ECE 4813: Mathematical Foundations of Data Science

Instructor

Kiryung Lee

Goal

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.

Textbooks

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

- G. C. Calafiore and L. El Ghaoui, *Optimization Models*, Cambridge University Press, 2014. [amazon]
- G. James et al, *An Introduction to Statistical Learning*, Springer, 2013. [amazon], pdf available online at no cost.

These books also make excellent resources for supplementary material:

- S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004. [amazon], pdf available online at no cost.
- T. Hastie, R. Tibshirani, J. Friedman, *Elements of Statistical Learning*, Springer, 2009. [amazon], pdf available online at no cost.
- D. G. Luenberger, *Information Science*, Princeton University Press, 2006. [amazon]
- S. Raschka, *Python Machine Learning*, Packt Publishing, 2015. [amazon]
- J. Grus, *Data Science from Scratch: First Principles with Python*, O'Reilly Media, 2015. [amazon]

Grading

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

Prerequisites

MATH 1553/1554 or MATH 2605, and ECE 3077 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, Python, or a related language.

Learning Objectives

In this course, the students are expected to ...

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

Course Educational 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 multi-layer 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.

Academic Integrity

Academic dishonesty will not be tolerated. This includes cheating, lying about course matters, plagiarism, or helping others commit a violation of the Honor Code. Plagiarism includes reproducing the words of others without both the use of quotation marks and citation. Students are reminded of the obligations and expectations associated with the Georgia Tech Academic Honor Code and Student Code of Conduct, available online at www.honor.gatech.edu.

Learning Accommodations

If needed, we will make classroom accommodations for students with documented disabilities. These accommodations must be arranged in advance and in accordance with the Office of Disability Services (http://disabilityservices.gatech.edu).

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. The course is organized by model-type, and

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
(f) Case study topics: regression for fitting functions to data, restoring images using denoising, topic modeling in text using PCA.

2. **Linear Programming**
   (a) Basic concepts: linear inequalities, feasibility, boundedness
   (b) Simplex algorithm
   (c) Case study topics: MaxFlow for network throughput, scheduling, optimal resource allocation

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
   (c) Case study topics: maximum likelihood estimation for logistic and poisson regression, Bayesian estimation

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
   (c) Case study: large-scale supervised classification of images

6. (As time permits) **Graphical Models**
   (a) Basic concepts: graphs, adjacency matrices, Markov models, random walks, mixing time
   (b) Case study topics: PageRank, clustering