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

Prof. Michael Coughlin, Dr. Michael Steinbach

Spring 2021

E-mail: cough052@umn.edu, stei0062@umn.edu
Class Hours: Mon/Wed 14:30-16:25 pm
Office Hours: Wed 11-12 (Michael S.); Fri 1:00-2:00 (Nico); Fri 3:30-4:30 (Michael C.)
Website: https://github.com/mcoughlin/ast8581_2021_Spring

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

Mid-Term 1: March 1-5 (Take home exam; replacement for homework)

Mid-Term 2: April 26-30 (Take home exam; replacement for homework)

Final Project

Monday, Feb. 15 by 2:30 pm (start of class) Part I: 1-page (double-spaced) project proposal due to my email. The project can be in any area of astrophysics, as long as data analysis (and a lot of it) is involved. For the proposal, it should explain why your topic is of interest, including the main goals/motivating questions that caused you to choose it. It should also explain how your proposed project is appropriate for demonstrating the skills learned in this class. For criteria of success, it should explain what "success" looks like, or if you want, what you expect an "A"-level project would be.

Monday, March 15 - Monday, May 3 during class Part II: 15 minute presentation (with 5 minutes for questions) to the class. The presentation should explain the project in a clear, logical and organized way, such that the listener should be able to easily follow. In the question section, the speaker should be able to respond and summarize when needed. For those demonstrating an easily digestible analysis, I might suggest showing off preliminary versions of these, if they exist.

Friday, May 7 by 5 pm Final Project: project (content or some link to it is fine) due to my email.

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

Material	Total Points	% of Grade
Final Project Total	400	40% **See Note Below!**
Problem Sets - **DUE FRIDAY 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 proposal, 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 examination. Discrepancies should be reported to Prof. Coughlin.

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 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.

Homework

Late Work: You are allowed two 48-hour late passes to use as you see fit on any homework assignment, barring the Final Project and Exams. You do not need to email me – just write at the top of the problem set that you intend to use one of them. Submissions within 15 minutes of the due date are acceptable for "tech fudge time."

Extensions for projects work best with a conversation, so please talk to me if you will have a conflict with a larger assignment. *Reach out in advance, day-of requests will not be accommodated outside of truly exceptional circumstances.* Swing by office hours to talk about work arounds or send me an email. Deadlines are set to try to avoid work piling up in this class, but sometimes there are other conflicts, and I will work with you to ameliorate the situation given enough notice.

Tentative Course Schedule

Class Dates	Topic	Due Dates
Jan 20	First steps, crash course in python	**No HW**
Jan 25, 27	Intro to Big Data and Machine Learning	HW 1
Feb 1, 3	Probability distributions, Astrophysics Datasets	HW 2
Feb 8, 10	Statistical inference - Classical, Databases - Overview	HW 3
Feb 15, 17	Statistical inference -Bayesian, MCMC, Databases - SQL	HW 4 Project Proposal due Feb 17 by class
Feb 22, 24	Time Series Analysis - Introduction	HW 5
March 1, 3	Time Series Analysis - Periodicity	**Mid-Term Exam 1 April 5**
March 8, 10	ML - Clustering	HW 6 Project Presentations Begin
March 15, 17	ML - Regression and Model Fitting	HW 7
March 22, 24	ML - Gaussian processes	HW 8
March 29, 31	ML - Dimensionality Reduction	HW 9
April 12, 14	ML - Classification	HW 10
April 19, 21	ML - Intro to Recommender Algorithms / Graphic Models	HW 11
April 26, 27	Visualization; Outliers, imbalanced, and missing data	**Mid-Term Exam 2 April 29**
May 3	Special Topics	–
May 7	**Final Project Due - 5 pm**	

Frequently Asked Questions:

Q: What if I miss a lecture? Things go by too fast for me to write everything down.

A: PDFs of all lectures will be available on the Canvas site, although they will differ somewhat from what is presented in lecture. You can print out pdf versions of them beforehand to take notes on, minimizing the amount you have to write.

Q: I think that the lectures are going too fast.

A: Studies have shown that when students read the textbook before the corresponding lecture, problems with the lecture going too fast evaporate.

Q: Where are the exams given?

A: The midterm exams are performed as Take Homes in lieu of homework for that week.

Q: What do I do if I can't make it to an exam?

A: Contact me as soon as possible. All makeups require my written permission before the exam.

Q: Are the midterms graded on a curve?

A: For each midterm you are given a certain number of points. At the end of the course, these points are all added up and you are assigned a grade. Thus, there is no assignment of grades for each midterm.

Q: How can I find out my exam and lab scores? My on-line scores?

A: All scores will be posted on Canvas.

Q: All of my assignments are given points, not grades. How can I estimate my grade in the course?

A: Add up all of your grades, divide by the total available at any point in the course, and compare to the preliminary grade distribution given in your syllabus.