BST 261 Data Science II
Spring 2019 MW 9:45-11:15am Kresge G2

Instructor

Heather Mattie
Instructor of Data Science
Office: Building 1, 4th floor, room 421A
Email: hemattie@hsph.harvard.edu
Phone: (617) 432-5308
Office Hour: Wednesdays 1 - 3 pm or by appointment

Teaching Assistants

Matt Ploenzke
PhD Candidate, Biostatistics
ploenzke@g.harvard.edu
Office hour: Fridays 11:15am - 12:15pm, Kresge 202A
Labs: Fridays 9:45am - 11:15am except March 29 and May 17, LL6

Aaron Sonabend
PhD Candidate, Biostatistics
asonabend@g.harvard.edu
Office hour TBD

Course Description

Deep learning is a subfield of machine learning that builds predictive models using large artificial neural networks. Deep learning has revolutionized the fields of computer vision, automatic speech recognition, natural language processing, and numerous areas of computational biology. In this class, we will introduce the basic concepts of deep neural networks and GPU computing, discuss basic neural networks, convolutional neural networks and recurrent neural networks structures, and examine biomedical applications. Students are expected to be familiar with linear algebra and machine learning.

Course Objectives

Upon successful completion of this course, you should be able to:

- Understand the state of the art deep learning algorithms
- Understand the pros and cons of different approaches
- Implement deep machine learning applications using cloud GPU servers
- Become familiar with ways to optimize deep learning methods for biomedical applications
- Appreciate the strengths and limitations of deep learning applications

Course Prerequisites

Students should have taken a semester of linear algebra and multivariable calculus, and be comfortable programming in Python. It would be helpful if students are also familiar with machine learning. Linear algebra and Python review slides will be available but not presented in class. The foundations of machine learning will be covered in class.

Credits

This is a 2.5 credit course.

Course Structure and Grading

Course grades will be determined on the basis of three homeworks and a group project proposal.

- Homework #1 (25%), due April 8th by 11:59pm
- Homework #2 (25%), due April 22nd by 11:59pm
- Homework #3 (25%), due May 6th by 11:59pm
- Group project proposal (25%), due May 15th by 11:59pm

Homework

All homework assignments must be written in Python and submitted in the form of a Jupyter notebook on the course Canvas site by the due dates listed above. A template notebook will be provided for each homework on the course Canvas site and the course GitHub repository. Students are encouraged to work together on the assignments, but each must submit their own notebook and unique response to open-ended questions. The assignments will be related to the material presented in class and the lab sessions will help to answer questions in each assignment.

Group project proposal

Due to the short length of the course, only a project proposal, and not an entire project, will be due on May 15th. Students may work individually or in a team of no more than 4 students. The proposal must be in the form of a Word doc or pdf and be submitted on Canvas. Only one proposal needs to be submitted per group. The proposal must contain the following:

1. Project title
2. All group member names
3. A short literature review (no more than 1 page)
    - Remember to cite your sources by adding a bibliography at the end of the proposal
    - You should also cite the data source
    - Several resources are provided below for some inspiration
4. The knowledge gap the project will be filling
    - What is the goal of the project? (What is the task?)
    - Why is the project important? (Please note that "because it's worth 25% of my grade" is not an appropriate answer)

5. The data needed
    - Where will the data come from?
    - What is the outcome of interest?
    - Is the outcome binary, categorical or continuous?
    - What are the features (predictors) you'll be using?
    - Will any feature engineering need to be done?
    - How large is the data set?
    - Is the data publicly available?
6. Methods
    - What kind of model(s) would you use for this project?
    - Describe the architecture of the model. If you are thinking about multiple models, please describe each one.
        - How many layers will the model(s) have?
        - What kinds of layers? (fully connected, convolutional, pooling, LSTM, etc.)
        - Which activation function?
        - Which loss function?
    - What will the train/validation/test split be?
    - What measure(s) of accuracy will you use?
    - How will you work to reduce any overfitting?

Late Day Policy

Each student is given six late days for homework at the beginning of the course. A late day extends the individual homework deadline by 24 hours without penalty. No more than two late days may be used on any one assignment. Late days are intended to give you flexibility: you can use them for any reason, no questions asked, and you do not need to tell us when you use them. You don't get any bonus points for not using your late days. Also, you can only use late days for the individual **homework** deadlines. No late days may be used for the project proposal. Although each student is only given a total of 6 late days, we will be accepting homework from students that pass this limit. However, we will be deducting 2 points for each extra late day.

Course Materials

Course Canvas

The Canvas site is an important learning tool for this course where students will access course materials, **submit course assignments** and share other resources with the class. Course announcements will be posted on the site and students will be required to check the course site on a weekly basis.

Course GitHub

All course materials (slides, in-class examples, labs, homework assignments) will also be available on the course GitHub repository.

Textbooks

- Deep Learning, Goodfellow and Bengio, 2016
  Freely available online
- Deep Learning with Python, Chollet, 2017
  The first few chapters are available online

Project Resources

Papers

- CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning
- Multitask Learning and Benchmarking with Clinical Time Series Data
- Learning to diagnose with LSTM recurrent neural networks
- Evaluating deep variational autoencoders trained on pan-cancer gene expression
- Semi-supervised learning of the electronic health record for phenotype stratification
- U-net: Convolutional networks for biomedical image segmentation
- Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs
- Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning
- Deep Learning Predicts Tuberculosis Drug Resistance Status from Whole-Genome Sequencing Data

Data

- Large repository of publicly available medical data. Includes links and descriptions of datasets.
- MIMIC-III
- NIH Chest X-ray database
- Data and Specimen Hub from the NICHD
- Gene Expression Omnibus - Database of over 1 million gene expression samples

Other Resources

Python

- Jupyter Notebook
- Using Python for Research Videos

Google Cloud Platform

- Google Cloud Platform
- BST 261 GCP Tutorial

Machine Learning and Deep Learning

- fast.ai Course
- Google Machine Learning Crash Course
- A Review Article on Deep Learning
- Opportunities and obstacles for deep learning in biology and medicine
- Deep Learning 101

Git and GitHub

- git Reference
- Understanding git Conceptually
- GitHub git Desktop Client
- githug

Linear Algebra, Statistics and Machine Learning

- fast.ai
- Stanford CS229 Linear Algebra Primer
- fast.ai Linear Algebra
- Additional Linear Algebra
- Notes on linear algebra needed for Deep Learning
- Fast Hamiltonian Monte Carlo Using GPU Computing

Tools

- Keras
- Tensorflow

Harvard Chan Policies and Expectations

Inclusivity Statement

Diversity and inclusiveness are fundamental to public health education and practice. Students are encouraged to have an open mind and respect differences of all kinds. I share responsibility with you for creating a learning climate that is hospitable to all perspectives and cultures; please contact me if you have any concerns or suggestions.

Bias Related Incident Reporting

The Harvard Chan School believes all members of our community should be able to study and work in an environment where they feel safe and respected. As a mechanism to promote an inclusive community, we have created an anonymous bias-related incident reporting system. If you have experienced bias, please submit a report here so that the administration can track and address concerns as they arise and to better support members of the Harvard Chan community.

Title IX

The following policy applies to all Harvard University students, faculty, staff, appointees, or third parties:

- Harvard University Sexual and Gender-Based Harassment Policy
- Procedures For Complaints Against a Faculty Member
- Procedures For Complaints Against Non-Faculty Academic Appointees

Academic Integrity

Each student in this course is expected to abide by the Harvard University and the Harvard T.H. Chan School of Public Health School's standards of Academic Integrity. All work submitted to meet course requirements is expected to be a student's own work. In the preparation of work submitted to meet course requirements, students should always take great care to distinguish their own ideas and knowledge from information derived from sources.

Students must assume that collaboration in the completion of assignments is prohibited unless explicitly specified. Students must acknowledge any collaboration and its extent in all submitted work. This requirement applies to collaboration on editing as well as collaboration on substance.

Should academic misconduct occur, the student(s) may be subject to disciplinary action as outlined in the Student Handbook. See the Student Handbook for additional policies related to academic integrity and disciplinary actions.

Accommodations for Students with Disabilities

Harvard University provides academic accommodations to students with disabilities. Any requests for academic accommodations should ideally be made before the first week of the semester, except for unusual circumstances, so arrangements can be made. Students must register with the Local Disability Coordinator in the Office for Student Affairs to verify their eligibility for appropriate accommodations. Contact Colleen Cronin ccronin@hsph.harvard.edu in all cases, including temporary disabilities.

Religious Holidays, Absence Due to

According to Chapter 151c, Section 2B, of the General Laws of Massachusetts, any student in an educational or vocational training institution, other than a religious or denominational training institution, who is unable, because of his or her religious beliefs, to attend classes or to participate in any examination, study, or work requirement on a particular day shall be excused from any such examination or requirement which he or she may have missed because of such absence on any particular day, provided that such makeup examination or work shall not create an unreasonable burden upon the School. See the Student Handbook for more information.

Course Evaluations

Constructive feedback from students is a valuable resource for improving teaching. The feedback should be specific, focused and respectful. It should also address aspects of the course and teaching that are positive as well as those which need improvement.

Completion of the evaluation is a requirement for each course. Your grade will not be available until you submit the evaluation. In addition, registration for future terms will be blocked until you have completed evaluations for courses in prior terms.

Course Schedule

Date	Meeting Type	Topics	Deliverables
March 25	Lecture	Introduction to course Brief history of deep learning Brief review of machine learning	Create GCP account Request MIMIC III Data
March 27	Lecture	Backpropagation and MLPs	
April 1	Lecture	MLPs continued	
April 3	Lecture	Universal ML workflow Performance measures Bias/Variance tradeoff	
April 5	Lab	Basics of Python, Keras, Linear Algebra and GCP	
April 8	Lecture	Convolutional neural networks (CNNs)	Homework #1 due
April 10	Lecture	Convolutional neural networks (CNNs)	
April 12	Lab	CNNs: basics	
April 15	Lecture	Convolutional neural networks (CNNs)	
April 17	Lecture	Recurrent neural networks (RNNs)	
April 19	Lab	CNNs: advanced	
April 22	Lecture	Recurrent neural networks (RNNs)	Homework #2 due
April 24	Lecture	Recurrent neural networks (RNNs)	
April 26	Lab	RNNs: basics	
April 29	Lecture	Guest Lecture: Matt Ploenzke Applications in genomics and model interpretability	
May 1	Lecture	Guest Lecture: Matt Ploenzke Deep generative learning: GANs, VAEs, adversarial attacks	
May 3	Lab	Deep learning for genomics	
May 6	Lecture	Advanced topics	
May 8	Lecture	Advanced topics	
May 10	Extended Office Hours	Project proposal help	Homework #3 due
May 13	Lecture	Advanced topics	
May 15	Lecture	Advanced topics	Group project proposal due