CAS CS 506
Computational Tools for Data Science
Spring 2018

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`

Name	Role	Email (Use Piazza if possible)	Office	Office Hours
Adam Smith	Instructor	`ads22@bu.edu`	MCS 135F	Tue 11:00am–12:30pm, Thu 12:00pm–1:30pm
Sofia Nikolakaki	TF	`smnikol@bu.edu`	MCS B08	Tuesday 1:00-2:30pm, Wednesday 1:00-2:30pm
Ziba Cranmer	Spark! coordinator	`zcranmer@bu.edu`		
Dora Erdos	Co-instructor (projects)	`edori@bu.edu`		

Action items you should complete today:

- Read the syllabus
- Sign and return academic honesty policy page
- Add yourself to the course Piazza site and set notification preferences
- Make sure you have a github.com account

Over the next few days:

- Register for Top Hat (instructions forthcoming on Piazza)
- Install python and git (see notes for Lecture 02)
- Fill out the background survey (to be posted on Piazza, due by Tuesday, January 23)
- Get started on HW 0!

Overview of the Course

This course is targeted at students who require a basic level of proficiency in working with and analyzing data. 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). Lectures aim to help students (a) understand the practical settings in which these methods are useful and (b) interpret the results (and assess the significance) of their analyses.

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*. Familiarity with basic concepts of probability (indendence, random variables, expectation, variance) and statistics (point estimates, regression, confidence intervals, hypothesis tests).
- *General scientific mathematics*. Calculus, elementary combinatorics, boolean logic.

Other useful background:

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

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 discuss the limitations of their methodology. They will also understand and be able to articulate efficiency and systems issues related to working on very large datasets.

Slides

There is no text for the course. Lecture notes will be posted online. Many of the slides 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.)

Additional Reading

Some other recommended texts:

- *Introduction to Data Mining,* by Tan, Steinbach and Kumar. Some of the lectures are based on this book.
- *Python for Data Analysis* (`http://shop.oreilly.com/product/0636920023784.do`). This is the definitive text for *Pandas*.
- Larry Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, Springer, 2004. A survey of major statistical concepts, written for scientists outside of statistics.
  We will post links to additional reading on Piazza as the smester proceeds.

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, you will need a github account (create one if you don't already have it—they are free).

You are expected to work individually on homeworks.

2. There will be a final project to be conducted in teams. 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. That report will be revised based on feedback from course staff. 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.

   See the lecture schedule for project deadlines.

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 detail about our expectations.

Top Hat

TopHat is a software platform for me to ask you multiple-choice questions in class. It provides a similar functionality to "clickers". I use Top Hat to help students participate, and to gauge whether the pace of the class is right. There is a subscription fee of about $26 (but no textbooks to buy). We will post details of how to sign up for this class on Piazza.

Course and Grading Administration

45% Homework assignments
5% Class participation (in-class discussion, Piazza, Top Hat)
50% Final Project

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

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.

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

Lecture Schedule

We will maintain a detailed lecture schedule here:

`https://tinyurl.com/cs506-spring2018`

Tentative topics:

- 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

CS 506 Academic Honesty Policy—Sign and Return this Page

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/`.

I, the undersigned, have read and understand the academic honesty policy.

Signature: _____

Date: _____

Your name (print): _____

BU ID number: _____

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

1. It is against the course honesty policy (see above).
2. 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.
3. 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.