2018

CSC 59970 Introduction To Data Science

Grant Long
CUNY City College

NYC Tech-in-Residence Corps

CSC 599.70 - Introduction To Data Science

Course Title	Introduction To Data Science
Time	Mondays, 6:30pm-9:00PM
Location	NAC 5/110
Credits & Hours	3 credits, 3 hours
Instructor	Grant Long
Email	gml252@nyu.edu
Office	Varies
Office Hours	Varies, by appointment
Course Website	grantmlong.com/teaching

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 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.

Pre-requisites

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

	Weight
Group Project	40%
Homework / Quizzes	30%
Midterm Exam	20%
Attendance / Class Partipation	10%

- **Project**. The bulk of the course grade will be a group project that will be due in December (exact date TBD). Students will be expected to work on the project during the second half of the class and will be required to present their progress throughout the course of the semester. 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 project will discussed in the second week of class.
- **Assigments**. This class includes short, frequent assignments to check comprehension. All assignments and quizzes will be graded on a 5-point scale. All quizzes will be announced in advance of class.
    - **No late assignments accepted**. Assignments not turned in by the set deadline will

be scored as 0/5. Exceptions will be granted only as mandated by CUNY policy.
    - **Worst two assignments dropped**, includes missed assignments.
- **Exam**. A short midterm exam will be held in October 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.
- **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.

Course Project

Specifics of the course project will be announced in the second week of the course.

Homework Grading

Homeworks and quizzes will be graded on a 10 point scale. The lowest 2 homeworks and quizzes will be dropped from the calculation of the final grade.

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 have:

1. A laptop capable of running git, bash, python 3.5 + libraries (eg. Anaconda 4.1.1)
2. Github account (free educational account ok)

Recommended Texts and Materials

- **Required Text**: *Data Science from Scratch*, Joel Grus. 2nd Edition, April 2015 (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**: *Python for Data Analysis*, Wes McKinney. 2nd Edition, October 2017 (O'Reilly). Available online.
    - **Recommended Text**: *Elements of Statistical Learning*, Trevor Hastie, Robert

Tibshirani and Jerome Friedman. 2nd Edition, 2009 (Springer). Available free online here.

Tentative Schedule: Fall 2018

Week	Date	Topic
1	August 27	Course Intro: What is Data Science and Why Does It Matter?
2	**September 5**	Data Exploration 1: Loading, Summarizing, and Visualizing Data
3	September 17	Data Exploration 2: Dirty Data
4	September 24	Data Exploration 3: Storytelling and Statistics
5	October 1	**Project Teams Formed.**, Models 1: Intro to Regression and Classification
6	October 15	**Project Proposals Due via Email** Models 2: Regularization, Variance/Bias
7	October 22	**Midterm Exam.**, ML 1: Trees, Feature Selection
8	October 29	*TBD Guest Lecture*
9	November 5	**First Project Update.** ML 2: Ensemble Models, Evaluation
10	November 12	ML 3: Bayes Rule and Bag of Words Methods for Text
11	November 19	ML 4: Unsupervised Learning
12	November 26	**Second Project Update.** ML 5: Bayesian Analysis and Scalable Data Science
13	December 3	TBD. Options: Big Data and the Cloud, Intro to Deep Learning, Recommender Systems, Advanced NLP, Data Ethics

14	December 10	Course Review, Careers in Data, Data Ethics

CUNY Policy on Academic Integrity

The CUNY Policy on Academic Integrity: 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.