ECE 6254: Statistical Machine Learning
Spring 2017 Syllabus

Summary

This course will provide an introduction to the theory of statistical learning and practical machine learning algorithms with applications in signal processing and data analysis. In contrast to most traditional approaches to statistical inference and signal processing, in this course we will focus on *how to learn effective models from data* and *how to apply these models to practical signal processing problems*. We will approach these problems from the perspective of statistical inference. 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: Centergy One, Room 5212
Phone: (404) 894-2881
Office Hours: I will typically hold scheduled office hours the day before homework is due. More details will be provided soon. I am also available to meet in Centergy 5212 by appointment.

Teaching Assistants

TBA

Text

The material for this course will come from several different sources. I will not require you to purchase any specific text, but the primary sources for the course are:

- Abu-Mostafa, Magdon-Ismail, and Lin, *Learning from Data*, 2012. (Hardback available for $28 on amazon. See also the related online course.)
- Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning*, 2011. (Available online as a pdf, free and legal.)
- Murphy, *Machine Learning: A Probabilistic Perspective*, 2012. (Hardback available for $\sim$ $90 on 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-spring2017

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 may also use T-square as an additional option to distribute some of the same 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 T-square or via:

https://piazza.com/gatech/spring2017/ece6254/home

Piazza can also provide a good platform for you to work with your fellow students to discuss problems, find study groups, sketch out project ideas, etc. Please direct any questions you might have to Piazza before trying to contact me via e-mail.

Finally, as this course also has an online enrollment, all of my lectures are being recorded. These lectures will be made available to the on-campus students as well and can be accessed through T-square. I am providing access to the recorded lectures because they might be useful as study materials and to help out when missing class is unavoidable. I am *not* meaning to suggest that you can simply watch the online lectures in lieu of attending class. If I feel that the on-campus students are abusing this privilege, I reserve the right to remove access to these materials.

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 8 ± 1 homework assignments. They will consist of exercises, proofs, and Python implementations. I expect your write-ups to be very clear; I do not just want you to produce correct answers — I want you to demonstrate that you understand the material and write your solutions as if you were explaining your answer to a colleague. Style matters and will be a factor in the grade.

  In the absence of prior arrangements or a valid documented excuse, late homework will get zero credit.

  You are encouraged to discuss the homework with other members of the class. However, everything that you turn in must be your own work. **You must write up the assignments (and accordingly the Python code) by yourself, citing any outside references you use to arrive at your solution.**

- **Midterm exam (20%):** The midterm exam will occur in-class at about mid-semester (tentatively scheduled for **March 9**) and will cover the same material as the homeworks.

- **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: **May 2, 6:00pm–8:50pm**.

- **Final project (25%):** 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 4–5 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 16**. The presentations will be in the format of a poster session, tentatively scheduled for the last day of class (**April 25**). 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.

- **Participation (5%):** This part of your grade is based on my assessment of your engagement in the course. This will be based on factors such as attendance, participation in classroom discussions, engagement outside of the classroom (such as during office hours and/or on Piazza), and on peer evaluations from project groups.

Unauthorized use of any previous semester course materials, such as tests, quizzes, homework, and any other coursework, is prohibited in this course. Furthermore, redistributing materials from this semester (e.g., contributing to test banks, CourseHero, or similar sites) is also prohibited. For any questions involving these or any other Academic Honor Code issues, please consult me or www.honor.gatech.edu.

Distance learning students

Distance learning students (Sections Q and QSZ) will generally be required to complete the same assignments as the on-campus students, but with a delay of up to one-week to accommodate any delays in the posting of the lectures. These delays will apply to all homework/project/exam dates **except that I would like the final projects and final exams to be completed and submitted by 11:59pm on May 4.** This will allow me to grade the projects/exams for all sections at the same time and ensure a timely release of your grades.

Distance students must also form project groups, although alternate arrangements for handling the group project presentations will be considered as necessary. All lecture materials/handouts will be available through some combination of GT Courses, T-Square, and the course website. Please contact GTPE or the appropriate professors on the Shenzhen campus for any questions regarding exam scheduling/procedures.

Outline

- Introductory supervised learning
  - Concentration inequalities and generalization bounds
  - The Bayes classifier and the likelihood ratio test
  - Nearest neighbor classification and consistency
  - 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
  - Theory of generalization (Part I)
    - Vapnik-Chervonenkis (VC) dimension
    - VC generalization bounds
  - Regression
    - least-squares
    - regularization
    - the LASSO
    - kernel ridge regression
  - Theory of generalization (Part II)
    - overfitting
    - bias-variance tradeoff
    - 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
    - bootstrap aggregating (bagging)
    - boosting
    - stacking
  - Random forests
  - Multi-layer neural networks and backpropagation
  - Deep learning
- Further topics (Important things that we will probably not get time to cover)
  - Graphical models
  - Reinforcement learning
    - Markov decision processes
    - optimal planning
    - learning policies

A *very tentative* schedule can be viewed at bit.ly/2iAHMmM. This schedule is subject to complete and arbitrary revision at any moment, but it should give you a rough idea of what to expect this semester.