DAT Class Syllabus

Basic Info

Instructor: Jonathan Bechtel

Head TA: Andrew Riddle

Instructor Contact Info: jonathanbechtel@gmail.com

Class Time: 7-10 PM EST, M/W

Class Duration: 02/22 – 04/28

Class Slack: q1ectechdata2021.slack.com

Office Hours: Mon 6-7 EST, Wed 6-7 EST

Github Repo: https://bit.ly/dat-02-22

Class Website: https://www.jonathanbech.tel/data-science-part-time-class

Class Schedule

Week	Class 1	Notes	Class 2	Notes	Reading
Unit 1: Python Foundations					
1, Week of Feb. 22		Intro, Class setup		Python Foundations	None
2, Week of Mar. 1		Dictionaries / API connections		Web Scraping	None
Unit 2: Pandas and Exploratory Data Analysis					
3, Week of Mar. 8		Pandas – Connecting, Syntax		Pandas – Common Operations	Py4DA: Ch.5, Ch.6
4, Week of Mar. 15		Pandas – Grouping, Dates		Pandas – Time Shifts / Data Hierarchies	Py4DA: Ch. 7, 9
5, Week of Mar. 22		Plotly Intro, Data Visualization		HW II Presentations / ML Intro	Py4DA: Ch. 7, 9
Unit 3: Model Building and ML Fundamentals					
6, Week of Mar. 29		GBM Intro		Data Prep / Pipelines	HOML Ch. 6
7, Week of Apr. 5		KFold / Model Validation		Partial Dependence	HOML – pgs 199-207

8, Week of Apr. 12		Classification / xgboost intro		HW III Presentations / Classification continued	HOML – pgs 23 - 33
Unit 4: Deep Learning / Model Deployment					
9, Week of Apr. 19		Data Applications		Deep Learning / Neural Net Intro	Py4DL Ch. 6
11, Week of Apr. 26		Neural Network Deep Dive		Transfer Learning	Py4DL Ch. 6
12, Week of May 3		BONUS: Neural Net Application			Py4DL Ch. 6

Important: This schedule is tentative and subject to change. It's normal for the class to diverge from the set schedule to some degree, and class material will be modified in an appropriate manner to make sure the class itself is completed in the most satisfactory way. Some flexibility is built into the syllabus with the expectation that class material will naturally take its own unique direction.

Class Expectations

This class is pass and fail, and a student's ability to get a passing grade is dependent on their attendance and successful completion of homework projects.

Attendance: Students must have an 85% attendance mark throughout the course. Classes missed entirely are worth one unit of attendance, and uncommunicated tardiness or early absences count for a ½ unit of attendance.

In general, judging late arrivals and early dismissals will be lenient if reasons are communicated beforehand.

Likewise, if students have arrangements that come up that force them to miss additional amounts of time then I'm happy to make accommodations as long as the issue is communicated to me clearly and in a good faith manner.

Homeworks: Every homework must be successfully completed in order to receive a passing grade. Every homework is Pass/Fail. It's understood that students might have different levels of aptitude upon entering the class, so most homeworks give students a choice of projects to choose from depending on the amount of time they have available as well as their level of expertise in the particular subject.

Class Homework

The class is broken up into 4 modules, and each one has its own homework assignment at its completion. With the exception of the 1st homework, students will be expected to give an 8-12 minute presentation on their project in front of a group of students on the day that it is due.

Each homework assignment is Pass/Fail, and students will receive a written evaluation after each homework assignment approximately one week after the assignment is due with their final grade and feedback.

Successful completion of every homework assignment is required in order to receive a passing grade for the class.

Homework Descriptions

Homework 1

Due Date: March 8

Option A: Python Foundations Coding Challenge

Overview: The primary purpose behind this homework is to make sure every student has a minimally acceptable level of basic programming knowledge to get through the rest of the class. It's designed to mimic coding challenge type questions one might encounter at a job interview and is meant to test an ability to grasp basic types of problem solving one might encounter when trying to write functions.

Option B: Twitter API Coding Challenge (More Advanced)

Overview: The Twitter API challenge gives students to work with a modern API and harvest it to dynamically create datasets using different conditions. Students will write functions that, when called, will connect to the Twitter API and return results depending on various conditions given.

Option C: Web Scraping Challenge (More Advanced)

Overview: In this homework assignment, students will scrape Yelp for restaurant reviews in the greater London area. You have the option of completing the project with up to 5 levels of difficulty, and will be tasked with transforming website information into a structured dataset.

Homework 2: Exploratory Data Analysis with Pandas

Due Date: March 24

Option A: Data Cleaning IMDB or Chipotle Dataset

Overview: Homework 2 is designed to test a student's ability to take a dataset, perhaps a messy one, and be able to clean and format it in a manner that allows for coherent insight and analysis. Students will be given two different datasets to evaluate, and a number of prompts for them to answer about the data itself, with some additional leeway to provide their own interpretation.

Students will also have the choice to use their own on this assignment if they find it relevant.

Option B: Working With Large Files (More Advanced)

Overview: For students who want an additional challenge, this option gives students exposure to a variety of Pandas functionalities for dealing with very large files. What if your data set is too large to fit into memory? What if basic computations are unreasonably slow? This homework will guide you through the basic operations necessary to optimize performance with large, messy datasets.

Students will also learn how to manipulate and clean files without entirely loading them into memory.

Homework 3: Model Building and ML Fundamentals

Due Date: April 14

Overview: Homework 3 is designed to allow students to use different types of statistical techniques on a dataset in order to draw inferences from it and make useful prognostications about what might happen next.

Students will have their choice of five different datasets to choose from with varying levels of complexity, and they'll work through an ML project from beginning to end.

Homework 4: Independent Project

Due Date: April 28

Overview: The final homework assignment will allow students to choose a datset and project of their choosing and create their own learning path that seems appropriate for them. The idea is that at this point students should have a clear idea of how to use the skills learned in the course to best further their own learning agenda, and this project will give them a chance to accomplish this.

They will also have the bonus challenge of being able to tie in aspects of Unit 1 as well as deploying their work as an application over the web for bonus credit.

Class Errata

Class Style

Most classes will have a similar format:

- One 'big idea' that will be the main focus
- Usually about 45 minutes – 1 hour of presentation/slides, often with students doing a code-along
- An individual or group project that goes for about 25-50 minutes
- Overview of the material as well as wrap up and introduction to the next topic

General Teaching Philosophy

In general, I think my job in this course is broken down into two categories:

- I should explain complicated topics clearly, and make them more easily digested than they otherwise would be if using other avenues to learn.
- I should push you outside of your own comfort zone to learn topics and concepts you maybe don't see yourself as being able to, or being necessary for your own career path.

With regards to point 1, our time together should be breezy, enjoyable and (maybe) fun. With regards to point 2, the experience should be uncomfortable, frustrating and much less entertaining than other uses of your time.

If we're doing a good job, we should be spending a decent amount of time dipping our toes into both experiences. So while we should expect to have a good time together, please remember **that the most effective learning necessarily requires some discomfort.** You are only going to take this class once and it's imperative that we work together to make sure it has maximum impact for your career goals, which sometimes requires a certain appreciation of short-term pain in order to get some long-term gain.

Online Class Discussion

The class has its own slack channel, which every student is invited to at the beginning of class. This is intended to give students a chance to ask each other questions about homework assignments, discuss general topics, as well as provide a place for me to make announcements about class assignments or other class wide details.

Class Prompts

To make class material more 'complete', many classes will come with either suggested reading or pre-work that might help students better prepared for material, or allow more advanced students delve into more differentiated material to make class material more satisfying.

This material will be released to the class GitHub repo, and will be announced on the class Slack channel.

Completing these prompts is optional, but is meant to give students a more thorough, engaging learning experience that's appropriate for their level of expertise.

Class Material/GitHub

The class has a github repo that it uses to disseminate all class material. Students are expected to continually initiate pull requests to get new course material as it's released. **Class material will be dripped out class by class.** If a student is going to be absent arrangements can be made for them to get material beforehand.

Students are also expected to setup their own GitHub repo, which is where they will publish their homework assignments for me to grade.

Class Readings

There are no required class readings, and students aren't expected to do any class readings in order to follow along with the course, but if students want to do some additional background research or prep themselves for material optional readings are given from the following three books:

Python for Data Analysis, Wes McKinney

URL: bit.ly/dat-book-1

Description: This book was written by the primary author of Pandas and is a good overview for how to use its various nuts and bolts to accomplish data cleaning tasks. It's not available for free, and must be bought.

Hands On Machine Learning With SciKit Learn and Tensorflow

URL: https://bit.ly/dat-ml-book

Description: The most comprehensive ML book written in Python. It covers almost all major techniques, and all examples use libraries that will be covered in this class.

Deep Learning with Python

URL: https://bit.ly/dat-dl-book

Description: A very concise, well written book that covers the most up-to-date advances in deep learning. Has very thorough walk throughs that quickly introduce you to the latest advancements in Deep Learning: computer vision, natural language processing, and generative models.

Additional Points

- **Working In Teams:** Students have the option of completing up to two of their homework assignments as part of a team, which can have two people each, except for the final project, which can have up to 4 people. Please contact me beforehand to let me know of your arrangements
- **Class Mixers:** Twice during the class we'll have an optional class mixer that will go on for about 1 – 1.5 hours to allow students to get to know each other better. These are optional, and dates will be discussed during class.