CS S109A

Introduction to Data Science

Syllabus - Summer 2020

Welcome to S109A, Introduction to Data Science. This course is the first half of a one-year introduction to data science. The course focuses on the analysis of messy, real life data to perform predictions using statistical and machine learning methods.

The material of the course is divided into 3 modules. Each module will integrate the five key facets of an investigation using data:

1. data collection - data wrangling, cleaning, and sampling to get a suitable data set;
2. data management - accessing data quickly and reliably;
3. exploratory data analysis - generating hypotheses and building intuition;
4. prediction, statistical learning, and inference; and
5. communication - summarizing results through visualization, stories, and interpretable summaries.

The list of topics includes data exploration and visualization, *k*-NN, linear regression, LASSO and Ridge, Cross-validation, logistic regression, discriminant analysis, decision trees, random forests, boosting, stacking, neural networks, principal components analysis, imputation, and experimental design (AB testing).

Course Logistics

Prerequisites

You are expected to have programming experience at the level of CS 50 or above, math understanding at the level of Math 1a and 1b, and statistics knowledge at the level of Stat 100 or above (Stat 110 recommended). HW0 is designed to test your knowledge on the prerequisites. Successful completion of this assignment will show that this course is suitable for you. HW0 will only be graded for completion, but you are required to submit.

Accessibility

The Summer School is committed to providing an accessible academic community. The Accessibility Office offers a variety of accommodations and services to students with documented disabilities. Please visit http://www.summer.harvard.edu/resources-policies/accessibility-services for more information.

Academic Integrity

You are responsible for understanding Harvard Summer School policies on academic integrity (http://www.summer.harvard.edu/policies/student-responsibilities) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting the wrong draft, or being overwhelmed with multiple demands are not acceptable excuses. To support your learning about academic citation rules, please visit the Resources to Support Academic Integrity (http://www.summer.harvard.edu/resources-policies/resources-support-academic-integrity) where you will find links to the Harvard Guide to Using Sources (https://usingsources.fas.harvard.edu) and two free online 15-minute tutorials to test your knowledge of academic citation policy. The tutorials are anonymous open-learning tools.

Course Components

Lectures

The class consists of two weekly lectures and one lab, which is designed as a class activity. They will be live-streamed on Zoom feed and recorded (recording will be available within 24 hours) . We will have quizzes and coding exercises after each lecture online to assess and challenge your understanding of the material and to help us identify gaps.

Lecture will be broken into three or four 45 to 55 minute blocks (with ~5 minute break in between): two blocks will consist of standard lecturing from slides on Zoom, and the third and fourth blocks will be a Jupyter Notebook coding exercise (done in break-out rooms). This structure may vary slightly from lecture to lecture.

Labs

Attendance to labs is optional but **strongly encouraged**. Labs are designed as hands-on in-class activities. The instructor will go over practice problems similar to the homework problems and review difficult material. Lab will be held Fri 12-2pm.

Office Hours (Zoom & Calendly)

Check pinned post on Ed for current OH times

Assignments

There will be an initial self-assessment homework called HW0 and 5 more graded weekly homework assignments. You will be working in Jupyter Notebooks which you can run on your own computer. **HW 1-5 can be done in pairs (optional); HW0 is individual. HW0 will be published on June 12.**

Quizzes

Quizzes and short coding exercises will be available right after lecture and be available for 36 hours (the content will be based on what was discussed in lecture). 1/3 of the quizzes will be dropped from your grade. *Note: coding exercises will be treated as separate quizzes. Some lectures will have both.

Final Exam

There will be an individual open-notes, take-home final exam due Mon, Aug 3.

Recording

Lectures and labs will be live-streamed, and will be recorded and made available within 24 hours later via Canvas (typically much quicker turn around then that).

Recommended Textbook

An Introduction to Statistical Learning by James, Witten, Hastie, Tibshirani.

The book is available here. There will be assigned readings from the text leading up to each lecture:

Free electronic version: http://www-bcf.usc.edu/~gareth/ISL/ (Links to an external site).

HOLLIS: http://link.springer.com.ezp-prod1.hul.harvard.edu/book/10.1007%2F978-1-4614-7138-7

Amazon: https://www.amazon.com/Introduction-Statistical-Learning-Applications-Statistics/dp/1461471370 (Links to an external site).

Course Policies

Getting Help

For questions about homework, course content, package installation, JupyterHub, and after you have tried to troubleshoot yourselves, the process to get help is:

1. Post the question in **Ed** and hopefully your peers will answer. Note that in Ed questions are visible to everyone. You can access Ed from the Navigation bar on the course Canvas page (note: you must be logged into Canvas).

2. Go to **Office Hours**: this is the best way to get individualized help.

3. For private matters send an email to the **Helpline**: [s109a2020@gmail.com]. The Helpline is monitored by all the teaching staff.

4. For personal and private matters send an email to the **instructor and head TF**.

Questions on Graded Homework and Regrading Policy

We take great care in making sure all homework are graded properly. However if you feel that your assignment was not fairly graded you may:

1. Contact the grader by emailing the helpline with subject line "Regrade HW1: Grader=johnsmith" within 2 days.

2. If still unhappy with the initial response, then submit a reason via email to the Helpline with subject line "Regrade HW1: Second request" within 2 days of receiving the initial response. Note: once regrading is done, you may receive a grade that is higher or lower than the initial grade.

Late Day Policy

You are allowed up to 2 days of late homework submissions, maximum of 1 day on any single assignment, no questions asked (this does not applky to the final exam). No homework will be submitted more than 24 hours late. Solutions will be posted one day after the due date. Late homework submissions will not be accepted after 24 hours past the due date. If you exceed your 2 late days, 1 point (20%) will be deducted for late days after that. Late minutes count as a whole day, e.g. if you submit 30 minutes late, this will count as a 1 day.

Communication from Staff to Students

Class announcements and official communication from staff will be through **Canvas**. All homework and quizzes will be posted and submitted in Canvas.

MAKE SURE you have your settings set so you can receive emails from Canvas. No official communication or announcements will be done via Ed.

Submitting an assignment

You are to work all homework in a Jupyter Notebook. When you are done, convert your notebook in a pdf and submit **both** the .ipynb file and the .pdf file. You can submit multiple times up to the deadline.

You are encouraged but not required to submit in pairs. We will be using the Groups function in Canvas to do this, details to be announced later.

All assignments will due on Tuesdays at 11:59pm in Canvas and will be posted one week in advance.

Collaboration Policy

We encourage you to talk and discuss the assignments with your fellow students (and on Ed), but you are not allowed to look at any other students assignment or code outside of your pair. Discussion is encouraged, copying is not allowed.

Grading Guidelines

Homework will be graded based on 1) how correct your code is (the Notebook cells should run, we are not troubleshooting code), 2) how you have interpreted the results - we want text not just code, it should be a report, and 3) how well you present the results. The scale is 1-5.

Software

We will be using Jupyter Notebooks, Python 3 and various python modules. You can access the notebook viewer either in your own machine by installing the Anaconda platform (Links to an external site) which includes Jupyter/IPython as well all packages that will be required for the course, or by using the SEAS Jupyter Hub from Canvas. Details in class.

Course Grade

Your final score for the course will be computed using the following weights:

Homeworks 60%

Quizzes and Coding Exercises 15%

Final Exam 25%

Total 100%

A final letter grade will be given in accordance with the Summer School's grading policy: https://www.summer.harvard.edu/resources-policies/grades