CSE 6740, Fall 2023, Georgia Tech

Computational Data Analysis: Course outline

22nd August 2023

Instructor: Nisha Chandramoorthy (nishac@gatech.edu)

In this course, we will learn the mathematical and computational foundations of machine learning methods with the goal of understanding i) how and when they do work and do not; and ii) how to use them in a principled manner. Besides neural networks, we will cover classical statistical models and data analysis methods that pre-date deep learning and are still widely used and/or are relevant for our fundamental understanding of learning, prediction and estimation with data.

We will start with an overview of the following topics and try to connect them with the state-of-the-art in ML/statistics research:

- **Statistical foundations of learning, learning models and algorithms**: Empirical risk minimization, regression models, classifiers, PAC learning, boosting, decision trees, clustering, support vector machines (SVMs), neural networks
- **Optimization methods and statistics** Convex optimization, Stochastic gradient descent and variants, generalization, kernel methods, model selection and cross-validation, bias-variance tradeoff
- **Additional estimation/inference/learning models** Multiclass ranking, compressed sensing, principal component analysis, generative modeling, graphical models

1 General information

- 3 credits, two lectures per week, 4 homeworks, 2 midterms and 1 final project.
- Class time and location: Tuesdays and Thursdays 12:30-1:45 pm, East Architecture 123.
- Office hours: 30 minutes after each lecture
- Instructor email: nishac@gatech.edu
- TAs name and email: **Darryl Jacob** (djacob30@gatech.edu), **Atharva Ketkar** (aketkar30@gatech.edu), **Chengrui Li** (cnlichengrui@gatech.edu), **Akpevwe Ojameruaye** (aojameruaye3@gatech.edu), **Yusen Su** (ysu349@gatech.edu), and **Mithilesh Vaidya** (mithilesh.vaidya@gatech.edu)

2 Prerequisites and tips for success

A strong background in linear algebra, probability and statistics as well as mathematical maturity are necessary to succeed in this course. You must be skilled at Python/Julia programming for ML/data science. Additionally, you must be motivated to gain a foundational understanding of data analysis methods and interested in their principled application. In this case, you can definitely fill in gaps in your math and computing background by working through textbook material in linear algebra and statistics and through programming assignments. Good starting points are the course textbook appendices (see section 4) and the books "Numerical Linear Algebra" by Trefethen and Bau and "Introduction to Probability" by Bertsekas and Tsitsiklis.

We emphasize that helping each other understand the concepts and enjoy the mathematical and computational aspects of learning (pun intended!) together is the main goal of this course. Spending time building a strong foundation by learning classical techniques will help in forming a good mental model of this vast field, and help us keep up with (and not be intimidated by) the proliferating research in the area of data science methods.

3 Learning outcomes

Students who attend all the lectures and complete the homeworks and the final project will

- get an overview of supervised learning models for regression and classification, and some unsupervised learning models and methods
- understand the mathematical and statistical foundations of learning and data mining
- gain experience implementing machine learning methods using Pytorch or other standard libraries
- use their foundational understanding to select, analyze and interpret results from machine learning and optimization methods in a principled manner taking into account application needs.

4 Resources (not exhaustive)

- The textbook for the course will be "Understanding machine learning: from theory to algorithms" by Shalev-Shwartz and Ben-David. Other books we will cover material from are "Foundations of machine learning" by Mohri, Rostamizadeh and Talwalkar and "Probabilistic machine learning" (parts I and II) by Murphy. These are available online. Some lectures will be based on research articles and other books, and these will be cited during class. As modern ML methods grow, so do the mathematical and computational questions around them. Hence, it is important to remember that this course is only a limited view of a vast landscape. Apart from similar courses offered across Georgia Tech (CS ML/ISyE 6740/7750), there are other freely available course materials that will certainly enhance this view. Here are a couple: MIT 6.867 and MIT 6.860.

- Main website will be Canvas. We will use Piazza for discussions. Any technical question you have, chances are others have it too, and still others know how to solve, and so post publicly on Piazza for everyone's benefit.

5 A note on the CSE qualifying exam

If you are taking the CSE "Computational Data Analysis" qualifying exam, please prepare all the topics listed in the CSE graduate handbook by taking additional courses or through self-study. Some of the listed topics (such as non-negative matrix factorization) will not be covered in this class, and some topics that will be covered are beyond the scope of the qualifying exam.

6 Tentative course schedule

Please note that the plan below is subject to change, both in terms of the content and order.

Part 1 - Learning: Foundations, models, algorithms

Week 1 Least-squares regression, Compressed sensing, LASSO, Logistic regression, perceptron algorithm, Empirical risk minimization

Week 2 Continuation of linear models, Halfspaces and linear programming, Gaussian mixtures

Week 3 Neural network models, generalization, computational complexity of learning, PAC learning

Week 4 Boosting algorithm, Support Vector machines, Decision trees, multi-class classifiers

Week 5 Kernel methods, kernel trick, basis expansions, PCA, ICA

Week 6 Clustering, k-means, spectral clustering, graphical models

Part 2: Statistics and optimization

Week 6 Kernel density estimation, Model selection and cross-validation, Bayesian inference

Week 7 Variational inference, parameter estimation methods, probabilistic classifiers

Week 8 Gradient and stochastic gradient descent variants, regularization, overfitting, generalization revisited

Week 9 Margin theory, Bias-complexity tradeoff, generalization bounds, algorithmic stability

Part 3: Other learning models and modalities

Week 10 Reinforcement learning, stochastic optimal control

Week 11 SDEs, MCMC revisited, Score-based generative models/diffusion models.

Week 12 Variational Autoencoders, Generative adversarial neural networks

Week 13 Learning dynamical systems, operator learning

Other possible topics, if time permits:

- Optimal transport, mean-field games
- Graph neural networks
- Transformers, LSTMs
- Recurrent neural networks

7 Grading information and late policy

The final grade will be determined by performance on:

- Final project: 40%
- Homework: 30%
- Midterm I and II: 30%

Final project: single most important contribution toward the grade. You can choose to do individual projects or in groups of 2. The final project submission includes a proposal (due mid November), code, accompanying report and a 5-minute presentation in the last week of class. A final project rubric and a set of guidelines will be posted on canvas before the proposal due date.

Homeworks: there will be 4 homework assignments (due dates TBD, but spread out evenly through the semester before the final project) that will include programming assignments and theoretical questions. You are welcome to discuss with other students and use online resources to solve the questions. After that, however, all the submitted work should be your own. Please submit typed up homework solutions (handwritten solutions are often illegible and will not be graded) on Canvas as a pdf.

Midterms: One in-class midterm will be held on October 19th. Another mid-term will be take-home and due on November 7th in class. You are allowed to use a single page cheat sheet of your own notes during your midterms and **no** other material. We require that you follow the honor code for both midterms: no copying or cheating will be tolerated (see 8).

Late policy – only applies to homeworks and final project proposal: there is a late penalty of 25% for a submission late by up to 24 hours, 50% for a submission delayed beyond 24 hours and up to 48 hours.

8 Honor code

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 this link or this one.

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.

Ultimately, learning and engaging with the material, and having fun in the process is most important! Assessments and homeworks are no more than good motivators to keep you accountable in the learning process. It serves no purpose to violate the honor code.

9 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 through their website, 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.

10 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. See the catalog for an articulation of some basic expectation that you can have of me and that I have of you. 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.