CS 506 - A1

Computational Tools for Data Science

Course Repository

Location: CAS B12
Time: M/W 4:30-5:45

Instructor	**Teaching Fellow**	**Teaching Fellow**
Lance Galletti galletti@bu.edu **Office Hours:** W 6-7PM MCS 203	Saurav Chennuri (A3 & A5) saurav07@bu.edu **Office Hours:** Wed 3-4 PM 730 Commonwealth 302	Muhammad Fahad Farid (A2 & A4) fahadf@bu.edu **Office Hours:** Mon 6-7 PM 730 Commonwealth 302

Overview

The goal of this course is to provide students a hands-on understanding of classical data analysis techniques and to develop proficiency in applying these techniques in modern programming languages (Python) while also learning about the social and ethical challenges of collecting and mining data by studying real world examples.

The course introduces students to a wide range of techniques that are commonly used in the analysis of data, such as clustering, classification, regression, and network analysis. Broadly speaking, the course breaks down into three main components, which we will take in order of increasing complication: (a) unsupervised methods; (b) supervised methods; and (c) methods for structured data.

Lectures will present the fundamentals of each technique and aim to help students understand the practical settings in which these theoretical/analytical methods are useful. In class, we will also study use cases and go over relevant Python packages that will enable students to perform hands-on experiments with their data. Class discussion will, for the most part, be extended office hours, review, or extra coding exercises. However, **this is not a Python course**, so self-study will be necessary for those students who do not already know the language.

Prerequisites

Students taking this class **must** have some prior familiarity with programming at the level of CS 105, 108, or 111, or equivalent. CS 132 or equivalent (MA 242, MA 442) is **required**. CS 112 is also helpful.

Workload

There will be 6 homework assignments, 1 midterm exam, and 1 final project (no final exam)

Homework

The **homework** assignments will be due throughout the semester as relevant material is covered. There will be 6 assignments in total. The homework **instructions are intentionally vague -** you must ask clarifying questions (in class or on Piazza) or make reasonable assumptions and justify your decisions.

Late Policy

You are allowed a **one-time** late homework with no penalty (up to 3 days) for the semester. Otherwise late homeworks will not be accepted.

Collaboration

You may discuss questions but you must submit individual code. You must list your collaborators in the homework.

Midterm

The **midterm** will be a Kaggle Data Science competition among the students in the class with a live leaderboard. Students will need to submit predictions based on a training dataset and a report detailing the methods used and decisions made. 80% of the grade will be based on the report and 20% will be based on the competition score related to the quality of the predictions made.

Final Project

Teams can have **3-5 students**. These will be assigned based on a project preference form that you will be asked to submit after Pitch Day.

Spark Project

BU Spark! offers students an opportunity to work on technical projects provided by companies or organizations in the Greater Boston area through our experiential learning lab (X-Lab). For this semester, Spark! has partnered with CS506 to offer a diverse selection of external data science projects scoped to support the course's learning outcomes and enhance the student experience. To learn more about Spark!, please visit their website: https://www.bu.edu/spark/

For students working with Spark! on projects for external partners, your project team will be led by one of the Spark! Project managers: Hong Xin, Grace Yoon, Della Lin, Savannah Majarwitz, Hope Ruse, Anqi Lin, and John Chai.

Spark! projects are a great opportunity for students to get real-world project experience to highlight on their github and CV. These projects have already been curated and will be presented during "Pitch Day". Project descriptions will be made available at the start of the semester.

See project expectations below.

Individual Projects

Students may choose their own projects but **must submit a project proposal by Feb. 2nd** (see Piazza Resources tab for a Project Proposal Template). You must share this project proposal in the form of **a google doc** with comment permissions to all the instructors. You must address all comments and suggestions before the proposal can be approved. Every project should contain: Data gathering / aggregating, Data processing, Visualization, and Data analysis.

Final Project Grading

Grading will be based on the deliverables and your performance in your team **throughout the semester.** See below on project expectations.

Grading

20% midterm
40% assignments
40% final project
5% extra credit

Letter Grades

A:	95% and above
A-:	90% - 95%
B+:	87% - 90%
B:	83% - 87%
B-:	80% - 83%
C+:	77% - 80%
C:	73% - 70%
C-:	70% - 73%
D:	60% - 70%
F:	below 60%

Expectations

Project Expectations

Please read through How to Navigate Group Projects.

Spark Help / Consultations

https://www.bu.edu/spark/resources/consultations/

Extra Credit

Extra credit can be earned by *consistently*:

- Submitting weekly project notes / updates to your final project repository on GitHub.
- Asking and answering questions on Piazza.
- Submitting PRs to our class repository with code or class notes.
- Contributing to our class repository by fixing typos, providing clarification edits etc.

Emails

Do not expect direct messaging from emails or private Piazza questions and allow for up to 24h.

Re-Grades

If you notice an issue with a grade you've received, you must email your TA **within 48h** of receiving this grade. Anything beyond 48h will not be accepted for a re-grade.

Availability

If you cannot attend Office Hours please email me with your availability and I will send you a calendar invite to meet remotely.

Tools / Platforms

We will be using:

1. Piazza for questions
2. GitHub and GradeScope for homeworks, midterm, and project submission
3. Kaggle will be used for the midterm

Note: we are not using blackboard.

Getting Started Checklist

1. Join Piazza: piazza.com/bu/spring2022/cs506a1
2. Create a GitHub account: https://github.com/
3. Create a Kaggle account: https://www.kaggle.com/
4. Fill out this form with your GitHub and Kaggle account username
5. Install Python: https://www.python.org/about/gettingstarted/
6. Sign up for GradeScope: https://www.gradescope.com/courses/356949 (code: 2RXY6J)

Get Ahead Checklist

1. Complete the Git Workshop: CS506-Spring2022/tree/master/00-git
2. Review: https://hmc-cs-131-spring2020.github.io/howtos/assignments.html

Course Schedule

Date	Lecture	Assignments
01/24	Overview + Effective Programming	
01/26	Git	
01/31	Intro to Data Science + Distance & Similarity	
02/02	Spark Pitches	Individual (non-Spark) Project Proposal due (share google doc with instructors and TAs) ~~Git Fundamentals due~~

02/07	Clustering I (k-means)	Homework 1 due
02/09	Clustering II (hierarchical clustering) + Clustering III (Density Based Clustering)	
02/14	Clustering III cont’d + Probability Review	
02/16	Clustering IV (Soft Clustering + Clustering Aggregation)	
02/21	NO CLASS	Homework 2 due
02/22	MONDAY SCHEDULE Singular Value Decomposition	
02/23	Classification I (Learning from data + Overfitting & Underfitting + KNN)	
02/28	Classification III (Naive Bayes + Decision Trees)	Homework 3 due Project Deliverable 0 due
03/02	Classification III (SVM + Bagging & Boosting) + Recommender Systems	
03/14	Midterm Help	Midterm START Homework 4 due
03/16	Regression I (Linear Regression)	
03/21	Regression II (Logistic Regression)	Midterm END midnight
03/23	Regression III (Recap + Model Evaluation)	Project Deliverable 1 due
03/28	Gradient Descent	
03/30	Neural Networks I	Project Deliverable 2 due Homework 5 due
04/04	Guest Lecture TBD	
04/06	Neural Networks II	
04/11	**Early insight presentations**	Project Deliverable 3
04/13	Guest Lecture (Day in the Life of Red Hat Data Scientists) - 4 speakers	
04/18	NO CLASS	Homework 6 due
04/20	Networks & Graphs	
04/25	Why Data Privacy Matters	Project Deliverable 4

04/27	EXTRA TOPICS	
05/02	**Final Project Presentations**	
05/04	**Final Project Presentations**	Final Project Report due

EXTRA TOPICS: time permitting, could include Fourier Analysis, Differential Privacy, Computational Learning Theory, more in depth Neural Networks and other advanced topics.

Academic Honesty

You may discuss homework assignments with classmates, but you are solely responsible for what you turn in. Collaboration in the form of discussion is allowed. Please be sure to include the list of your collaborators in the homework you turn in. All forms of cheating (copying parts of a classmate's assignment, plagiarism from books or old posted solutions) are NOT allowed. We -- both teaching staff and students -- are expected to abide by the guidelines and rules of the Academic Code of Conduct (which is at http://www.bu.edu/academics/policies/academic-conduct-code). Graduate students must also be aware of and abide by the GRS Academic Conduct code at http://www.bu.edu/cas/students/graduate/forms-policies-procedures/academic-discipline-procedures.