

CDS DS 122

Foundations of Data Science III

Spring 2023

Lecture Meeting Place: Kenmore Classroom Building (KCB) 101

Meeting Time: MWF 09:05 am – 09:55 am

Discussion Meeting Place: A2: Fuller Building (FLR) 121

A3-A5: ENG Product Innovation Center (EPC) 206

Meeting Times: A2: M 1:25 pm – 2:15 pm, A3: M 2:30 pm – 3:20 pm,
A4: M 3:35 pm – 4:25 pm, A5: M 4:40 pm – 5:30 pm

Instructor: Prof. Lisa Wobbes

- **Office Hours:** See calendar on Piazza
- **Office Hours Location:** CCDS 1401
- **Email:** edwobbes@bu.edu

Teaching Assistants: Anush Veeranalala, Benjamin Cox

- **Office Hours:** See calendar on Piazza
- **Office Hours Location:** See calendar on Piazza
- **Email:** jun@bu.edu, benrcox@bu.edu

Course Assistants: Leo Perez, Qingyuan Kong

- **Office Hours:** See calendar on Piazza
- **Office Hours Location:** See calendar on Piazza
- **Email:** leoperez@bu.edu, qykong@bu.edu

Overview of the Course

CDS DS 122 is the third in a three-course sequence (with CDS DS 120 and CDS DS 121) that introduces students to theoretical foundations of data science. DS 122 covers topics in probability (including common probability distributions, conditional probability, independence, Bayes theorem, prior and posterior distributions, sampling, and the central limit theorem), statistics (including maximum likelihood), and basic numerical optimization (including gradient descent methods). Knowledge of a programming language (such as Python) is expected. Effective Spring 2022, this course fulfills a single unit in each of the following BU Hub areas: Quantitative Reasoning II, Critical Thinking.

Prerequisites

DS 120 is *required*, and DS121 is *strongly recommended*. Prior knowledge of Python is *required*.

Getting Set Up

You will need to set up access to the following (online) materials. Instructions for how to do all of those setups are below.

Required: Python on your laptop for homeworks.

Required: Piazza for discussion of assignments and course material.

Required: Gradescope for assignment submission.

Textbook: No textbook is required for this course! The online lecture notes are available at <https://mcrovella.github.io/DS122-Foundations-of-Data-Science-III>

Programming Environment

We will use `python` as the language for teaching and for assignments that require coding. You are expected to know python and to use it for all coding assignments.

Setup: Instructions for installing and using Python are on Piazza.

Piazza

We will be using Piazza for class discussion. The system is well tuned to getting you help fast and efficiently from classmates, and instructors. Rather than emailing questions to the teaching staff, please post your questions on Piazza. We will also use Piazza for distributing materials such as homeworks and helpful resources.

When someone posts a question on Piazza, if you know the answer, please go ahead and post it. However, please *don't* provide answers to homework questions on Piazza. It's OK to tell people *where to look* to get answers, or to correct mistakes; just don't provide actual solutions to homeworks.

Setup: Our class Piazza page is at <https://piazza.com/bu/spring2024/ds122/info>. If you registered before the semester started, you should have been automatically enrolled. If you are adding the class late, go to Piazza at that link and enroll yourself. If you have any problems, please contact a TA or a CA.

Gradescope

Assignments will be submitted via Gradescope (<https://www.gradescope.com/>). Graded assignments will be returned to you via Gradescope as well.

Setup: If you registered before the semester started, you should have been automatically enrolled. If you are adding the class late, go to Gradescope at the link above, and enroll yourself using the entry code XX5WP3. If you have any problems, please contact a TA or a CA.

Homeworks

1. There will be 10 homework assignments. The top 9 grades will be used to compute the total homework credit.
2. Homeworks will be assigned on Tuesdays.
3. Homeworks are due at 09:00 am on the following Tuesdays.
4. You can discuss homeworks in section meeting on Mondays. But don't expect that TAs will be going into detail – instead, they will answer specific questions!
5. You need to consistently work the problem sets each week. Plan to set aside a regular time each week to do them.
6. Some homeworks will contain bonus questions. You must earn full participation credit to activate the bonus points.
7. Homeworks will be submitted via *Gradescope*. See the next section.
8. **IMPORTANT:** You are responsible for submitting homework electronically on Gradescope by the stated due date and time. Assignments will be accepted on Gradescope up to 24 hours late for a 20% grade reduction. Later assignments will not be accepted.

Submitting Homework

For showing your analytical / mathematical work, there are three options, in increasing order of quality:

1. You can scan handwritten notes into PDF. Note that these must be **clear** and **neat** because the grader will simply read them as best they can – if the grader cannot understand your handwriting easily, you may lose points on the assignment. If you use this option, you can scan from your mobile device if it comes out clearly enough. There are instructions on Piazza for how to scan and submit your homework via Gradescope.
2. You can write up your work in Word, using the built-in equation editor for the mathematics. Then save as PDF, and follow the same instructions for how to submit to Gradescope. Added benefit: no trees are destroyed.
3. You can learn and use \LaTeX . This is the tool that produces a professional, publishable PDF document. It is what hardcore computer scientists use. You can learn to use it quickly – I recommend starting with the cloud based system called Overleaf at

https://www.overleaf.com/learn/latex/Learn_LaTeX_in_30_minutes/.

If you want to install L^AT_EX on your own computer (to use offline, for example) there are instructions at <http://www.latex-tutorial.com/>. Eventually you will find it useful for lots of your coursework (L^AT_EX is required for DS 320!), so it makes sense to learn it now.

For submitting the code portions of the homeworks, you will use Gradescope as well. These can be submitted in Jupyter Notebook format. Your TAs are a great resource for making sure you are up to speed on Jupyter Notebooks.

Course and Grading Administration

Final grades will be computed based on the following:

50% Homework assignments.

5% Participation.

10% Midterm 1.

10% Midterm 2.

25% Final (Cumulative).

Participation

To earn full participation credit, you need to attend at least 11 out of 14 discussion labs. However, up to 2 out of 11 discussion labs can be replaced by active participation on Piazza, in class, or by actively helping other students in office hours.

Final

I will determine grade cutoffs after all assignments and exams have been graded. Grade cutoffs will take into account my assessment of the difficulty level of the assignments and exams, and my assessment of what is expected for each letter grade.

Regrade policy

You have the right to request a regrade of any homework or exam question. All regrade requests must be submitted via Gradescope, and must describe a factual error in our assessment. If you request a regrade for one question, then we have the right to review the entire homework or exam. Beware that this may potentially result in a lower grade.

Important Dates

- Midterm 1: March 1.
- Midterm 2: April 12.

The final exam will take place between May 6 and May 10. Mark this week in your calendars!

Readings

You are expected to actively use the lecture notes throughout this course. In general, the topics in the course schedule directly correspond with the chapters in the textbook.

Office Hours

There are 8 office hours each week. The schedule for office hours is on Piazza.

Accommodations for Students with Documented Disabilities

If you are a student with a disability or believe you might have a disability that requires accommodations, please contact the Office for Disability Services (ODS) at (617) 353-3658 or access@bu.edu to coordinate any reasonable accommodation requests. ODS is located at 25 Buick Street on the 3rd floor.

Course Feedback

Your feedback about any material you would like to cover or suggestions for how to improve the course is welcome at any time. All suggestions will be considered, though it cannot be guaranteed that the feedback will be incorporated. Please send your comments as a private Piazza note to Prof. Lisa Wobbes.

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, but 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/dos/policies/student-responsibilities/>).

You can probably, if you try hard enough, find solutions for homework problems online. Given the nature of the Internet, this is inevitable. Let me make a couple of comments about that:

1. If you are looking online for an answer because you don't know how to start thinking about a problem, talk to a TA, a CA, or myself, who may be able to give you pointers to get you started. Piazza is great for this – you can usually get an answer in an hour if not a few minutes.
2. If you are looking online for an answer because you want to see if your solution is correct, ask yourself if there is some way to verify the solution yourself. Usually, there is. You will understand what you have done *much* better if you do that.
3. If you are looking online for an answer because you don't have enough time and are getting close to the assignment deadline, think about this:
 - (a) what you are doing is intellectually dishonest,
 - (b) you are going to have to solve problems like this on the midterm and final,

- (c) you can drop one homework without penalty,
- (d) you can compensate for the lost points by submitting bonus questions.

So ... it would be better to simply submit what you have at the deadline (without going online to cheat) and plan to allocate more time for homeworks in the future. We care more about making an honest attempt on the homework than the final solution being exactly right.

Course Schedule

Week	Topics	Assigned	Due
3	Introduction		
4	Probability Review Distributions	HW 1	
5	Continuous Distributions Joint Distributions Functions of Random Variables	HW 2	HW 1
6	Frequentism Central Limit Theorem and Sampling Confidence	HW 3	HW 2
7	Hypothesis Testing Multiple Hypothesis Testing and FDR Parameter Estimation	HW 4	HW 3
8	02/19: No class; Discussions on 02/21 Parameter Estimation	HW 5	HW 4
9	Review 03/01: Midterm 1		HW 5
10	Bayes's Theorem Bayes's Theorem for Distributions Estimating Proportions		
11	No class		
12	Estimating Counts Poisson Processes Bayesian Testing	HW 6	
13	Bayesian Comparisons and Classification Bayesian Inference	HW 7	HW 6
14	Conjugate Priors Markov Chains MCMC	HW 8	HW 7

See next page.

Week	Topics	Assigned	Due
15	Review 04/12: Midterm 2		HW 8
16	04/15: No class Hidden Markov Models	HW 9	
17	Hidden Markov Models in NLP Gradients Gradient Descent	HW 10	HW 9
18	Improving Gradient Descent Review	HW 10	