Introduction to Data Science (Fall 2024)

CS 1090a, AC 2090a, Stat 109a, or CSCI E-109a

Instructors

Pavlos Protopapas (SEAS) and Natesh Pillai (Statistics)

Preceptor

Chris Gumb (SEAS)

Meeting Times

Lectures: Mon & Wed 9am-10:15am [Science Center, Hall B]*

Labs: Friday 9am-10:15am [Science Center, Hall B]*

**Due to high registration numbers, the course has been moved to the Cambridge campus (hence the earlier start time of 9am).*

Office Hours: TBD

Course Introduction

Welcome to CS1090a/AC209a/STAT109a, also offered by the DCE as CSCI E-109a, Introduction to Data Science. The course will focus on the analysis of messy, real-life data to perform predictions and inferences using statistical and machine learning methods.

Material covered will integrate four key facets of an investigation using data:

1. data wrangling - web scraping, data cleaning;
2. exploratory data analysis – generating hypotheses and building intuition;
3. prediction, inference, or statistical learning; and
4. communication – justifying decisions and analyzing results.

This course is the first in a two-part series. In the spring semester, the curriculum builds upon the content of the fall course, diving deeper into unsupervised learning, deep neural networks for computer vision and language modeling, transformers, generative models, and Bayesian inference. Students are strongly encouraged to enroll in both the fall and spring courses within the same academic year.

Prerequisites

A foundational knowledge of Python programming is required for this course. Specifically, students should be comfortable with the following Python concepts:

- Conditionals, loops, and data structures (e.g., lists, dictionaries, etc.)
- Functions
- File I/O and string parsing
- Classes, methods, attributes, and general OOP principles

Additionally, students should possess an understanding of the following:

- Random variables and common probability distributions, such as the normal and binomial distributions.
- Basic concepts in probability (e.g., independence, joint and conditional probabilities, etc.).
- Calculus at an introductory level. (Note: Multivariable calculus is not required.)

Topics

- Web scraping
- Manipulating tabular data: Pandas
- Exploratory data analysis (EDA)
- kNN & linear regression
- Multiple & polynomial regression
- Model selection & cross-validation
- Regularization (LASSO and Ridge)
- Maximum likelihood estimation (MLE)
- Bootstrap, confidence intervals, & hypothesis testing
- Missing data & imputation
- High dimensionality & principal component analysis (PCA)
- Classification & logistic regression
- Decision trees
- Ensemble methods: bagging, random forest, & boosting
- Causal inference

Differences Between CS1090A & AC2090A

- **Readings:** There may be further readings assigned to 209A students.
- **Assignments:** Homework assignments for 209A students may have addition components.
- **Exams:** The quiz questions and coding portions of the two midterms will differ slighly between 1090A and 2090A students.
- **Projects:** Project groups with one or more 209A students will perform some self-directed outside reading to inform their projects. The goal is to make use of some approach or method not explicitly covered in class and communicate an understanding of the method(s) and its applicability to the problem of focus through the final report and presentation.

NOTE: Unlike previous years, there will be no separate advanced sections for 209A students.

Course Components

Lectures and labs will be live-stream for Extension School students and can be accessed through the 'Zoom' section on Canvas.

Recordings will be made available to all registered students with 24 hours and can be accessed through the 'Course Videos' section on Canvas.

Lecture

The class meets for lectures twice a week (Mon & Wed). Attending and participating in lectures is a crucial component of learning the material presented in this course. Students may be asked to complete short readings before certain lectures. On-campus students are expected to attend all lectures and labs. For every 8 classes attended, on-campus students earn an extra late day. See the homework section below for more information on late days.

Lab

Lab will be held every Friday and will present deep-dives into the software libraries used to implement the methods described in lecture. They prepare students to excel on the homework assignments, group projects, and coding portions of the exams. Attendance is required for on campus students.

Readings Lectures and labs may assign readings to be completed either before or after class. Pre-class readings bring important context and framing for the day's class while post-class readings dive in deeper.

Readings may include review of mathematical concepts, articles discussing machine learning models, ethical case studies, and code tutorials.

Students will be responsible for roughly one hour of reading per week. Additional reference may also be provided but will be marked as 'bonus' readings.

Quizzes A short quiz will be released after each lecture. They will cover content from the lecture as well as any assigned pre-class readings.

Quiz solutions will be revealed after the following lecture.

Quizzes provide feedback on students' current level of understanding and prepares them for the types of questions that will appear on the multiple choice component of the midterm exams.

The lowest 1/3 of quiz grades will dropped.

Exercises Exercises are short Jupyter notebooks containing discussion and code demonstrating ideas recently introduced in class. They may focus on algorithms, visualization, or features of useful libraries like Numpy, Pandas, Matplotlib, Seaborn, and Scikit-learn.

Exercises are hosted on Ed. Many exercises have fill-in-the-blank sections with test cases you can run your notebook against to assess one's understanding.

Exercises are only included in the final grade calculation if they would improve the grade. Otherwise, the weighting of the exercises is shifted onto the quizzes (i.e., from 8% to 10%).

Homework

Homework 0 will be released sometime in June and will be due on the 11th of September. It serves to evaluate students' preparedness for the course, covering the prerequisites listed above. Students finding this material too challenging should consider registering for the course in a future semester, after having spent more time with the prerequisites.

After homework 0, there will be 6 homework assignments, each due two weeks after being assigned. For these assignments, students have the option to collaborate and submit their work in pairs. Please consult the collaboration policy below.

Midterms

The course will hold two midterms. Each midterm consists of an in-class quiz, and a coding take-home assignment.

Important Dates: - Midterm 1: - Quiz: Friday October 18th, normal class time - Coding Take-home: Released Friday October 18th 12 pm; Due Sunday October 20th 10 pm - Midterm 2: - Quiz: Wednesday December 11th, normal class time - Coding Take-home: Released December 11th 12pm; Due Friday December 13th 10 pm

Quiz Component Similar to the lecture quizzes in format, with each covering content from the first and second halves of the course respectively.

Coding Take-Home Similar to the exercises but a larger scale challenge.

Project

Students may propose a project topic or adapt an example provided by the teaching staff. Students self-assemble into groups and submit their preferences over the approved set of project topics. Groups are assigned a project and a mentor from the teaching staff who will provide guidance and feedback along the remaining milestones.

Milestones

0. Project proposals
1. Preferences & group formation
2. Data check, refined problem statement, and course-of-action
3. Data processing, EDA, and baseline analysis
4. Minimum viable project
5. Final Presentations & Report

Details regarding each of the milestones will be released during the first few weeks of the course. A rubric will be made available to the students outlining expectations for the final presentation and report.

Grading

Final grades for the course will be computed using the following weights:

Assignment	Final Grade Weight
Homework: 0	1%
Homework: 1-6	34%
Lecture Quizzes	8%
Exercises	2%
Exams: Quiz	20%
Exams: Coding	10%
Project: Milestones 1-4	10%
Project: Presentation Video	5%
Project: Final Report	10%
Total	**100%**

This course uses the grading system outlined in the FAS Student Handbook:

A, A– Earned by work whose excellent quality indicates a full mastery of the subject and, in the case of the grade of A, is of extraordinary distinction.

B+, B, B– Earned by work that indicates a good comprehension of the course material, a good command of the skills needed to work with the course material, and the student's full engagement with the course requirements and activities.

C+, C, C– Earned by work that indicates an adequate and satisfactory comprehension of the course material and the skills needed to work with the course material and that indicates the student has met the basic requirements for completing assigned work and participating in class activities.

D+, D, D– Earned by work that is unsatisfactory but that indicates some minimal command of the course materials and some minimal participation in class activities that is worthy of course credit toward the degree.

E Earned by work that is unsatisfactory and unworthy of course credit toward the degree.

Numerical scores in the class will be converted to letter grades at the end of the course by the instructors.

Resources

Recommended Textbook

The book for the course is *An Introduction to Statistical Learning* and is available at statlearning.com.

Software

Our primary software tools will be Python 3, Jupyter notebooks, and various 3rd party Python libraries. A set-up guide will be included in the to-be-released HW0. SEAS also provides the FASOnDemand service which is accessible through Canvas. This a remote Python environment where students can run Jupyter notebooks. The environment has all packages used in the course pre-installed.

Late Work Policy

Extension School Late Days **Extension School** students are allocated a total **4 late days** with **at most 2 days applied to any single homework**.

On-Campus Students Late Days **On-campus students** students are initially allocated a total of **3 late days**. with the possibility of acquiring more through attendance (see attendance policy below). **At most 2 late days can applied to any single homework**.

General Late Day Policies If a student has exhausted all their late days, late homework will not be accepted unless there is a medical (if accompanied by a doctor's note) or other official, University-excused reasons. There is no need to ask before using one of your late days.

Late days cannot be applied to quizzes, exercises, midterm components, or project milestones.

Attendance Policy

Attendance at lectures and labs is required for all on-campus students**. The teaching staff will record on-campus attendance. For every 8 sessions attended (i.e., lecture or lab), on-campus students will earn 1 additional late day.** Any effort to misrepresent attendance will be considered a violation of the honor code and be dealt with accordingly.**

Academic Integrity

We expect you to adhere to the Harvard Honor Code at all times. Failure to adhere to the honor code and our policies may result in serious penalties, up to and including automatic failure in the course and reference to the ad board.

DCE Academic Integrity Policy

If you are an Extension School student, you are responsible for understanding Harvard Extension School policies on academic integrity (https://extension.harvard.edu/for-students/student-policies-conduct/academic-integrity/) and how to use sources responsibly. Stated most broadly, academic integrity means that all course work submitted, whether a draft or a final version of a paper, project, take-home exam, online exam, computer program, oral presentation, or lab report, must be your own words and ideas, or the sources must be clearly acknowledged. The potential outcomes for violations of academic integrity are serious and ordinarily include all of the following: required withdrawal (RQ), which means a failing grade in the course (with no refund), the suspension of registration privileges, and a notation on your transcript.

Using sources responsibly (https://extension.harvard.edu/for-students/support-and-services/using-sources-effectively-and-responsibly/) is an essential part of your Harvard education. We provide additional information about our expectations regarding academic integrity on our website. We invite you to review that information and to check your understanding of academic citation rules by completing two free online 15-minute tutorials that are also available on our site. (The tutorials are anonymous open-learning tools.)

Student Collaboration

If you work with a partner on an assignment make sure both parties solve all the problems. Do not divide and conquer. You are expected to be intellectually honest and give credit where credit is due. In particular:

- if you work with a fellow student and want to submit the same notebook you need to form a group prior to the submission. Details in the assignment. Not all assignments will permit group submissions.
- you need to write your solutions entirely on your own or with your collaborator
- if you worked with a fellow student on a paired assignment but decide in the end to submit different notebooks individually, include the name of the other student as a comment at the top of your notebook.
- you are welcome to take ideas from code presented in labs, lecture, or sections but you will need to change it, adapt it to your style, and ultimately write your own. Simply copying verbatim will rarely be successfully.
- if you use code found on the internet, books, or other sources you need to cite those sources.
- you should not view any written materials or code created by other students for the same assignment.
- you may not provide or make available solutions to individuals who take or may take this course in the future. If you are using a remote git repository such as GitHub to work on your assignments **you must make it private.**

Use of AI Models

Purpose of Policy: This policy outlines the acceptable use of AI models, including but not limited to ChatGPT, in completing assignments for this course.

Policy Guidelines:

1. **Original Work:** Students are expected to complete assignments using their original thoughts and interpretations. AI models can be used to help understand concepts, generate ideas, or learn about different perspectives, but they should not write or complete assignments for students.

2. **Collaboration with AI:** Students may use AI models for brainstorming or generating preliminary ideas, but the final work submitted must be substantially their own. Students should be able to explain their reasoning, logic, and conclusions without relying on the model's output.

3. **Restrictions for Specific Assignments:** There may be specific assignments (e.g. quiz part of the midterms) or parts of the course where the use of AI models is entirely prohibited. These restrictions will be clearly stated in the assignment guidelines.

4. **Ethical Considerations:** Students are encouraged to approach the use of AI with ethical considerations in mind, including issues related to privacy, bias, and authenticity.

Consequences for Non-Compliance: Failure to adhere to this policy may result in academic penalties as outlined in the course's academic integrity policy.

Questions and Clarifications: If students have questions about the appropriate use of AI models in an assignment, they should consult the course instructor or teaching assistants before proceeding.

Please refer to the University's policy for further information.

Accommodations for Students with Disabilities

Harvard students needing academic adjustments or accommodations because of a documented disability must present their Faculty Letter from the Accessible Education Office (AEO) and speak with the professor by the end of the second week of the term, (fill in specific date). Failure to do so may result in the Course Head's inability to respond in a timely manner. All discussions will remain confidential, although Faculty are invited to contact AEO to discuss appropriate implementation.

Harvard Extension School is committed to providing an inclusive, accessible academic community for students with disabilities and chronic health conditions. The Accessibility Services Office (ASO) https://www.extension.harvard.edu/resources-policies/accessibility-services-office-aso offers accommodations and supports to students with documented disabilities. If you have a need for accommodations or adjustments in your course, please contact the Accessibility Services Office by email at accessibility@extension.harvard.edu or by phone at 617-998-9640.

Diversity and Inclusion Statement

As educators, we aim to build a diverse, inclusive, and representative community offering opportunities in data science to everyone. We will encourage learning that advances ethical data science, exposes bias in the ways data & data science can be (and all too frequently is) used, and advances research into fair and responsible data science.

We need your help to create a learning environment that supports a diversity of thoughts, perspectives, and experiences, and honors your identities (including but not limited to race, gender, class, sexuality, religion, ability, etc.) To help accomplish this:

- If you have a name and/or set of pronouns that differ from those in your official Harvard records, please let us know!
- If you feel like your performance in the class is being impacted by your experiences outside of class, please do not hesitate to come and talk with us. We want to be a resource for you. Remember that you can also submit anonymous feedback (which will lead to us making a general announcement to the class, if necessary, to address your concerns). If you prefer to speak with someone outside of the course, you may find helpful resources at the Harvard Office of Diversity and Inclusion.
- We (like many people) are still learning about diverse perspectives and identities. If something was said in class (by anyone) that made you feel uncomfortable, please talk to us about it.
- As a participant in course discussions, you are expected to respect your classmates' diverse backgrounds and perspectives.

Our course will discuss diversity, inclusion, and ethics in data science. Please contact us (in person or electronically) or submit anonymous feedback if you have any suggestions for how we can improve.

For additional resources, guidance, and support related to diversity and inclusion, please refer to the Harvard Office for Equity, Diversity, Inclusion, & Belonging.

Auditing

To request to audit the course, send an email to cs1090a2024@gmail.com with your HUID (required) and a statement of agreement to the terms below. **Note:** Please make sure you are not currently enrolled in the course when you send your request. You can't be added as an auditor in Canvas if you are currently listed there as an enrollee.

All auditors must agree to abide by the following rules:

- Auditors must attend class in person. This is a Harvard policy. Auditors who do not confirm their presence during the first week of in-class instruction will lose course access.
- Auditors are held to the same standard of academic honesty as our registered students. Please do not share homeworks or solutions with anyone. Violations will be reported to the Harvard Administrative Board.
- Auditors are not permitted to take the course for credit in the future.
- Auditors should **not** submit HWs or midterms, or participate in projects.
- Auditors should refrain from using any course and TF resources that are designed for our registered students like Ed, FASOnDemand, and office hours.