CAS CS 506
Computational Tools for Data Science
Fall 2017

Lectures: Tuesdays and Thursdays, 3:30–4:45 PM in Photonics 211
Lecture schedule: `https://tinyurl.com/cs506-spring2018`
Piazza site: `https://piazza.com/bu/spring2018/cs506`
Lecture materials: `https://github.com/adamdavisonsmith/BU-CS506-Spring2018`
Instructor: Adam Smith

- **Office:** MCS-135F
- **Office Hours:** Tuesdays 11:00AM–12:30pm, Thursdays 12:00PM–1:30PM
- **Email:** ads22@bu.edu (but use Piazza)

Teaching Fellow: Ms. Sofia Nikolakaki

- **Office Hours:** TBD.
- **Office Hours Location:** Undergrad Lab, EMA 302
- **Lab Tutoring Hours:** TBD.
- **Email:** smnikol@bu.edu (but use Piazza)

Co-instructors (projects) : Dora Erdos (`edori@bu.edu`), Ziba Cranmer (`zcranmer@bu.edu`)

Action items you should complete today:

- Read the syllabus
- Send Sofia your github account name (you may need to create one)
- Add yourself to the course Piazza site
- Fill out the course survey (see Piazza for link)
- Install python, git (see notes for Lecture 02)

Overview of the Course

This course is targeted at students who require a basic level of proficiency in working with and analyzing data. The course emphasizes practical skills in working with data, while introducing students to a wide range of techniques that are commonly used in the analysis of data, such as clustering, classification, regression, and network analysis. The goal of the class is to provide to students a hands-on understanding of classical data analysis techniques and to develop proficiency in applying these techniques in a modern programming language (Python).

Broadly speaking, the course breaks down into three main components, which we will take in order of increasing complication: (a) unsupervised methods and summaries; (b) supervised methods; and (c) methods for structured data.

Lectures will present the fundamentals of each technique; the focus is helping students (a) understand the practical settings in which these methods are useful and (b) interpret the results (and assess the significance) of their analyses. Class discussion will study use cases and will go over relevant Python packages that will enable the students to perform hands-on experiments with their data.

Prerequisites

Students taking this class **must** have

- *Experience programming*, at the level of CS 105, 108, or 111, or equivalent.
- A solid understanding of linear algebra. CS 132 or equivalent (MA 242, MA 442) is **required**. In particular, students should be comfortable with the notions of linear independence, rank, eigenvalue and eigenvector.
- *Probability and statistics*. We will assume familiarity with basic concepts of probability (indendence, random variables, expectation, variance) and statistics (point estimates, regression, confidence intervals, hypothesis tests).
- *General scientific mathematics*. We will assume familiarity with basic concepts of probability (independence, random variables, expectation, variance) and statistics (point estimates, confidence intervals, hypothesis tests), as well as calculus.

Other useful background:

- Data structures and algorithms (CS 112, 131)
- Probability for computer scientists (CS 337)
- Vector calculus

Learning Outcomes

Students who successfully complete this course will be proficient in basic data acquisition, manipulation, and analysis. They will be able to understand and carry out the most commonly used methods of clustering, classification, and regression. They will be able to interpret their results, and understand the limitations of their methods. They will also understand and be able to articulate efficiency and systems issues related to working on very large datasets.

Readings

There is no text for the course. Lecture notes will be posted online.

Some of the lectures are based on *Introduction to Data Mining,* by Tan, Steinbach and Kumar. This is a good place to go for more detail if something is not clear.

Some other recommended texts:

- *Python for Data Analysis* (`http://shop.oreilly.com/product/0636920023784.do`). This is the definitive text for *Pandas* which we will use quite a bit.
- Larry Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, Springer, 2004.
- *Programming Collective Intelligence* (`http://shop.oreilly.com/product/9780596529321.do`)

Web Resources

Many of the slides for the course are actually executable python scripts, using the `jupyter notebook`. You can download and execute the lectures on your own computer, and you can modify them any way you'd like, play around with them, experiment, etc.

The slides I use in lecture are published on `github`. The repository is `https://github.com/adamdavisonsmith/BU-CS506-Spring2018`. If you want to access the repository using `git`, please feel free. If you find a bug, submit a pull request (alternatively, ask about it on Piazza.)

Homeworks and Project

1. There will seven to nine homework assignments. In a typical assignment you will analyze one or more datasets using the tools and techniques presented in class.

   Homeworks will be submitted via `github`. For this, we need your github account (create one if you don't already have it). After you have created it, fill out the form at **TBD** (was `https://goo.gl/forms/8W0SOdvMn07UKdip2`) to let us know what it is.

   You are expected to work individually on homeworks.

2. There will be a final project. For the project you will extract some knowledge or conclusions from the analysis of dataset of your choice. The analysis will be done using a subset of the methods we described in class. The final project will require a proposal, two progress reports, and a final presentation in poster form.

   The project will have three essential components: 1) a data collection piece (which may involve crawling or calls to an API, combining data from different sources etc), 2) a data analysis piece (which will involve applying different techniques we described in class for the analysis) and 3) a conclusion component (where the results of the data analysis will be drawn). The students will submit a 5-page report explaining clearly all the three components of their project. Finally a poster presentation will be required where the students will be prepare to present their effort and results in front of their poster.

   As an example, you may choose to collect data from Twitter related to a specific topic (e.g., Ebola virus) and then measure the intensity of posts about a topic in different areas of the world. Other examples of projects may include (but are not limited to): analysis of MBTA data, analysis of census data from your favorite country, crawling of YouTube (or other social media data) and analysis of social behavior like trolling and bullying.

   The main project report is due in early April; a revised report is due at the end of April. See the lecture schedule page for exact dates. The project presentations will be given in the form of a final poster explaining components 1, 2 and 3 of the project.

   You are expected to work in teams of two on the final project. I will leave it up to you to form teams on your own, but everyone must work in a team.

Piazza

Piazza is a website that allows you to ask questions, either to instructors or course-wide. We will be using Piazza for almost all course communication outside of the classroom. Please sign up, and set appropriate email notification options so that you make sure to receive announcements.

`https://piazza.com/bu/spring2018/cs506`.

Piazza allows you to ask questions that are visible only to instructors, but it also allows you to ask questions to the entire class, and answer others' questions. When someone posts a question on Piazza, if you know the answer, please go ahead and post it. However please *do not* 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. Also, be polite. See the post "Ethics and Etiquette on Piazza" for more detail.

Programming Environment

We will use `python` as the language for teaching and for assignments that require coding. Instructions for installing and using Python are on Piazza.

Course and Grading Administration

Homeworks are due at 7pm on Fridays. Assignments will be submitted using `github`. Ms. Nikolakaki will explain how to submit assignments.

IMPORTANT: Late assignments **WILL NOT** be accepted. However, you may submit **one** homework up to 3 days late. You **must** email Ms. Nikolakaki before the deadline if you intend to submit a homework late.

Final grades will be computed based on the following:

50% Homework assignments.

50% Final Project

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

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 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/`.

You can probably, if you try hard enough, find solutions for homework problems online. That said,

1. If you are looking online for an answer because you don't know how to start thinking about a problem, talk to Ms. Nikolakaki 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. We will enforce the collaboration policy.

Course Schedule

We will maintain a detailed lecture schedule here:
`https://tinyurl.com/cs506-spring2018`
Topics (tentative):

- Introduction
  - Python and essential tools (Git, Jupyter Notebook, Pandas)
  - The process of “data science”
  - Probability and Statistics Refresher
  - Linear Algebra Refresher
- A sampling of techniques
  - Distance and Similarity Functions, Timeseries
  - Clustering
  - Assessing significance: correlations and clustering
  - Singular Value Decomposition and Dimension reduction
  - Web Scraping
  - Classification and Regression
- Interpretation, assessment, and confidence
  - statistical validity
  - p-hacking and multiple hypothesis testing
- Ethics of data: privacy, transparency, accountability, representation, fairness
- Advanced technical topics
  - Parallel architectures and Map Reduce
  - Collaborative filtering and recommender Systems
  - Analyzing Graphs and Networks