Welcome

Instructor:
Theodore J. LaGrow, **tlagrow@gatech.edu (mailto:tlagrow@gatech.edu)**

Head TAs:
Dan Boros, **boros@gatech.edu (mailto:boros@gatech.edu)**

(mailto:boros@gatech.edu) John Mansfield, **jmansfield6@gatech.edu (mailto:jmansfield6@gatech.edu)**

(mailto:jmansfield6@gatech.edu) Sunmin Lee, **sunmin@gatech.edu (mailto:sunmin@gatech.edu)**

(mailto:sunmin@gatech.edu) Danyang Cai, **dcai38@gatech.edu (mailto:dcai38@gatech.edu)**

Creators of Recorded Material:
Charles Isbell
Michael Littman

Office Hours:
See Ed Discussion for details.

Required Text:
Machine Learning
by Tom Mitchell, McGraw Hill, 1997

Interestingly, now that he completely owns the copyright, Tom has made the textbook *free* and online:
http://www.cs.cmu.edu/afs/cs.cmu.edu/user/mitchell/ftp/mlbook.html (http://www.cs.cmu.edu/afs/cs.cmu.edu/user/mitchell/ftp/mlbook.html)

Video Lectures:
"Modules" in Canvas
Beta: "Ed Lessons" in Canvas

General Information

Machine Learning is a three-credit course on the study and application of the field of Machine Learning. Machine Learning is that area of Artificial Intelligence that is concerned with computational

artifacts that modify and improve their performance through experience. The area is concerned with issues both theoretical and practical. This particular class is a part of a series of classes in Machine Learning and takes care to present algorithms and approaches in such a way that grounds them in larger systems. We will cover a variety of topics, including statistical supervised and unsupervised learning methods, randomized search algorithms, and Bayesian learning methods The course also covers theoretical concepts such as inductive bias, the PAC and Mistake-bound learning frameworks, minimum description length principle, and Ockham's Razor. In order to ground these methods, the course includes some programming and involvement in a number of projects.

Due to the reduced timeline of the Summer term, we have removed reinforcement learning as required material.

Objectives

There are four primary objectives for the course:

- To provide a broad survey of approaches and techniques in ML
- To develop a deeper understanding of several major topics in ML
- To develop the design and programming skills that will help you to build intelligent, adaptive artifacts
- To develop necessary skills to communicate research and practicum in ML

The last objective is the core one: you should develop enough background that you can pursue any desire you have to learn more about specific techniques in ML, either to pursue ML as a research career or to apply ML techniques in other research areas.

Prerequisites

The official prerequisite for this course is an introductory course in artificial intelligence. In particular, those of you with experience in general representational issues in AI, some AI programming, and at least some background in statistics and information theory should be adequate. Any student who did well in an introductory AI course should be fine. You will note that most semi-modern AI courses suggests at least some tentative background in some machine learning techniques as well. Of course, having said all that, the most important prerequisite for enjoying and doing well in this class is your interest in the material. We say this in every semester and in every course, but it's true. In the end, it will be your own motivation to understand the material that gets you through it more than anything else. If you are not sure whether this class is for you, please contact either of the instructors.

Resources

- **Readings.** The textbook for the course is *Machine Learning* by Tom Mitchell. We will follow the textbook quite closely for most of the semester, so it is imperative that you have a copy of the book. We will also use supplemental readings as well, but those will be provided for you.
- **Computing.** Even though you absolutely will not need it, you will have access to CoC clusters for your programming assignments. You can test your code on the Shuttles cluster, using your GT username and password to log in - you will not need a CoC account for this course. More info can be found here: **https://support.cc.gatech.edu/facilities/general-access-servers (https://support.cc.gatech.edu/facilities/general-access-servers)** .
- **Web.** We will use the class canvas page and Ed Discussions to post last-minute announcements, so check them early and often.

Statement of Academic Honesty

At this point in your academic careers, we feel that it would be impolite to harp on cheating, so we won't. You are all adults and are expected to follow the university's code of academic conduct (**honor code (https://policylibrary.gatech.edu/student-life/academic-honor-code)**). Furthermore, at least some of you are researchers-in-training, and we expect that you understand proper attribution and the importance of intellectual honesty.

We should also point out that "proper attribution" does not absolve the writer of the "intellectual honesty" that comes from original writing. While it is definitely the case that copying text without attribution is considered plagiarism, it is also the case that copying too much text even with attribution betrays a lack of intellectual honesty. Too many quotes of more than, say, two sentences will be considered plagiarism and a terminal lack of academic originality. Do not overthink this issue, but do not under think it either.

Please note that unauthorized use of any previous semester course materials, such as tests, quizzes, homework, projects, videos, and any other coursework, is prohibited in this course. In particular, you are not allowed to use old exams. Using these materials will be considered a direct violation of academic policy and will be dealt with according to the GT Academic Honor Code. Furthermore, we do not allow copies of my exams outside of this course. Just as you are not to use the previous material you are not to share current material with others either now or in the future. Our policy on that is strict. If you violate the policy in any shape, form or fashion you will be dealt with according to the GT Academic Honor Code.

Readings and Lectures

The online lectures are meant to summarize the readings and stress the important points. You are expected to critically read any assigned material. Your active participation in the material, the lectures, and office hours are crucial in making the course successful. We want to stress that the more you put into the material, the more you will get out. The full teaching staff is to assist you in learning and growing in the area of machine learning.

To help you to pace yourself, we have provided a nominal schedule (see the Syllabus link) that tells you when we would be covering material if we were meeting twice a week during the term. We recommend you try to keep that pace.

Scoring and Grading

Your final grade is determined by how you do on several components: a reading/writing quiz, a hypothesis quiz, 3 comprehensive assignments, and a cumulative final exam.

- **Reading/Writing Quiz.** In an effort to improve the course, I have included a quiz to help with reading and writing formal reports. At a graduate level in a terminal program, it is not enough to be able to apply advanced algorithms but understand why you make specific design choice and comment on nuanced caveats on your results. This is a key aspect of the course you will not find in other online Machine Learning classes. The teaching staff prides ourselves in providing detailed feedback and we consistently hear that this skill can be immediately applied to many positions students currently hold. Therefore, I have included a lecture and quiz to help establish many of these skill you will practice in your reports over the course of the semester. You will have unlimited attempts to earn the points on the quiz. The lecture material and quiz will be available at the beginning of the first week.
- **Hypothesis Quiz.** To help supplement the Reading/Writing Quiz, I have provided an exercise on hypothesis interpretation and development. There is a large disparity in understanding term-to-term, so this should be great practice with feedback before the first assignment is due. The quiz will be available at the end of the first week.
- **Assignments.** There will be three scored assignments, one for the first section and two for the second sections. They will be about programming and analysis. Generally, they are designed to give you deeper insight into the material and to prepare you for the exams. The programming will be in service of allowing you to run and discuss experiments, do analysis, and so on. In fact, the programming is incidental, as you shall see.

  When your assignments (projects and exams) are scored, you will receive feedback explaining your errors (and your successes!) in some level of detail. This feedback is for your benefit, both on this assignment and for future assignments. It is considered a part of your learning goals to internalize this feedback. This is one of many learning goals for this course, such as understanding how to analyze data or the differences between each algorithm or bias/variances in each of them.

  If you are convinced that your score is in error in light of the feedback, you may ask for additional feedback on the assignment for clarify of comments. We will not be conducting rescores this term as the feedback follow up are significantly more beneficial to previous cohorts. Be concrete and specific as this will help the discussion.

- **Final Exam.** There will be a written, closed-book final exam at the end of the term. The final exam will also be administered via whatever our proctoring solution is this term.

** Please note, in previous semesters we have administered a midterm. I have made a decision to remove this from the course as the exam was not conducive to a healthy student workload balance. There is additional emphasize on the reports and communication this semester.*

*** Please note for Summer, we will not have an A4 which typically centers around reinforcement learning. Reducing the schedule from 17 weeks to 13 weeks prevents this final unit in this compact schedule. The material will be available but not required.*

Due Dates

All scored assignments are due by the time and date indicated. Here "time and date" means **Eastern Time**. If you are in another time zone, you should probably go to settings on Canvas and set your time zone appropriately. We will not accept late assignments or makeup exams. You will earn zero credit for any late assignment. The only exceptions will require: a **note** from an appropriate authority and **immediate notification** of the problem when it arises. Your excuse must be acceptable.

Numbers

Component	
Reading/Writing Quiz	5%
Hypothesis Quiz	5%
Assignments	60%
- A1	(20%)
- A2	(20%)
- A3	(20%)
Final	30%

Extra Credit

There are two opportunities to receive extra credit in this course. We intend to provide an additional comprehensive problem set that you will be able to turn in before the Final Exam. We will provide answers but will not score the set. The problem set will help prepare you for the exam. If all the problems are attempted and turned in, we will award 1% to your overall course percentage. These will be calculated on Canvas later in the term but before final grades are released.

Additionally, if there are significant contributions to the Ed Discussion board throughout the term, we will award 1% to your overall course percentage. "Significant contributions" will up to our discretion,

however interacting with Ed Discussions will only benefit you over the course of the term. These additional points will be calculated on Canvas later in the term but before final grades are released.

Office Hours and Other Channels

We love the assignments in this course. As you will discover they are wonderfully open-ended, much more so than many of you will be used to. It is therefore important that in addition to watching the lectures and comprehending the required readings that you attend office hours and regularly check Ed Discussions. We will record Office Hours but strongly suggest to interact both on Ed and in the live Office Hours. These are only for your benefit. You should consider your participation in both required.

Disclaimer

We reserve the right to modify any of these plans as need be during the course of the class; however, we won't do anything capriciously, anything we do change won't be too drastic, and you'll be informed as far in advance as possible.

Reading List

Required Text:

- **Tom Mitchell, Machine Learning. McGraw-Hill, 1997. (http://www.cs.cmu.edu/afs/cs.cmu.edu/user/mitchell/ftp/mlbook.html)**

Optional Text:

- **Larry Wasserman, All of Statistics. Springer, 2010. (http://www.stat.cmu.edu/~larry/all-of-statistics/)** (Read Part 1 for an intro to Probability Theory)
- Richard Sutton and Andrew Barto, Reinforcement Learning: An introduction. (for Reinforcement Learning) (**Nov 5, 2017 version (http://incompleteideas.net/book/bookdraft2017nov5.pdf)**)

A List:

- Linear Algebra
  - **Linear Algebra and Eigenproblems (https://github.com/pushkar/4641/raw/master/downloads/Eigenproblems.fm.pdf)**
- Lessons (Sidebar: Ed Lessons)
  - ML is the ROX
    - Mitchell Ch 1
  - Decision Trees
    - Mitchell Ch 3
  - Regression and Classification
  - Neural Networks

    - Mitchell Ch 4
- Instance-Based Learning
    - Mitchell Ch 8
- Ensemble Learning
    - **Schapire's Introduction (https://github.com/pushkar/4641/raw/master/downloads/boosting.ps)**
    - **Jiri Matas and Jan Sochman's Slides (https://github.com/pushkar/4641/raw/master/downloads/adaboost_matas.pdf)**
- Kernel Methods and SVMs
    - **An introduction to SVMs for data mining (https://www.cc.gatech.edu/classes/AY2008/cs7641_spring/handouts/yor12-introsvm.pdf)**
    - **Christopher Burges tutorial on SVMs for pattern recognition (https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/svmtutorial.pdf)**
    - **Scholkopf's NIPS tutorial slides on SVMs and kernel methods (https://github.com/pushkar/4641/raw/master/downloads/svm-scholkopf.ps)**
- Computational Learning Theory
    - Mitchell Ch 7
- VC Dimensions
    - Mitchell Ch 7
- Bayesian Learning
    - Mitchell Ch 6
- Bayesian Inference
- Randomized Optimization
    - Mitchell Ch 9
    - **No Free Lunch Theorem (https://ml-cs7641.s3.us-east-1.amazonaws.com/nfl-optimization-explanation.pdf)**
- Clustering
    - Mitchell Ch 6
    - **Intuitive Explanation of EM (http://www.cc.gatech.edu/~dellaert/em-paper.pdf)**
    - **Statical View of EM (https://github.com/pushkar/4641/raw/master/downloads/em.pdf)**
    - **Jon Kleinberg's Impossibility Theorem for Clustering (https://www.cs.cornell.edu/home/kleinber/nips15.pdf)**
- Feature Selection
    - **ICA: Algorithms and Applications (http://mlsp.cs.cmu.edu/courses/fall2012/lectures/ICA_Hyvarinen.pdf)**
    - **Restructuring High Dimensional Data by Charles and Paul Viola (https://www.cc.gatech.edu/~isbell/papers/isbell-ica-nips-1999.pdf)**
- Feature Transformation
- Information Theory

        - Charles Isbell's Note on Information Theory (https://www.cc.gatech.edu/~isbell/tutorials/InfoTheory.fm.pdf)
        - An Introduction to Information Theory and Entropy (https://github.com/pushkar/4641/raw/master/downloads/gentle_intro_to_information_theory.pdf)

The following are Optional but still available. These are outside the scope of the Summer term and will not be include as 'comprehensive' on the Final Exam.

- Markov Decision Processes
- Reinforcement Learning
    - Mitchell Ch 13
    - Richard Sutton and Andrew Barto, Reinforcement Learning: An introduction. MIT Press, 1998. (http://incompleteideas.net/book/bookdraft2017nov5.pdf)
    - Reinforcement Learning: A Survey (https://github.com/pushkar/4641/raw/master/downloads/kaelbling96reinforcement.pdf)
- Game Theory
    - Andrew Moore's slides (http://www.cs.cmu.edu/~awm/tutorials.html) (http://www.cs.cmu.edu/~awm/tutorials.html)
- Outro

Software

- WEKA (https://ml.cms.waikato.ac.nz/weka) Machine learning software in JAVA that you can use for your projects
- Data Mining with Weka (https://weka.waikato.ac.nz/) A MOOC Course
- ABAGAIL (https://github.com/pushkar/ABAGAIL) Machine learning software in JAVA. This is hosted on my github, so you can contribute too
- scikit-learn (http://scikit-learn.org/stable/) A popular python library for supervised and unsupervised learning algorithms
- pybrain (http://pybrain.org/) A popular python library for artifical neural networks
- Murphy's MDP Toolbox for Matlab (http://www.cs.ubc.ca/~murphyk/Software/MDP/mdp.html)
- MATLAB Clustering Package (http://www.cc.gatech.edu/~dellaert/FrankDellaert/Software.html) By Frank Dellaert (http://www.cc.gatech.edu/~dellaert/FrankDellaert/Frank_Dellaert/Frank_Dellaert.html)

Datasets

- UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) An online repository of data sets that can be used for machine learning experiments.
- Stanford Large Network Dataset (http://snap.stanford.edu/data/) Dataset of large social and information networks.

- **Vision Benchmark Suite** (http://www.cvlibs.net/datasets/kitti/index.php) Autonomous car dataset

Version Control

- 05/06/24: TJL updating syllabus for the reduced Summer term.