AST 8581 / PHYS 8581 / CSCI 8581: Big Data in Astrophysics

Prof. Michael Coughlin

Spring 2022

E-mail: `cough052@umn.edu`
Class Hours: Mon/Wed 9:00-10:55 pm   Class Room: Keller Hall 4-250 Zoom
Office Hours: TBD   Office: Tate 275-02
Website: https://github.com/mcoughlin/ast8581_2022_Spring
TAs: Laura Salo (salo0090@umn.edu); Sai Sundar (sunda127@umn.edu)

Please read the entire syllabus carefully; you are responsible for all of the requirements and procedures described here. You are also responsible for all announcements, assignments, changes, etc., whether or not you are in class.

This course will introduce key concepts and techniques used to work with large datasets, in the context of the field of astrophysics. In the first 4 weeks of the course the focus will be on the modern approaches to creating and manipulating large data sets, with the focus on time series analyses and Bayesian methods applied to astrophysics survey data. The remaining part of the course will focus on a range of machine learning techniques for processing data: classification algorithms (supervised and unsupervised learning), clustering algorithms, regression problems, recommender systems, graphic models and others. The course will dedicate about 2 weeks to each algorithm type: the algorithms will first be introduced in 1-2 lectures, and the emphasis will then be placed on team projects in which the students will apply the algorithms (and already available packages) to astrophysical data sets to answer specific astrophysics questions. The course will assume familiarity with basic concepts in astrophysics, but it will include brief reviews as needed to demonstrate the use of modern data analysis techniques.

Due Dates

Exams

<u>Mid-Term 1</u>: Feb 28 - Mar 5 (Take home exam; replacement for homework)

<u>Mid-Term 2</u>: April 11 - 15 (Take home exam; replacement for homework)

Final Project

For the projects, the plan of action is:

1. Look at the git repo created for your team - course staff will be viewing progress and meeting with you at least once to discuss your plan

2. In your repo, maintain a statement of work, with what each team member is working on

3. Divide up the tasks in your project however you see fit, but each of you needs to be writing code and pushing to git - no making one person do all the work

4. You should have multiple meetings as a group, to be sure things are on-track (and commit to the repo the minutes of these meetings)

5. You get to do a presentation (20 minutes + 5 minutes for questions) - All of you will switch off and present sections of the final presentations, but everyone else gets to ask anyone in your team questions about each section. The presentation days are April 27 and May 2; we will draw lots for the slot sign-up.

6. You get to do a write-up (1000 - 1500 words, i.e. 2-3 pages, + 1 page for figures and captions). The final write-up and link to the Github with the pushed notebook(s) are due 5 pm on May 7.

7. Use jupyter as we've been doing throughout class, but this means that final presentations are more conversational than formal final presentations - think more like journal club

8. For the write-ups, teams are strongly encouraged to use latex / Overleaf. Please send a pdf file to the course staff by the due date.

9. Your whole team gets the same grade - this might be frustrating if you are used to individual accomplishment, but science doesn't actually work that way anymore.

10. If you are having issues with your team members, please first try to work it out amongst yourselves; if this is not successful, please contact the course staff.

Required Texts/Materials

Primary Textbooks

There is no required textbooks for the course, although we list suggested options below that we expect to pull optional reading from.

Supplementary Textbooks

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

- Python Data Science Handbook, J T. VanderPlas

- Mining of Massive Datasets, J. Leskovec, A. Rajaraman, and J. Ullman, Cambridge University Press, 2014.

- Introduction to Data Mining, P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, 2019.

- Pattern Recognition and Machine Learning, C. Bishop

- Worked notebooks from the LSST Data Science Fellowship Program.

Course Requirements and Grading

Course Policies and Procedures

Special Needs

Any students with special learning needs must contact their professor during the first two weeks of class.

Student Mental Health Services

As a student you may experience a range of issues that can cause barriers to learning, such as strained relationships, increased anxiety, alcohol/drug problems, feeling down, difficulty concentrating and/or lack of motivation. These mental health concerns

Material	Total Points	% of Grade
Final Project Total	400	40% **See Note Below!**
Problem Sets - **DUE SATURDAY NIGHTS MIDNIGHT**	300	30%
Class Participation (showing up to $\geq 80\%$ of classes will give you full credit)	100	10%
Mid-terms	2 @ 100	20%
Total for the Course	1000	100%

NOTE! In order to receive a passing grade in the class you must earn at least 50% of the total available lab points (120/240) AND at least 50% of the total available class points (65/130). In addition, you must take both exams and turn in the final project.

Grading will be assigned approximately as follows based on past experience: A: 900 - 1000; B: 800 - 899; C: 700 - 799; D: 600 - 699; F: 0 - 599 (You must receive a "C-" or better to receive a grade of "S.")

For the final project, 10% of the project will be graded on the project plan / statement of work, 30% on the project presentation, and 60% on the project itself.

Keep copies of all materials upon which you are graded (homework, final project materials, and examinations) until the end of the semester. Grades for each assignment, lab, and exam will be posted to the Canvas site as soon as scoring has been completed. Students are expected to review their grade summaries for accuracy periodically during the semester and after the final project. Discrepancies should be reported to Prof. Coughlin.

or stressful events may lead to diminished academic performance or reduce a your ability to participate in daily activities. University of Minnesota services are available to assist you with addressing these and other concerns you may be experiencing. You can learn more about the broad range of confidential mental health services available on campus via the Student Mental Health Website at http://www.mentalhealth.umn.edu

Academic Standards

The scholastic conduct and classroom procedures of the Office of Community Standards will be followed. You are responsible for being familiar with these. Students are welcome to work together, exchange ideas, etc. For the on-line assignments, you must log in individually, and provide your own answers, even if you talk things over with another student.

Tentative Course Schedule

Class Dates	Topic	Due Dates
Jan 19	First steps, crash course in python	**No HW**
Jan 24, 26	Probability distributions, Astrophysics Datasets	HW 1
Jan 31, Feb 2	Statistical Inference - Classical, Bayesian	HW 2
Feb 7, 9	Statistical Inference - MCMC, Sampling, Statistical Inference - Posterior-predictive checks	HW 3
Feb 14, 16	Time Series - Introduction, Time Series Analysis - Fourier Series and Fourier Transform, Lomb-Scargle	HW 4
Feb 21, 23	Time Series Analysis - Variability and Modeling, Time Series Analysis - Stochastic and Autoregressive Processes	HW 5
Feb 28, Mar 2	ML - Gaussian Processes (intro)	**Mid-Term Exam 1 Feb 28 - Mar 5**
March 14, 16	ML - ML - Gaussian Processes (continued), ML - Trees / Regression	HW 6
March 21, 23	ML - Dimensionality Reduction, Principal Component Analysis	HW 7
March 28, 30	ML - Clustering, ML - Anomaly Detection	HW 8
April 4, 6	ML - Classification metrics, ML - Bayesian Classification, ML - Deep Learning (intro)	HW 9
April 11, 13	ML - Deep Learning (continued)	**Midterm Exam 2**, **April 11 - 15**
April 18, 20	Databases	–
April 25, 27	Databases (continued), Project Presentations	–
May 2	Project Presentations	–
May 4	**Final Project Due - 5 pm**	