BST 261: Data Science II
Spring II 2021 MW 9:45-11:15am

Instructor

Heather Mattie
Lecturer on Biostatistics
Office: Building 1, 4th floor, room 421A
Email: hemattie@hsph.harvard.edu
Phone: (617) 432-5308
Office hour: Mondays 8 - 9pm EST or by appointment

Teaching Assistants

Beau Coker
PhD Candidate, Biostatistics
beaucoker@g.harvard.edu
Office hour: Thursdays 1 - 2pm EST

Gopal Kotecha
PhD Candidate, Biostatistics
gkotecha@g.harvard.edu
Office hour: Wednesdays 4:30 - 5:30pm EST

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 including public health, medicine and computational biology. In this class, we will introduce the basic concepts of deep neural networks, discuss basic neural networks, convolutional neural networks and recurrent neural networks structures, and examine biomedical applications. Students are expected to be familiar with calculus, linear algebra, machine learning and Python.

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.

Credits

This is a 2.5 credit course.

Course Structure and Grading

Course grades will be determined on the basis of two problem sets, a 5-minute individual presentation, and a group project proposal. An overall grade of 70% is needed to pass this course.

- Problem Set #1 (20%)
- Problem Set #2 (30%)
- Paper presentation (25%)
- Group project proposal (25%), due May 14th by 11:59pm

Problem Sets

All problem sets must be written in Python and submitted in the form of a Jupyter notebook on the course Canvas site. A template notebook will be provided for each problem set on the course Canvas site in the form of a link to a Google Colab notebook. 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.

Paper Presentation

Each student is expected to read and present a journal article in a 5-minute presentation. Presentations should be recorded and uploaded to the Paper Presentations discussion board on canvas along with the link to the paper. Slides are encouraged but not mandatory. Students may sign-up for a particular article from this list or ask Heather for approval if they find an article of interest not on the list. Recordings should be uploaded before class on the date specified on the list. In addition to a paper presentation, each student is expected to watch and comment on 7 other presentations (not including their own). Comments do not have to be long but should include something specific about the video.

If you are unable to record your presentation or need help posting it to the discussion board, please reach out to Heather and the TAs as soon as possible.

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 14th. 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 four 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 **problem set** deadlines. No late days may be used for the project proposal or paper presentation. Although each student is only given a total of 4 late days, we will be accepting problem sets from students that pass this limit. However, we will be deducting 20% of the problem set points for each additional late day. Students who have an emergency and are unable to submit their assignment(s) on time should email Heather as soon as possible.

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. **All lectures, lab sessions and office hours will be held via Zoom. The links to each Zoom meeting can be found on the course Canvas site in the menu on the left-hand side. Each meeting will be recorded and available to view throughout the course.**

Course GitHub

All course materials (slides, in-class examples, labs, problem sets) 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
  Freely available here 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

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 22	Lecture	Introduction to course Brief history of deep learning MLPs	
March 24	Lecture	Backpropagation and MLPs	
March 26	Lab	Backpropagation and MLPs	
March 29	Lecture	MLPs in Python with Keras	
March 31	Lecture	MLPs in Python with Keras continued Regularization	
April 2	Lab	MLPs continued	
April 5	Lecture	Introduction to Convolutional neural networks (CNNs) Data augmentation Using pre-trained networks	Problem Set 1 Due
April 7	Lecture	CNNs continued Transfer learning Visualizing what CNNs learn	
April 9	Lab	CNNs: basics	
April 12	Lecture	CNNs continued	
April 14	Lecture	CNNs continued	
April 16	Lab	CNNs: advanced I	
April 19	Lecture	CNNs continued	
April 21	Lecture	Guest Lecture	
April 23	Lab	CNNs: advanced II	
April 26	Lecture	Introduction to Recurrent neural networks (RNNs)	
April 28	Lecture	RNNs continued	
April 30	Lab	RNNs: basics	
May 3	Lecture	RNNs continued	
May 5	Lecture	RNNs continued	
May 7	Lab	RNNs: advanced	Problem Set 2 Due
May 10	Lecture	Advanced topics	
May 12	Lecture	Advanced topics	
May 14	**No Lab**		Group project proposal due