Title: Syllabus Slug: syllabus Date: 20120-8-12

Welcome to AC295. In this course we explore advanced data science practices. The course is divided into three major topics:

1. How to scale a model from a prototype (often in jupyter notebooks) to the cloud. In this module, we cover virtual environments, containers, and virtual machines before learning about containers and Kubernetes. Along the way, students will be exposed to Dask.

2. How to use existing models for transfer learning. Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task. It is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and natural language processing tasks. This could be very important, given the vast computing resources required to develop neural network models on these problems and the huge gains that these models can provide. In this part of the course, we will examine various pre-existing models and techniques in transfer learning.

3. In the third part, we will introduce several intuitive visualization tools for investigating and diagnosing models. We will be demonstrating a number of visualization tools ranging from the well established (like saliency maps) to recent examples that have appeared in https://distill.pub.

Lectures (online): Tuesday and **Thursday** 10:30-11:45am (and possibly depending on timezone of students repeat **Tuesday and Thursday** from 9:00-10:15pm)
TFs: Rashmi Banthia, William Palmer, Andrea Porelli
Office Hours: TBD

List of Contents

- Assignments
- Getting Help
- Course Policies

Prerequisites

Students are expected to be fluent in programming (Python), statistics knowledge at the level of Stat 110 or above, data science (or machine learning) at the level of AC209A and AC209B.

Software

We will be using a variety of software, primarily Python 3, Pytorch, Tensorflow, and Docker. More details in class.

Topics

The course is organized in three modules.

1. **Deploy data science** (integration + scalability)
   1a. Virtual Environments and Virtual Boxes 1b. Containers
   1c. Kubernetes
   1d. Dask

2. **Transfer learning and distillation**
   2a. Intro to Transfer Learning: basics and Convolutional Neural Networks (CNNs) review 2b. Transfer Learning for Images and SOTA Models 2c. Language Models and Transfer Learning with Text Data
   2d. Attention and Transformers 2e. Distillation and Compression

3. **Visualization as investigative tool**
   3a. Introduction and Overview of Viz for Deep Models: lime and shapley 3b. CNNs for Image Data, Activation Maximization and Saliency Maps 3c. Attention for Debugging Language Models

Course Activities

Each module is structured in three types of activities: **Lectures**, **Reading Discussion**, and **Practicum**. Each

activity requires the students to complete different assignments in the form of exercise/homework, discussions, reading assignments, and presentations (see Assignments below). During the first weeks of each module, students will attend a lecture on Tuesday and reading discussion on Thursday. The last week of each module will be Practicum. **Attendance is mandatory**.

1. **Reading List** consists of papers, blogs, and other reading material that will be released no later than the beginning of each week. This will be the base for all the activities during the week. See Readings Guidelines here link to guidelines.

2. **Lectures** are held online **Tuesdays** from 10:30-11:45am (and possibly depending on timezone of students repeat **Tuesday** from 9:00-10:15pm). During this activity, we will discuss and summarize the basic concepts of the material covered during the week.

3. **Journal Discussions** are held online on **Thursdays** from 10:30-11:45 am (and possibly depending on timezone of students repeat **Thursday** from 9:00-10:15 pm). During this activity, two groups will present one or two papers from the Reading List to the rest of the class and lead the discussion. See Paper Presentation Guidelines here link to guidelines.

4. **Practicum** are activities in the form of a project based on the material covered in the module. The students will work in groups and be expected to deliver a complete assignment in 10 days. There will be two practicums.

Assignments

The **final grade** will be calculated using the following weights for each assignment:

Exercises

There are eight (8) exercises to complete. They will be released at the end of each regular week *Lecture* and due a week later. The exercises are graded on a scale 1 to 5, where 5 is the highest grade.

Discussion Forum

There will be a discussion forum the day before the *Reading Discussion* on Thursday based on the reading from *Reading List*. All discussions will be on the Ed Platform (select Ed from the tab on the canvas course's main page).

Presentations

At every Reading Discussion, groups will present the reading material assigned at the beginning of the week.

Please see these on the presentations.

Practicums

There will be two practica during the first two modules (see the schedule for details).

Final Projects

There will be a final group project due during Exams period encompassing all the material learned in class.

Assignment	Final Grade Weight
Discussion Forum	10%
Exercises	10%
Presentations	15%
Practicums	45%
Final Projects	20%

|**Total**| **100%**|

Getting Help

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

1. Go to **Office Hours**; this is the best way to get help.
2. Post the question in **Ed Forum**, and hopefully, your peers will answer.

Course Policies

Collaboration Policy

We encourage students to talk and discuss the assignments with their fellow students. Discussion is encouraged. Presentation during Reading Discussion, Practicum, and Projects are group activities.

Communication from Staff to Students

Class announcements will be through **Ed Forum**.

Diversity and Inclusion Statement Data Science, like many fields of science, has historically only been represented by a small sliver of the population. Recent initiatives have attempted to overcome some barriers to entry. We would like to attempt to discuss diversity in Data Science from time to time where appropriate and possible.

Please contact me (in person or electronically) or submit anonymous feedback if you have any suggestions to improve the quality of the course materials. The best way to provide anonymous feedback is to use Ed, which allows you to provide comments anonymously.

Furthermore, I would like to create a learning environment for my students that supports a diversity of thoughts, perspectives and experiences, and honors your identities (including race, gender, class, sexuality, religion, ability, etc.) To help accomplish this:

If you have a name and/or set of pronouns that differ from those that appear in your official Harvard records, please let me know! If you feel like your performance in the class is being impacted by your experiences outside of class, please don't hesitate to come and talk with me. I want to be a resource for you. Remember that you can also submit anonymous feedback (which will lead to me making a general announcement to the class, if necessary to address your concerns). If you prefer to speak with someone outside of the course, you may find helpful resources at the Harvard Office of Diversity and Inclusion. We (like many people) am still in the process of learning about diverse perspectives and identities. If something was said in class (by anyone) that made you feel uncomfortable, please talk to us about it. (Again, anonymous feedback is always an option.) As a participant in course discussions, you should also strive to honor the diversity of your classmates.

Academic Honesty

Ethical behavior is an important trait of a Data Scientist, from ethically handling data to attribution of code and work of others. Thus, in AC295 we give a strong emphasis to Academic Honesty. As a student your best guidelines are to be reasonable and fair. We encourage teamwork for problem sets, but you should not split the assignments and you should work on all the problems together.

Accommodations for students with disabilities

Students needing academic adjustments or accommodations because of a documented disability must present their Faculty Letter from the Accessible Education Office(AEO) and speak with Pavlos by the end of the third week of the term: Friday, September 18. Failure to do so may result in us being unable to respond in a timely manner. All discussions will remain confidential.

Accommodations for students with disabilities