Syllabus for:

Python for Machine Learning and Data Analysis

Spring 2019

Time and place:	Thursday 6:00pm – 9:00pm	TBD
Instructor:	Handan Liu	h.liu@northeastern.edu
Office hours:		
TA:		

Course Rationale:

This course we will cover is broadly applicable, and has led to significant advances in many fields. Once you understand the basics of the technology of machine learning and data analysis, and the close connection between theory and practice, it's a very open field, where lots of progress can be made quickly. More important, you will transform your theoretical knowledge into practical skill using many hands-on labs in this course.

Prerequisites:

You should be comfortable programming in Python. Basic knowledge of Unix/Linux will be helpful.

Grading:

Homework:	40%	(HW1 10%, HW2 15% and HW3 15%)
Project:	30%	
Attendance:	10%	
Final exam:	20%	

Assignments:

Lectures are complemented by homeworks (programming assignments) to bridge the theory with the practice. The homeworks are associated with the three main parts of the course that mostly consist of programming assignments to exercise a technology or programming model.

Homework is due electronically. All assignments have a specific due date and time. Submissions will be accepted up to one day after the deadline with a 50% penalty. For example, an on-time submission might receive a grade of 80 points. The same assignment submitted after the deadline would receive 40 points (80 x 0.5).

Submission will be done through Blackboard.

Make-up Policy

Students who miss the presentation and the final exam will not, as a matter of course, be able to make up it. If there is a legitimate reason why a student will not be able to complete an assignment on time or not be present for the presentation or the exam, then they should contact the instructor beforehand. Under extreme circumstances, as decided on a case-by-case basis by the instructor, students may be allowed to make up assignments or exam without first informing the instructor.

Course Schedule:

Week	Topics	Assignments
1	• Course/Syllabus • Python Infrastructure and Development Tools: Anaconda, an enhanced interactive science-centric console; and the Jupyter Notebook, a web-based application that mixes code, plots, and rich media, making it ideal for sharing and publishing analyses with peers, etc. • Anaconda, conda and pip • *Lab 1*: install anaconda o Custom environment settings o Jupyter Notebook and JupyterLab o Script editor for Python	Install Software on your computer
2	Python Language Essentials: Introduction to Python's core language features and packages: Python's built-in data structures, including how and where each might be used and what trade-offs are present, and cover Python's looping and control flow constructs, etc. • Fundamental data types and data structures; Organizing code with functions, modules and packages; Loading packages, namespaces; Reading and writing data; Control flow • *Lab 2*: samples - writing python code and run it	
3	Numerical Analysis and Data Exploration with **NumPy** Arrays: NumPy is a critical tool for rapidly manipulating and processing large data sets. • The NumPy array; Selecting data using slicing and logical indexing; Efficient numerical processing with multi-dimensional arrays; Expressive array operations and manipulations; Access larger-than-RAM data using memory mapped arrays • *Lab 3*: run samples	

4	Data Visualization with **Matplotlib** • 2D plotting with Matplotlib: line plots, scatter plots, histograms, labeling, and more. • *Lab 4*: run samples	HW1
5	**Pandas**, the Python Data Analysis Library is a powerful package for working with tabular data for data aggregation and reorganization, such as: • Loading from CSV and other structured text formats; Accessing data stored in SQL databases; 1D and 2D data structures; Stripping out extraneous information; Normalizing data; Dealing with missing data; Data manipulation (alignment, aggregation, and summarization); Group-based operations: split-apply-combine; Statistical analysis; Date and time series analysis with Pandas; Visualizing data • *Lab 5*: run samples	
6	***Lab 6***: Hand-on Lab: • Setup computers to be able to access Discovery Cluster • Learn how to submit and manage jobs on the HPC cluster • Write our first program and run on HPC cluster	
7	Parallel programming in Python: • Python multiple threads: **multiprocessing** • Python MPI: **mpi4py** *Lab 7*: run python parallel code on HPC cluster	HW2
8	Introduction to Machine Learning: This section starts with a short conceptual introduction to machine learning and explain how it works, and what kinds of problems it's best suited to solve. And the section covers the frameworks and tools provided by **scikit-learn**, a widely used library for machine learning, such as: • Linear and nonlinear models; Constant and variable learning-rates; Cost functions, regularization methods, and other constraints; Fitting, transforming, and predicting • *Lab 8*: run samples	
9		

10	Machine Learning: Numeric Data: • Logarithmic and curvilinear transforms; Data scaling; Outliers; Linear regressors; l1 and l2 normalization; Support vector machines (SVM) Machine Learning: Categorical Data: • Contrast encoding; Missing values; Categorical rebinning; Linear classifiers; Tree-based classifiers; Ensemble methods; Boosting methods; Unbalanced designs Machine Learning: Image Data: • Image storage formats; Scikit-image; Smoothing and denoising; Edge detection; Feature-based segmentation; K-means clustering *Lab 9*: run samples	HW3
11		Project
12	Brief Introduction to the Python Deep Learning Library **TensorFlow** and **Keras**; *Lab 10*: run python code of ML and DL on HPC cluster for parallel on CPU and GPU	
13	Final exam; debrief	
14	Presentation of the project by every student	

Note: This schedule and contents will be adjusted as needed throughout the semester.

Resources:

Required Textbook: None

Other Material:

- Python: https://www.python.org/
- Anaconda: https://www.anaconda.com/
- Numpy: http://www.numpy.org/
- Matplotlib: https://matplotlib.org/
- Pandas: https://pandas.pydata.org/
- Scikit-learn: https://scikit-learn.org/
- TensorFlow: https://www.tensorflow.org/
- Keras: https://keras.io/

Other information will be completed later.