CSC 599.70: Introduction To Data Science

Grant Long

Fall 2019

E-mail: itds.ccny@gmail.com
Office Hours: By Appointment
Office: StreetEasy, 130 Fifth Ave

Web: **grantmlong.com**
Class Hours: Monday 6:30-9pm
Class Room: NAC 5/110

Course Description

This course consists of a survey of analytical tools and concepts in data science, with goal of equipping students with an understanding of the best practices used by professional data scientists and analysts in top companies in technology, finance, and media. The course begins with an overview of fundamentals in data handling and exploratory data analysis, followed by an introduction to core concepts in statistical modeling and machine learning, and concludes with a brief introduction to advanced concepts in data science.

Students will work with a wide variety of real world data sets throughout the course in order to gain hands on experience. Emphasis will be placed on frequent practice through writing and reviewing code each week. In addition, students will be assigned and expected to discuss short reading assignments ranging from academic reviews of popular topics in analytics as well as data science and engineering blog posts from companies such as Airbnb, Facebook, and Spotify. Tasks and readings will aim to demystify the work of data teams in the real world, and familiarize students with the concepts and resources needed to secure and succeed in analytical roles.

Prerequisites/Corequisites

Prerequisites: Intro to Programming (CSc102/103) or equivalent and Probability and Statistics (CSc217). The course assumes profiency in basic programming paradigms, data structures, and statiscal concepts. Students will also need a basic profiency in the *Python* computing language to be acquired independently in the early weeks of the course.

Course Objectives

Through this course, students should be able to:

1. Explain the key steps in a data science project.
2. Apply Python to load, clean, and process data sets.
3. Identify key elements of and patterns in a data set using computational analysis and statistical methods.
4. Explain and visualize empirical findings using with Python and other resources.
5. Explain fundamental principles of machine learning.
6. Apply predictive algorithms to a data set.
7. Work effectively in a team dedicated to analyzing data.

Course Policies

Grading

Element	**Weight**
Group Project	30%
Homework	30%
Midterm Exam	30%
Attendance, Quizzes, & Class Participation	10%

Project

The bulk of the course grade will be a group project that will be due in advance of the final class on December 9. Students will be expected to work on the project throughout the semester and will be required to provide evidence of their progress. Grades will be assigned on the basis of overall project quality, demonstration of core principles taught in the class, and individual contributions to the group's effort. More details on the course project will be announced in the first weeks of the course.

Assignments

This class includes short, frequent assignments to provide necessary background for the course lectures. All assignments will be graded on a 10-point scale; most will be assigned through DataCamp.

- **Late assignments**. Assignments not turned in by the set deadline are eligible to be completed for half credit by the final class on December 9. Exceptions will be granted only as mandated by CUNY policy.

Exam

A short midterm exam will be held in November and will focus on broad concepts the course has surveyed thus far. The format will mimic the style of questions frequently asked in interviews for data-related roles.

Attendance, Quizzes, & Participation

Students are expected to attend class and be active participants in discussion. This includes, but is not limited to, discussing assigned readings and videos and sharing ideas during classroom exercises. The instructor may give impromptu quizzes to assess attendance and collect feedback, though these will not generally be designed to test comprehension.

Deadlines

Projects and homeworks must be turned in on time, with exceptions and extensions only granted in extraordinary circumstances as outlined by College policy. Students are expected to use their ability to drop the lowest two homeworks and quizzes judiciously.

Resources

Students are expected to bring a wifi enabled computer to each class. If a student is unable to bring a laptop to class on a regular basis, they should contact the professor as soon as possible. In addition, students should have:

- A Github account (free account ok).
- A DataCamp account. Educational access is available free of charge and will be provisioned by the second week of class.
- Access to a cloud-based Jupyter notebook service. Examples include Google Colaboratory (preferred), Binder, and Microsoft Azure.
- *Required*: Access to a computer with a internet connection during class. If you do not have access to a computer for class, reach out to the professor as soon as possible.
- *Recommended*: Access to a computer with a standard data science stack installed, including Anaconda Python 3.6 or greater and Jupyter.

Recommended Texts and Materials

- **Required Text**: *Data Science from Scratch*, Joel Grus. 2nd Edition, May 2019 (O'Reilly). Available online.
- **Additional required readings and videos** will be made available to students in advance of each week's assignments. All will be availble online at no cost.
- In addition to the required materials, students may find the following resources helpful in supplementing course materials:
  - **Recommended Text**: *Foundations of Data Science*, Avrim Blum, John Hopcroft, and Ravindran Kannan. January 2018. Available free online here.
  - **Recommended Text**: *Elements of Statistical Learning*, Trevor Hastie, Robert Tibshirani and Jerome Friedman. 2nd Edition, 2009 (Springer). Available free online here.

- **Recommended Text**: *Python for Data Analysis*, Wes McKinney. 2nd Edition, October 2017 (O'Reilly). Available online.

Tentative Schedule: Fall 2019

Subject to revision. Latest version available on the course page.

Week	Date	Topic
1	September 5 (Thur)	Course Intro: What is Data Science and Why Does It Matter?
2	September 9	Data Exploration 1: How to Get Data
3	September 16	Data Exploration 2: Processing and Cleaning Data
4	September 23	Data Exploration 3: Statistics and Stories We Tell Ourselves
5	October 7	Models 1: Intro to Regression and Classification
6	October 16 (Wed)	Models 2: Regression and Classification, Part 2
7	October 21	ML 1: Trees, Bias vs. Variance Tradeoffs
8	October 28	ML 2: Performance Evaluation and Ensemble Models
9	November 4	Guest Lecture: TBD
10	November 11	Midterm & Coding Demo
11	November 18	ML 3: NLP, Text as Data, and Bayes Rule
12	November 25	ML 4: Unsupervised Learning
13	December 2	Big Data
14	December 9	Life in Data

CUNY Policy on Academic Integrity

The CUNY Policy on Academic Integrity is available here. The policy, as adopted by the Board, is available to all students. Academic dishonesty is prohibited in the City University of New York and is punishable by penalties, including failing grades, suspension, and expulsion.

Accommodations for Disabilities

Reasonable accommodations will be made for students with verifiable disabilities. In order to take advantage of available accommodations, students with disabilities seeking accommodations and/or support services at The City College of New York are required to register with the AccessAbility Center/Student Disability Services (AAC/SDS). For more information, visit www.ccny.cuny.edu/accessability.