Sample Syllabi – Subject to Change
Pathways in Data Science

Instructors: Phillip Lo

There is no course textbook, or any other course materials you need to purchase. Course notes will be posted after every lecture.

Grading Scheme: Homework will be assigned daily, and the third week of the course will consist of a final project.

Accommodations: If you require any accommodations for the course, please contact Student Disability Services (https://disabilities.uchicago.edu/) as early as possible. Please feel comfortable reaching out to one of the instructors as well, so that we can work out an accommodation plan as early as possible.

Course Description:
In a world increasingly driven by data, *data science* has emerged as an all-encompassing term to describe the process of collecting, processing, analyzing, and making decisions with data. Being able to use and understand the tools needed to engage with an ever-exploding amount of digital data ensures that data-driven decisions are made as soundly as possible.

This course is structured to first introduce methods from a mathematical perspective, and then to apply these methods to real datasets using Python. The methods introduced will fall under the umbrella of *machine learning,* which is the mathematical and statistical study of algorithms that use data to make predictions. Understanding the mathematics of popular data science tools is an important way to ensure they are used correctly. Then, understanding how to collect real data, process it into a format that can be used by machine learning methods, and then actually implement the methods will help bridge the gap between the empirical world of data and the theoretical world of mathematics.

Course Objectives: By the end of the course, we will

- Understand of the goals of machine learning and what it can and cannot do
- Be able to describe, both in theory and practice, a few fundamental algorithms used in machine learning and be able to apply them to data sets
- Be able to obtain and work with real-world data sets in Python using pandas and numpy and visualize results using matplotlib/seaborn/plotly
- Apply machine learning techniques to glean insights from data sets
- Evaluate data insights and be able to draw and report conclusions from them
- Understand the role of ethics, privacy and fairness in data science

- Perform a real-world data analysis of your choosing to consolidate all of the above

Mathematical/Programming Prerequisites: Data science and machine learning lie in the intersection of mathematics and programming, and as such, require some prerequisites in both. It is strongly recommended that you are familiar with the following material at the start of the class. We will send out review material a few weeks before the start of the course for those unfamiliar with the content, and we will spend a day or two reviewing them very quickly at the beginning, but you should not rely on this as your first encounter with these topics:

- Computing derivatives in single-variable calculus, finding the maxima/minima of single-variable functions (if you have taken AP Calculus, that's more than enough)
- Vectors, how to add/subtract them, scalar multiplication, lengths/norms of vectors
- Basic conditional probability
- Familiarity with at least one programming language, *preferably Python*; **facility in programming is required**

Course structure: The first session (the "theory" section) will be focused on the mathematical aspects of data science, particularly the way certain machine learning algorithms work and the mathematics behind them. The second session (the "practice" session) will focus more on the actual implementation of these algorithms, addressing the questions of how we can take these algorithms and put them into practice, as well as issues of experimental design. Of course, the line between theory and practice is quite blurry, so the content of the two halves of the course will be closely coordinated, and constantly refer to each other.

The last week of the course will involve students working in small groups to design, and carry out, and write up a study of their own. Presentations will be given on the last day.

Course Calendar: The following calendar is provisional and almost certainly subject to change.

- Introduction to the course/logistics, overview of data science
- Review of prerequisite calculus/vectors, Python review
- Linear regression
- Empirical risk minimization, estimation and approximation error, overfitting, cross validation
- Trees, random forests, bagging/boosting
- Classification algorithms, K-nearest neighbors, normalizing/cleaning data
- Clustering algorithms, data visualization
- Ethics, fairness, and privacy in data science
- Buffer day
- Final projects and presentations