ECE 6254: Statistical Machine Learning
Spring 2024 Syllabus

Summary

This course will provide an introduction to the theory of statistical learning and practical machine learning algorithms. We will study both practical algorithms for statistical inference and theoretical aspects of how to reason about and work with probabilistic models. We will consider a variety of applications, including classification, prediction, regression, clustering, modeling, and data exploration/visiualization.

Prerequisites

Throughout this course we will take a statistical perspective, which will require familiarity with basic concepts in probability (e.g., random variables, expectation, independence, joint distributions, conditional distributions, Bayes rule, and the multivariate normal distribution). We will also be using the language of linear algebra to describe the algorithms and carry out any analysis, so you should be familiar with concepts such as norms, inner products, orthogonality, linear independence, eigenvalues/vectors, eigenvalue decompositions, etc. as well as the basics of multivariable calculus such as partial derivatives, gradients, and the chain rule. If you have had courses on these topics as an undergraduate (or more recently) you should be able to fill in any gaps in your understanding as the semester progresses. Finally, many of the homework assignments and the course projects will require the use of Python. Prior experience with Python is not necessary, but I am assuming a familiarity with the basics of scientific programming (e.g., experience with C, MATLAB, or some other programming language).

Instructor

Mark Davenport
Email: mdav@gatech.edu
Office: Coda, S1117
Phone: (404)894-2881
Office Hours: TBD

Teaching Assistants

TBD

Lectures

Lectures are Tuesdays and Thursdays from 9:30-10:45am and will be held in Clough Commons, Room 152. Lectures will also be available online live at (LINK TBD) and after a brief delay will be made available in Canvas (typically later that day).

Instruction Modality

In Spring 2024, ECE 6254 will be taught in a defacto hybrid mode. My intention is that all lectures will be delivered in-person. I would like to strongly encourage in-person attendance – I think that makes for a much better experience for both you and me. However, all lectures will be recorded and available to watch after class. Note: with the exception of distance students, exams will be in-person.

Grading

Your grade will be based on the following factors:

- **Pre-test (5%):** During the first week of class there will be a take-home "pre-test" which will review the basic concepts from calculus, linear algebra, probability theory, and programming that we will be using in this course. This is an open-book/internet test, so you should feel free to consult whatever outside resources you like, but **you must work through this test on your own.** The purpose of this pre-test is to help everyone get on the same page in terms of what you need to know in order to succeed in this course.
- **Homework (25%):** There will be 7 ± 1 homework assignments. They will consist of exercises, proofs, and Python implementations. See further details below.
- **Data challenges (10%)**: There will be 3 ± 2 "data challenges" where I will distribute a dataset and you will be tasked with making some kind of prediction. You will need to submit both an actual prediction on some test data and also a short description justifying the approach that you took. You will be judged primarily on the justification/reasoning behind your approach, but you will also receive bonus points for accuracy. These can be done individually or in groups.
- **Final project (20%):** A major component of this course consists of an in-depth project on a topic of your choosing. These projects will be done in groups of 3–5 students. The project will have several graded components, including a detailed (written) project proposal, a presentation, and a written report. The project proposal is tentatively scheduled to be due on **March 26**. The presentations will consist of a short pre-recorded video. The written report will be due at the end of the finals period. Further details about the project will be provided later in the semester.

- **Midterm exam (20%):** The midterm exam will occur in-class on **March 7**) and will cover the same material as the homeworks submitted before the exam.
- **Final exam (20%):** The final exam will be comprehensive – covering all the material as the homeworks throughout the semester – and will occur at the designated time during the finals period: **April 29, 8:00am–10:50am**. Note that I cannot accommodate requests to take the exam at alternative time slots.

Your final grade will be assigned as a letter grade according to the scale:

A: 90-100% B: 80-89% C: 70-79% D: 60-69% F: $\leq 59\%$

I may exercise the option to "curve" exam scores as necessary (by adjusting the grades higher, but not lower) if I determine that an exam was more difficult than intended.

Homework

Homework will be assigned weekly (approximately). **Homework will be turned in via Gradescope. Extensions are available on a case-by-case basis, but once the solutions are posted late submissions will not be accepted.** Each homework assignment will be graded out of 100 points. Over the course of the semester, **the maximum number of homework points that you can earn is** $(N-1)\cdot 100$**, where** N **is the number of assignments** – this serves a similar role to allowing you to drop one homework assignment, but should encourage you to still submit a partially completed one (and avoid panic if there is occasionally a problem that you do not finish in time.) Because of the large class size, the TAs will reserve the right to grade only a subset of the assigned problems as necessary.

The homework assignments will be hard; many of them will require significant amounts of time and effort to complete. But this is really where most of the learning takes place. You will get out of the assignments what you put into them. Students who complete all of the assignments in full will be rewarded with a deep understanding of machine learning algorithms and theory. Effectively, homework is worth much more than 25% of your grade. In teaching many courses over the years, **I have never seen a case where a student does not put effort into the homework assignments but does well on the exams**.

Students are *strongly* encouraged to discuss homework problems with one another. However, **each student must write up and turn in their own solutions written in their own words**. **Cases where solutions appear to be identical or nearly identical will be immediately referred to the Office of Student Integrity.**

Unauthorized use of any previous semester course materials, such as tests, quizzes, and homework, is prohibited in this course. Furthermore, redistributing materials from this semester is also prohibited. For any questions involving these or any other Academic Honor Code issues, please consult me or www.honor.gatech.edu.

Text

There is no required text. Course notes will be posted as they become available at the course website. These notes will be based on material sourced from several different texts. Some of the main texts I have found useful include:

- Abu-Mostafa, Magdon-Ismail, and Lin: *Learning from Data.* (Available at amazon. See also the related online course.)
- Hastie, Tibshirani, and Friedman: *The Elements of Statistical Learning.* (Available online as a pdf, free and legal.)
- Murphy: *Machine Learning: A Probabilistic Perspective* (Available at amazon.)
- Watt, Borhani, and Katsaggelos: *Machine Learning Refined* (Available at amazon.)

There are many other books and journal papers of interest which will be listed in the "Resources" section of the course web site.

Online resources

The course webpage is at:

http://mdav.ece.gatech.edu/ece-6254-spring2024

This page will provide general course information, copies of the lecture notes, resources (links to other site, books, and papers) that augment the lectures, and homework assignments. I will also use Canvas to distribute some additional materials.

In this course I also plan to make frequent use of Piazza to make announcements and answer questions. This site can be accessed via Canvas or via:

https://piazza.com/gatech/spring2024/ece6254

Piazza is a great platform for you to work with your fellow students to discuss problems, form study/project groups, etc. Please direct any questions you might have to Piazza (as opposed to my email, where it is likely to get lost). Unless your questions are personal in nature, please do not make private posts – if you have a question you are probably not the only one, and other students may benefit from seeing the discussion.

Distance learning students

Distance learning students will be required to complete the same assignments as the on-campus students, but with a one week delay on due dates. Note that no late submissions will be allowed as I would like to be able to post solutions and discuss problem without an excessive delay. Distance students must also form project groups and will be required to give presentations in the format of a pre-recorded video. Please contact GTPE for any questions regarding exam scheduling/procedures.

Course Expectations and Guidelines

Academic integrity
Georgia Tech aims to cultivate a community based on trust, academic integrity, and honor. Students are expected to act according to the highest ethical standards. For information on Georgia Tech's Academic Honor Code, please visit www.catalog.gatech.edu/policies/honor-code. Any student suspected of cheating or plagiarizing on a quiz, exam, or assignment will be reported to the Office of Student Integrity, who will investigate the incident and identify the appropriate penalty for violations.
Redistributing materials from this course and/or using external sites for assistance (e.g., contributing to test banks, CourseHero, Chegg, or similar sites) is prohibited.

Collaboration and group work
Students are *strongly* encouraged to discuss homework problems with one another. However, **each student must write up and turn in their own solutions written in their own words**. **Cases where solutions appear to be identical or nearly identical will be immediately referred to the Office of Student Integrity.**

Absences, late assignments, and missed exams
Active participation in the class discussions is expected. Please attend class (either in person or online) unless you have a compelling reason not to do so. However, you will not be penalized for any excused absences (e.g., due to illnesses, religious observances, career fairs, job interviews, etc.) **Late assignments cannot be accepted** in the absence of prior approval. In the event that an excused absence prevents you from submitting an assignment, your homework grade will be calculated on a pro-rated basis. **Exams will be completed in-person. If you expect to miss an exam, please contact me as soon as you realize this so we can make alternative arrangements.** We may consider options to take the exam at an alternate time or instead may adjust the grading allocation to place more emphasis on other exams, depending on the circumstances.

Accommodations for students with disabilities
If you are a student with learning needs that require special accommodation, contact the Office of Disability Services at (404)894-2563 or disabilityservices.gatech.edu, as soon as possible, to make an appointment to discuss your special needs and to obtain an accommodations letter. Please also e-mail me as soon as possible in order to set up a time to discuss your learning needs.

Student-Faculty expectations agreement
At Georgia Tech we believe that it is important to strive for an atmosphere of mutual respect, acknowledgement, and responsibility between faculty members and the student body. In the end, simple respect for knowledge, hard work, and cordial interactions will help build the environment we seek. Therefore, I encourage you to remain committed to the ideals of Georgia Tech while in this class. See www.catalog.gatech.edu/rules/22 for an articulation of some basic expectation that you can have of me and that I have of you.

Outline

- Theory of generalization
  - Introduction to classification
  - Concentration inequalities and generalization bounds
  - The Bayes classifier and the likelihood ratio test
  - Nearest neighbor classification and consistency
  - Vapnik-Chervonenkis (VC) dimension
  - VC generalization bounds
  - Bias-variance tradeoff
  - Overfitting
- Supervised learning
  - Linear classifiers
    - plugin classifiers (linear discriminant analysis, logistic regression, Naïve Bayes)
    - the perceptron algorithm and single-layer neural networks
    - maximum margin principle, separating hyperplanes, and support vector machines (SVMs)
  - From linear to nonlinear: feature maps and the "kernel trick"
  - Kernel-based SVMs
  - Regression
    - least-squares
    - regularization
    - the LASSO
    - kernel ridge regression
  - Model selection, error estimation, and validation
- Unsupervised learning
  - Feature selection
  - Dimensionality reduction
    - principle component analysis (PCA)
    - multidimensional scaling (MDS)
    - manifold learning
  - Latent variables and structured matrix factorization
    - non-negative matrix factorization
    - sparse PCA
    - dictionary learning
    - latent semantic indexing, topic modelling

        * matrix completion
    - Density estimation
    - Clustering
        * k-means
        * Gaussian mixture models and expectation-maximization
        * spectral clustering
- Advanced supervised learning
    - Decision trees
    - Ensemble methods
    - Random forests
    - Multi-layer neural networks and backpropagation
    - Deep learning
- Further topics (Important things that we may or may not get time to cover)
    - Graphical models
    - Reinforcement learning
        * Markov decision processes
        * optimal planning
        * learning policies