ECE 4813: Mathematical Foundations of Data Science

Summary

The purpose of this course is to introduce students to two fundamental pillars of data science: statistical inference and optimization. The algorithms, modeling techniques, and mathematics from these two fields will be introduced through a series of case studies that use real-world data.

Instructor

Mark Davenport

Office Hours: TBD

Teaching Assistant

TBD

Lecture

TBD

Prerequisites

MATH 1553/1554 or MATH 2605, and ECE 3077 or ISYE 3770 or a similar class on introductory probability and statistics.

Students are expected to have a working knowledge of linear algebra and probability. In particular, students should be familiar with basic matrix-vector computations, and have had exposure to the concepts of rank, subspaces, matrix factorization, eigenvalues/eigenvectors, and solving systems of linear equations. They should also be familiar with the concepts of conditional probability, joint density functions, moments (expectation and variance), and Bayes' rule.

Homeworks and projects will require basic programming experience in MATLAB or Python.

Course Objectives

As part of this course, students...

- Formulate inference problems in the language of linear algebra and optimization. [1]
- Analyze and compute the solutions to least-squares problems in the context of regression. [1]
- Become familiar with basic computational methods from optimization. [1,6]
- Map descriptions of real-world problems into quantitative computational problems. [1,6,7]

Learning Outcomes

Upon successful completion of this course, students should be able to ...

1. Apply appropriate models to solve classical (linear, logistic, Poisson) regression problems.
2. Implement algorithms for solving unconstrained and constrained optimization problems.
3. Identify and understand the differences between different types of convex optimization problems (linear, quadratic, etc).
4. Identify and understand the differences between convex and nonconvex optimization problems.
5. Describe the computational issues in solving different kinds of statistical inference problems, and how those issues scale with the amount of data available.
6. Implement basic machine learning algorithms, including support vector machines and multilayer neural networks.
7. Use cross-validation to perform model selection.
8. Describe the difference between training error and generalization error, and the effect of sample size on each.

Textbooks

Almost all the material in the course will come from these two text books:

- G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, 2013, available online (link1, link2)
- G. C. Calafiore and L. El Ghaoui, *Optimization Models*, Cambridge University Press, 2014.

These books also make excellent resources for supplementary material:

- S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004. pdf available online
- T. Hastie, R. Tibshirani, J. Friedman, *Elements of Statistical Learning*, Springer, 2009. pdf available online
- D. G. Luenberger, *Information Science*, Princeton University Press, 2006.
- S. Raschka, *Python Machine Learning*, Packt Publishing, 2015.
- J. Grus, *Data Science from Scratch: First Principles with Python*, O'Reilly Media, 2015.

Grading

- Homework 25%, $\sim$ 8 assignments
- Midterm Exam 25%
- Final Exam 25%
- Project 20%
- Participation 5%

For the first (roughly) half of the course, we will have weekly homework, followed by a midterm exam. After that, the assignment frequency will decrease so that you can focus on completing a final project in groups of 2–3. The project will consists of a short proposal, short progress report, an oral presentation, and a 6 page (maximum) paper. The presentations will be scheduled during the last week of class. More details regarding the project will be forthcoming.

Homework will be turned in at beginning of class. Late homework will get zero credit.

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.

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.

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 (e.g., contributing to test banks, CourseHero, Chegg, or similar sites) is also 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

Attendance and active participation in the lectures is a factor in your grade. However, you will not be penalized for any excused absences (e.g., due to illnesses, religious observances, career fairs, job interviews, etc.) In the event that an excused absence prevents you from submitting an assignment, your homework grade will be calculated on a pro-rated basis. **If you expect to miss a quiz or exam, please contact me as soon as possible to make alternative arrangements.** We may consider options to take the quiz at an alternate time or instead may adjust the grading allocation to place more emphasis on other quizzes/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 of Topics

The course will use a series of case studies to anchor the exposition. In each of the sections below, a real-world data science problem will be presented, and the tools needed to solve the problem (and understand its solution) will be presented in turn. There will be a significant emphasis on *modeling*, specifically in setting up statistical inference problems as optimization programs.

1. **Least Squares**
   (a) Solving systems of equations
   (b) Singular value decompositions
   (c) Iterative methods for least-squares (gradient descent)
   (d) Stability and regularization
   (e) Principal components analysis
2. **Linear Programming**
   (a) Basic concepts: linear inequalities, feasibility, boundedness
   (b) Simplex algorithm
3. **Quadratic Programming**
   (a) Basic concepts: quadratic forms, positive definite matrices, Lagrange multipliers and testing optimality
   (b) Projected gradient descent for solving QPs
   (c) Case study topics: data-driven linear classifiers (maximum margin), portfolio optimization, imaging with positivity constraints, model selection
4. **Unconstrained Optimization**
   (a) Basic concepts: Hessian matrices, local and global minima, convexity
   (b) Newton's method
5. **Nonconvex Optimization for Neural Networks**
   (a) Basic concepts: functional approximation, backpropagation, local minima in nonconvex optimization
   (b) Generalization, in-sample versus out-of-sample error, cross validation