CDS DS 122
Foundations of Data Science III
Fall 2023

Lecture Meeting Place: Sargent College (SAR) 101

Meeting Time: MWF 10:10 am – 11:00 am

Discussion Meeting Place: College of Arts and Sciences (CAS) 315

Meeting Times: M 1:25 pm – 2:15 pm, 2:30 pm – 3:20 pm, or 3:35 pm – 4:25 pm

Instructor: Prof. Pawel Przytycki

- **Office Hours:** See calendar on Piazza.
- **Office Hours Location:** CCDS 1539
- **Email:** `pawel@bu.edu`

Assistant Instructor: Prof. Lisa Wobbes

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

Teaching Assistants: Navya Jain (lead), Abhishek Malakar

- **Office Hours:** See calendar on Piazza.
- **Office Hours Location:** CCDS 5th floor
- **Email:** `jnavya@bu.edu`, `amalakar@bu.edu`

Course Assistant: Peiyang Liu

- **Office Hours:** See calendar on Piazza.
- **Office Hours Location:** CCDS 5th floor
- **Email:** `nickpliu@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, TFs, 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 pleased *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/fall2023/ds122`*. If you registered before the semester start, 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 CA.*

Gradescope

Assignments will be submitted via Gradescope (`https://www.gradescope.com/`). Graded assignments will be returned to you via Gradescope as well. If you have any questions about the grading you receive on Gradescope, please contact a TA.

Setup: *If you registered before the semester start, 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 WB3K7N. If you have any problems, please contact a TA or CA.*

Homeworks

1. Homeworks will be assigned on Wednesdays.
2. Homeworks are due at 10:10 am on the following Wednesday. This means they are due before the start of Wednesday's class.
3. 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!
4. Homeworks will be submitted via *Gradescope*. See the next section.

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 LaTeX 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 (LaTeX 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 will be distributed as Jupyter Notebooks which can then be saved as PDFs once the code has been run. Your TAs are a great resource for making sure you are up to speed on Jupyter Notebooks.

Course and Grading Administration

NOTE: IMPORTANT: Late homework assignments **WILL NOT** be accepted. However, your final grade will be based on the top 8 homeworks submitted (out of 10).

Final grades will be computed based on the following:

50% Homework assignments. The top 8 homework grades (out of the 10 assigned) will be used to compute this score.

5% Participation (Including in class, on Piazza, in discussion section, or helping other students in office hours. Submitting corrections to the book will also count toward participation)

10% Midterm 1.

15% Midterm 2.

20% Final (Cumulative).

The exact cutoffs for final grades will be determined after the class is complete.

You need to consistently work the problem sets each week. Plan to set aside a regular time each week to do them.

Office Hours

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

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 TF 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, and

   (c) you can miss up to two homeworks without penalty.

   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

Date	Topics	Assigned	Due
9/6	Introduction		
9/8	Probability Review		
9/11	Probability Review		
9/13	Distributions	H1	
9/15	Joint Distributions		
9/18	Frequentism		
9/20	Functions of Random Variables	H2	H1
9/22	Central Limit Theorem and Sampling		
9/25	Confidence		
9/27	Hypothesis Testing I	H3	H2
9/29	Hypothesis Testing II		
10/2	Multiple Hypothesis Testing and FDR		
10/4	Parameter Estimation (Estimators)	H4	H3
10/6	Parameter Estimation (Model Fitting)		
10/9	**No Class**		
10/10	Parameter Estimation (MLE) **(Substitute Monday)**		
10/11	Review Day		H4
10/13	**Midterm 1**		
10/16	Bayes's Theorem		
10/18	Bayes's Theorem For Distributions	H5	
10/20	Estimating Proportions		
10/23	Estimating Counts		
10/25	Poisson Processes	H6	H5
10/27	Bayesian Testing		
10/30	Bayesian Comparisons and Classification		
11/1	Bayesian Inference I	H7	H6
11/3	Bayesian Inference II		
11/6	Conjugate Priors		
11/8	Markov Chains	H8	H7
11/10	MCMC in Theory		
11/13	MCMC in Practice		
11/15	Review Day		H8
11/17	**Midterm 2**		
11/20	Bayesball; **No Discussion Sections**		
11/22	**No Class; Holiday**		
11/24	**No Class; Holiday**		
11/27	Hidden Markov Models I		
11/29	Hidden Markov Models II	H9	
12/1	Gradients and Gradient Descent		
12/4	Improving Gradient Descent		
12/6	Newton's Method I	H10	H9
12/8	Newton's Method II		
12/11	Review Day		
12/12	**No Class; Homework Due By Midnight**		H10