CSCI E-82 Advanced Machine Learning, Data Mining, and Artificial Intelligence
Fall 2020
Syllabus (Tentative – will be updated)

With the rapid rise of the interdisciplinary data science and big data fields, there has been a push for increased extraction of knowledge and insight from all types, forms, and shapes of data. Both of these new fields are built on algorithms that construct models of data and facilitate decision-making. Machine learning draws upon approaches found in the computer science, math and statistics fields making it less accessible for those without any technical training. Consequently, many new practitioners use these algorithms as black boxes without understanding their nuances or limitations. This course focuses on understanding how the primary machine learning algorithms work so students will be able to select appropriate methods, adapt the methods to solve specific problems, and work to overcome the limitations of the standard algorithms. This can provide a competitive advantage professionally for practitioners. The course will cover a range of current research fields and provide experience working on different types of data.

This course primarily builds upon CSCI E-63C: Elements of Data Science and Statistical Learning with R. We will assume that students are familiar with the concepts from the course and/or the Introduction to Statistical Learning text from Tibshirani and Hastie. This course will include:

- deeper coverage of some topics and theory
- additional clustering and classification methods
- new topics such as deep learning, outlier detection, time series, recommender systems, etc.
- practical and theoretical experience in applied areas such as text mining and image processing
- focus on python rather than R

Course Goals:

1) Provide a conceptual understanding of how the various algorithms work. This will largely be achieved through a combination of lectures that provide background and some theory, and programming exercises where students will learn by implementing various approaches
2) Provide hand-on exposure to some of the standard [mostly Python] software used by practitioners of the field with guidance from the course sections
3) Provide an entry point into different domains (text mining, images analysis, time series modeling, etc.) working with typical messy data that requires data cleaning
4) Empower students to correctly interpret and present results--a key skill set for practitioners of machine learning and data mining
5) Increase knowledge of field and be able to utilize journal articles in the field

Please note:

With these goals in mind, most students have found this course to be challenging. Each lecture will cover a lot of material quickly and past students have reported listening to the lectures twice. The exam will focus on understanding the key concepts of the lectures. Many of the homework problems are solving real problems rather than finding the simple 'right' answer. This makes the grading process as a comparison of how well many outstanding students have tackled the problems. Past students have strongly recommended that you allocate sufficient time to get through this course (see the FAQ at the end). We hope the outcome will be both fulfilling and rewarding.

Teaching Staff:

Instructor:	Peter V. Henstock, Ph.D.	machinelearnmine@gmail.com
Teaching Assistants:	Rashmi Banthia	rjain29@gmail.com
	Roman Budakov	br00ksrom@gmail.com
	Daniel Sauble	das3810@g.harvard.edu

Course Web sites:

Canvas course web site: https://canvas.harvard.edu/courses/TBD
Piazza for discussions: https://canvas.harvard.edu/courses/TBD
Zoom help:

Email Support = academictechnology@dce.harvard.edu
Phone Support = 617 998-8571

Course Meeting Times:

The lecture each week will be held 7:20-9:20 pm EST Thursdays via live web conference. We plan for students to present two times this semester: once during the semester and once for the final project meeting. Optional (but recommended) sections will be arranged to accommodate as many students as possible within the constraints of our teaching staff. In the past, sections have been held on weekend mornings. All lectures and sections will be recorded.

Course Philosophy:

There are two standard approaches to teaching machine learning and data mining. The first approach is well suited for advanced engineering students and statisticians where the emphasis is on deriving statistical learning equations to understand them. The second approach (which this course embraces) will present more conceptual, algorithmic and computational ideas of the field in lecture, and have students gain experience by coding and exploring different solutions. While the course will present mathematical details in most lectures, the emphasis will be on practical understanding how to leverage the ideas through coding. Students with a strong math/statistics backgrounds will certainly learn the field and gain new skills, but the approach may be different than you are accustomed to. Rather than deriving many equations, you will gain an understanding and also have to consider how to implement the equations efficiently.

Successful practitioners of the field have:

1) a solid understanding of what the various algorithms do and when to use them
2) ability to adapt the approaches to meet different needs specific to a project goal
3) research skills to locate and utilize the literature of the field

Consequently, we will have various types of assignments in this course. We will have individual "breadth-first" homework assignments that provide some hands-on familiarity with various content areas. These will be followed by a "depth-first" opportunity to work with a partner and explore different ideas and approaches to solve more open problems. Current plans for the latter include a project to work with your own data, and a Kaggle-like competition for a fixed data set. The final project will be an opportunity to extend any approach from the class one step further while working with a partner that will be presented to the class. Reading some journal and conference papers will be a component of the course to extend the content into areas of your interests.

Prerequisites:

a) Software programming/scripting skills in a major language including the ability to read/parse/write files and code mathematical calculations. The course will focus on Python as the standard machine learning language. Although we do not require prior experience with python, past students have strongly recommended learning it prior to the start of the course. This is a challenging course and if you're learning the programming aspects at the same time, it will be more demanding for you and also your partners. Consequently, we recommend learning the data structures (pandas), plotting, file I/O, etc. and general data manipulation using any of the edX, Coursera, Udemy, free resources, or any python book. The first section will quickly review some python basics by diving into pandas, libraries, and plotting tools. We recommend the Anaconda distribution of python and will be focusing more on Python 3.7 for most of the course.
b) General familiarity with linear algebra: understand matrix notation; how to add, multiply, and transpose matrices; and conceptually understand matrix inverses
c) Basic familiarity with probability and statistics will be helpful although we will review this. We recommend understanding probabilities (including joint and conditional), Bayes theorem, and understanding distributions.
d) We will release an ungraded Assignment 0 by early August so you can gauge your skills and perhaps improve brush up some areas before the course starts.

Who is this class for?

This course is designed for Harvard Extension students who take classes in the general Data Science/Software Engineering/IT/IMS program. Many wonderful Harvard Extension courses in the program teach how to program and modify data, transport and store data, and some focus on the theory. This course is intended as a complement to the other courses for analyzing and leveraging data to make decisions. As discussed in the Course Philosophy, the intended audience is a software coders who can use some math and statistics to try out and gain hands-on experience with different ideas, rather than a student from a strong mathematical/statistical background who might be comfortable with a different approach. Regardless of your background, we will [push you and] empower you to analyze a range of data types and solve what we hope are interesting and meaningful problems.

How much math/statistics do you really need?

We will routinely use math and statistics in the lecture to show how things work, and also since one equation is much easier to understand than as 20 lines of code. This will include an occasional derivative or integral but we will not require a derivation or specific mathematical manipulation in the homework or on the test. Some topics will include linear algebra (matrices) but the emphasis will be on coding and understanding the components rather than understanding how to manually compute an inverse, eigenvalues, etc. Probabilities, Bayes theorem, and basic linear algebra will be reviewed briefly in class.

How much programming do you need?

Given the philosophy, the homework will emphasize implementing different ideas in the python language mostly. If you have not taken any programming courses at the Harvard Extension or elsewhere, this course will likely be extremely challenging. You should be able to download and parse data files such as CSV or JSON formats, use control loops and recursion. If you have learned java, C"*", perl or other major languages, python will not be a problem. However, it takes time to get up to speed with a new language. Past students have stated that it was difficult to learn a lot of challenging material and learn python at the same time, yet alone take other classes and work. While we may allow other languages to be used for some projects (TBD), python is the standard in the computer science - machine learning area and is worth learning. Therefore we will require python code for many assignments. Furthermore, we strongly recommend that you learn some python perhaps through a Udemy course that includes the use of Pandas, Matplotlib, etc.

How much time does this course require?

Based on past years, students generally spend 2+ hours in lecture, 1-2 hours in section, and 12 hours a week working on homework. We appreciate this is a significant investment, but the goal is for you to have the skills to make an impact by combining your knowledge of how things work and the hands-on experience to implement your ideas.

What is this "live web conference"?

We hope to engage and interact with you throughout the course. We will meet 'live' via our electronic system for lecture each week. The sessions will be recorded. This year we are using Zoom and a separate mail from the Extension school describes how to get an account. Zoom offers better quality video, easier screen sharing, downloadable videos for multiple devices, and fewer lost connections. It also offers class videos for all of us and we plan to try this out throughout the course.

In preparation the Extension school recommends a head set and microphone for the audio. I bought their recommended Logitech H290 which is a lightweight USB set with a mute button with no other audio controls that costs $20. It is vastly superior to laptop and even free-standing (even professional) non-headphone head sets. Most laptop or USB cameras should be fine for video, but the Extension school recommended the Logitech 930 that I have interpreted as the C930e which is a pricy overkill at

$96.99 on Amazon but certainly works well. Per the Extension school policy, we will require that everyone be on video during class. I've tried teaching with and without cameras and the ability to detect body language and see people is a much more positive experience for everyone so we will follow the policy.

Grades:

45%	homework assignments (estimated 5-6) that will include individual homework assignments and paired projects with the latter being more open-ended
7%	in class topic presentation with two partners
5%	paper reviews (3 throughout semester)
14%	partner adjustment: corrects for unbalanced contributions between partners
2%	participation in class or online
2%	professionalism
4%	individual recorded paper summary
9%	timed online unproctored exam a week before Thanksgiving
12%	final project

Textbooks: none required

Policies

Accessibility:

The Extension School is committed to providing an accessible academic community. The Accessibility Office offers a variety of accommodations and services to students with documented disabilities. Please visit www.extension.harvard.edu/resources-policies/resources/disability-services-accessibility for more information.

Cheating & Plagiarism:

You are responsible for understanding Harvard Extension School policies on academic integrity (www.extension.harvard.edu/resources-policies/student-conduct/academic-integrity) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting the wrong draft, or being overwhelmed with multiple demands are not acceptable excuses. There are no excuses for failure to uphold academic integrity. To support your learning about academic citation rules, please visit the Harvard Extension School Tips to Avoid Plagiarism (www.extension.harvard.edu/resources-policies/resources/tips-avoid-plagiarism), where you'll find links to the Harvard Guide to Using Sources and two free online 15-minute tutorials to test your knowledge of academic citation policy. The tutorials are anonymous open-learning tools.

Coding and Answers:

For the homework assignments, we will sometimes ask you to write code to implement a given method and record the answer for the homework with support code. If you are using a snippet of code that you did not write, it must be appropriately cited within the code. Furthermore, copying the section code with minor tweaks is not considered your own work. On the exam, copying lecture note slides, copying their text verbatim, or copying items you googled but likely do not understand will not qualify as providing your own answer. Including a bibliographic reference alone is not a substitute for citing code for any submitted work including extra credit.

When working with a partner, some students have found it useful to create git repositories for their work. However, we consider posting your homework solutions to a public git repository that makes them available to other students as a violation of the student conduct.

Furthermore, we require that the code produces the answer that you are recording, i.e. show your work. In some cases, we may ask you to apply your coded algorithms to a new data set. Again, we require that your code produce the answer that you provide.

Late homework policy:

You have 7 total grace period days for the semester that can be applied in 1-day increments to any homework, but cannot be applied to presentations, the exam, or the final project. A maximum of 3 days can be used for any single homework. After the grace period has been used, homework grades drop by 25% per day. For team/partner work, all team members are charged as if it was an individual homework. This has led to some interesting issues where one partner has 3 days remaining and the other has 1 day remaining, so plan ahead and avoid this issue.

In class topic presentation with two partners

The course has a selected set of topics that complement the prerequisites to provide a solid foundation into the primary techniques of the field. However, machine learning is an area of active research with new developments almost weekly. We will have students prepare in small teams of 3 and present a topic that will not be covered in the course or its prerequisites. We will approve and guide topics through a sign-up since they will be in flux throughout the semester. Each team will prepare their work earlier in the semester and give a 15 minute presentation during one of the lectures. The topic could be slides or a notebook with significant text/bullets (not comments) intended to educate the classmates as well as original not-copied code demonstrating the concept. Both will be made available to fellow students as a resource perhaps for the final project. Please send your slides and/or code/notebook that you prepared to one of the TAs on the day of your presentation.

Participation:

We encourage active participation throughout the course. To provide a minor incentive, we will have Piazza, lectures, and sections for questions and discussions. We will not take roll call in any lecture or section, but we expect you to find a way to speak/chat live or through Piazza by asking questions, providing answers, sharing articles, etc. as a way of contributing to the community of CSCI E-82.

Professionalism:

We're trying out this concept. We expect almost everyone will receive the full points. However, we have had encountered some instances recently where individuals failed to show appropriate respect for their fellow students and teaching staff that we expect from Harvard Extension students.

Partner work & Partner adjustment:

Many of the open-ended assignments are challenging and will be difficult to complete individually. With many different aspects of machine learning and different approaches to solve problems, we expect partners to contribute equally. Ideally, this will always happen but it is important to fairly assess contributions, regardless. We will provide a standard template for submitting feedback on your partner's efforts relative to your own. Since both partners receive the same grade for the assignment, the Partner adjustment score will be used to correct for an imbalance of efforts.

For a few outstanding students:

Some focused students are taking this course with a particular application in mind, or perhaps an intended publication. We are willing to work with you to adjust some of the assignments to help you analyze your application using multiple techniques. We cannot offer this to a large number because it will be time consuming to provide feedback and grade.

Extra Credit:

Past students have frequently requested extra credit, often toward the end of the course. Since the goal of the class is to put machine learning to work, we will provide up to 5% extra credit total grade for one or two extra credit submissions combined. The first submission deadline is November 1. The second submission deadline is Thanksgiving so that we have time to grade them and since the focus of the rest of the semester should be on the final project. To set a standard, applying a homework assignment to a different data set or switching out a method here and there is likely not worth submitting.

Paper Summary:

All students will provide a 5 minute video summary of a research paper in the field. These are intended to focus on new research in the field. We will have a sign-up for these and everyone will review a different paper of your choosing that is related to the course. We recommend conference papers from KDD, NIPS, ICDM, ICML but journal or other papers are welcome. We ask that you avoid field-dependent or applied work (such as the application of K-Means to chemical compounds) and focus more on new techniques. Ideal papers are extensions of methods presented in lecture. These will be due on a rolling basis throughout the semester.

Paper Reviews:

Your fellow students will be summarizing papers. Following their video release, you will review up to 3 papers throughout the semester. Your task is not to summarize the paper, but to critique the content of the paper. Your critique should identify a limitation of their work and a way that you would approach it differently as in a final project or follow-on paper. Critiques of insufficient data or not enough comparisons to other techniques since they are always true. You will have to think about other aspects. The total length of the critique should be ½ page single spaced. We will have deadlines for these critiques throughout the semester so that you can receive feedback before going to the next one. You will also sign up to critique a paper so everyone doesn't select the same paper.

Currently Planned Topics

This is still in flux as we're trying to avoid topics that overlap with other courses and provide the most value to everyone.

- Introduction
  - Overview of Machine Learning field
- Statistics
  - Regression
  - Time series analysis
- Clustering
  - Distance metrics, leader, Jarvis-Patrick, scaling hierarchical clustering
  - Self-organized maps, GMM, and more advanced methods
- Dimensionality Reduction: PCA mathematical review, Sammon's, t-SNE, UMAP
- Supervised Methods
  - Classification: boosting, bagging, ensemble methods, random forests
  - Support vector machines review
  - Neural networks review
  - Deep learning: CNN
  - Genetic algorithms & genetic programming
  - Active learning
- Recommender systems: Matrix-based methods and factorization
- Frequent Pattern mining: APRIORI algorithm
- Application Areas
  - Text mining: topic detection, retrieval, classification, sentiment analysis
  - Image processing
  - Outlier detection
- Additional areas:
  - Network analysis
  - Reinforcement learning (probably briefly)

Introductions to Standard Tools (TBD):

Python with Pandas, MatplotLib, IPython
Python Natural Language Toolkit NLTK
Keras
XGBoost

Frequently Asked Questions (FAQ)

Do I have enough math background?

Although we will cover a fair range of math including statistics, calculus including partial differential equations and linear algebra, the point of the lectures is to understand the concepts. If your math is a bit rusty, you should be fine. If you've never been exposed to the topics, it would be challenging to follow the concepts through the math and you may consequently not grasp the concepts. However, most of the coding you will be doing will involve more algebra and statistics than the more advanced topics

Is the class difficult? How much time will it take?

We hope to provide you with background and experience in this course so that you will have confidence to tackle meaningful projects in the course and in your future. This will require an investment of time on your part that we estimate to be 12-15 hours/week depending on your skills. Past students have suggested it's on the high side of that range and some have reported spending much more time. We're aiming for challenging and rewarding.

I will be traveling during week X…

Lectures and sections will be recorded. Plan to attend the final 1-2 lectures for final projects where you will definitely be presenting. The paper presentation scheduling can be negotiated based on your topic and possible switching with classmates. The exam will not require a proctor and can be done anywhere there is internet and it has a ~2 hour period within a ~2 day window.
How does it fit with CSCI E-81 and CSCI E-63C?

I took CSCI E-81 a few years ago. Can I take this course too?

I taught CSCI E-81 (Machine Learning & Data Mining) a few years ago and it had a significant overlap with CSCI E-63C. The new course is designed to be a sequel to the CSCI E-63C that is based around a standard textbook used in courses across the world. Due to significant overlap with the CSCI E-81, we are currently not allowing students to take both CSCI E-81 and CSCI E-82 for credit.

How does it fit with other machine learning and data science courses?

As these very active fields grow and develop into various areas, the courses tend to adapt and follow the fields. Harvard Extension now has several courses in the space with different emphases and overlaps that change each offering with the field. Many machine learning courses can provide significant depth but little hands-on. Data science courses tend to cover a wide range from data wrangling to visualization to machine learning with limited depth. This course focuses on the theory in the lectures and the practice in sections giving it a balance. That said, the emphasis is on conceptually understanding a set of algorithms and how to apply them across fields. I don't have much additional information on specific courses.

What if I am unable to enroll as one of the first X students?

There will be a waiting list and that will probably fill as well preventing further enrollment. We will work with the Extension school to include as many as we can, while still offering as high quality as a new class. The enrollment tends to change over the first week of the course and many have been able to enroll by the end of the first week. The Extension school controls the waiting list so I cannot alter the order of who is included.