasedness & the Cramér–Rao bound, consistency, asymptotic limits, mean-squared errors (ICVG Ch. 4, FB Ch. 3)
- **WEEK 3**
  Maximum likelihood estimation and applications, ranting about minimizing χ^2 (ICVG Ch. 4)
- **WEEK 4**
  Regression & Inference: ordinary least squares, generalized least squares, orthogonal distance regression vs generative modeling of data (ICVG Ch. 8, FB Ch. 7)
- **WEEK 5**
  Bayes in practice, sampling and Markov Chain Monte Carlo methods (ICVG Ch. 5)
- **WEEK 6**
  Building models, effective sampling techniques, estimating parameters & uncertainties, posterior predictive checks, other MCMC wizardry (ICVG Ch. 8). Midterm exam posted.
- **WEEK 7**
  Visualization (VdP Ch. 4), Midterm exam due. No homework assignment because it's spring break and I'm not that mean.
- **WEEK 8**
  Time-series analysis (ICVG Ch. 10, FB Ch. 11), Gaussian processes (ICVG Ch. 8.10, readings from Rasmussen & Williams)
- **WEEK 9**
  Probabilistic Graphical Models (PGMs) & hierarchical Bayes (Readings from Hilbe, de Souza & Ishida)
- **WEEK 10**
  The ABCs of not having a likelihood function (Readings from Hilbe, de Souza & Ishida)
- **WEEK 11**
  Intro to Machine learning, tree methods (ICVG Ch. 9, VdP Ch. 5)
- **WEEK 12**
  Gaussian mixture models, density estimation, unsupervised clustering techniques, and dimensionality reduction (ICVG Ch. 6, 7, FB Ch. 9 and bits of Ch. 6, VdP Ch. 5)
- **WEEK 13**
  Dealing with outliers, imbalanced, and missing data, supervised machine learning techniques
- **WEEK 14**
  supervised ML continued, putting it all together. Final exam posted on May 4th.
- **May. 10th**
  Final exam due by Noon.