CAS CS391 (future CS365)

Foundations of Data Science

Spring 2021

This semester the course is offered in the LfA format, for details please see the appropriate sections below.

TODO:
please sign up for the course Piazza and Gradescope here:
Piazza: `piazza.com/bu/spring2021/cs391/home`
Gradescope: `https://www.gradescope.com/courses/232562` code: ERV7B2

Course staff

Instructor: prof. Dora Erdos

Office: 111 Cummington Mall MCS214
Office Hours (via Zoom): Mon 9:30-11 am ET, Wed 3-4 pm ET
Email: edori@bu.edu

Teaching Fellow: Hannah Catabia

Office Hours (via Zoom): Mon 3:30-4:30 pm ET, Tues 2:30-4:30 pm ET
Email: catabia@bu.edu

Overview of the Course

This course lays the foundation towards more advanced data-intensive classes, such as Data Science, Machine Learning, Data Mining. The course provides an understanding of the fundamentals and the practical implications of concepts. It covers both theoretical skills as well as working/practical knowledge. It introduces to students to key concepts that are necessary when taking more advanced data-science classes. Therefore, the goal of the class is to provide an understanding of the fundamentals and the practical implications of concepts such as: tools from probability and statistics, operations on feature spaces, Markov Chains and random walks, learning from data, classification as well as data-management techniques such as basics of SQL etc.

Broadly speaking, the course breaks down into four main components, which we will take in order of increasing complication: (a) fundamental concepts (b) unsupervised methods (c) supervised methods (d) data-management fundamentals.

Prerequisites: Students taking this class **must** have taken CS 112, CS 131 (MA293), CS 132 (MA242) and CS 237 (MA581) or equivalent. Or must obtain the consent of the instructor. CS 330 is recommended.

Note, that for CS majors CS391 counts towards the **group D** requirements.

Learning Outcomes

Students who successfully complete this course will be proficient in data acquisition, manipulation, and analysis. They will have good working knowledge of the most commonly used methods of clustering, classification, and regression. They will also understand the efficiency issues and systems issues related to working on very large datasets.

Readings

We will loosely follow the material from the book Foundations of Data Science by Avrim Blum, John Hopcroft, and Ravindran Kannan. The book is available online here: `https://www.cs.cornell.edu/jeh/book.pdf`.

We may assign readings from other resources as well in which case the link or pdf will be clearly posted on Piazza.

Course Format

The course meets twice a week for lectures and once for lab. The lectures are offered in the LfA format; on any given day you have a choice to either attend in person or join via Zoom. You don't have to inform the course staff which option you choose. The lectures and labs will also be recorded and posted online. (Zoom links will be posted on Piazza.)

Lectures:
Tues, Thur 11-12:15 ET in CAS216 or Zoom

Labs:
A2: Wed 8-8:50 ET
A3: Wed 9:05-9:55 ET
A4: Wed 10:10-11 ET

Course communication

We will be using **Piazza** for class discussion.

All relevant course material and links (e.g. links to Zoom, slides, recordings) will be posted here.Use Piazza to ask questions on any topic; course material, homework and lab problems, logistics. We strongly prefer communicating via Piazza over emails to the course staff.

When someone posts a question on Piazza, if you know the answer, please go ahead and post it. However, *don't* provide full solutions to homework assignments on Piazza. It's OK to tell people *where to* *look* to get answers, or to correct mistakes. If in doubt about posting a certain content publicly, you can create a private post only visible to you and the instructors.

Sign up to Piazza here: `piazza.com/bu/spring2021/cs391/home`

Course staff will also be holding office hours via Zoom.

Homework Assignments

There will be one homework assignment **per week**. A typical assignment will consist of solving one paper-and-pencil type problem as well as a practical assignment to be completed in Python. This latter will be some application of techniques discussed in class and applied to some data. Homework solutions will be posted the day after the homework deadline.

Homework will be submitted via **Gradescope**.
Sign up Please use your real name so that we know which student the submission belongs to. `https://www.gradescope.com/courses/232562` code: ERV7B2

grading and submission policy: At the end of the semester the two lowest homework grades will be dropped. If you don't submit an assignment, it will be counted as 0 points and will be one of the dropped grades. In light of this drop policy late assignments will **not** be accepted. Be mindful that sometimes it' s fine to submit a partial assignment if you weren't able to complete everything by the deadline.

deadline: Homework is due every week **Tuesday 11:59 pm ET** (Boston time!). No late assignments will be accepted.

regrade policy: If after carefully reading the posted solutions and your answer you believe that we have made an error in grading please submit your regrade request via Gradescope.

Grading

There will be two exams, a midterm (most likely **Thursday March 4th**, but subject to change) and a final exam (during the assigned final exam slot). Your grade will be based on the two exams as well as the homework assignments.

Final grades will be computed based on the following:

30%	Homework assignments.
30%	Midterm
40%	Final

Academic Honesty

studying together and collaboration on homework: You are encouraged to form study groups and discuss course content with your peers. You may discuss homework assignments with classmates, but first you must give it at least 45 minutes of thought by yourself. However, the homework submission has to be your **individual work**. You have to write down the solutions or write code on your own, without consulting others. The teaching staff can ask you at any point to explain your solution to them.

You are explicitly **forbidden to commit plagiarism** of any form. This includes copying parts of a classmate's assignment, plagiarism from books or online resources or old posted solutions.

Collaboration of any form on exams is strictly forbidden and will be prosecuted at all times.

We – both teaching staff and students – are expected to abide by the guidelines and rules of BU's **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/`.

Course Schedule (tentative)

Date	Topics
Week 1	Probability and statistics (review). Random variables, conditional, prob. distributions, expected value, variance, confidence intervals, bounds.
Week 2	Week 1 topics continued. Fitting distribution to real life data. EM-clustering.
Week 3	Notion of learning/fitting data, parameters. predictors, hypothesis.
Week 4	Feature space, linear algebra refresher. Matrix operations, decompositions.
Week 5	Dimensionality reduction. Low rank matrix approximation. SVD and PCA.
Week 6	SVD and PCA applications. midterm.
Week 7	Fitting to data, linear regression, gradient descent, SGD. mathematical tools.
Week 8	Classification, model evaluation, ML examples.
Week 9	Various ML algorithms (e.g. k-nn, bayes, decision trees). Bias-variance trade off.
Week 10	Clustering (k-means, dbscan).
Week 11	Walks and graph structure, mtx operations on graphs. Rnd walks on graphs, Markov chains. Page rank.
Week 12	Graph generation models. Fit graph models to data
Week 13	Relational data bases, relational model.
Week 14	Review