15.095 Machine Learning under a Modern Optimization Lens

Place-Time: E51-345, MW: 4:00-5:30PM

Instructor:

Kimberly Villalobos Carballo, Tel.: (617) 388-9911; Office hours: T 10:00-11:30am, E62-350; e-mail: kimvc@mit.edu

Teaching Assistants

Vassilina Stoumpou (head TA), email: vasstou@mit.edu, Office hours: W 2:30-4pm, E51-063

George Margaritis, email: geomar@mit.edu, Office hours: M 2:30-4pm, E51-063

Lisa Everest, email: leverest@mit.edu, Office hours: TR 2:30-4:00PM, E51-061

Recitation:

F E51-345, 11:00AM-12:00PM.

Course Page

The course page is available at https://web.mit.edu/canvas/

Piazza

Rather than emailing questions to the teaching staff, we will use Piazza for addressing all questions. It is also a great place for online discussion with classmates. Please sign up for class at https://piazza.com/mit/fall2024/15095_fa24. The access code is: `mlopt_15095_2024`.

Gradescope

We will use Gradescope for submissions and grading. Join the class using entry code: **VD42WB**.

Course Content and Objectives:

For over a century, Statistics has used probability models as a primitive assumption. In the data poor environment of the past, this was justifiable, but in the data rich world of today this is not only unnecessary, but it may lead to inaccurate predictions and suboptimal decisions. Machine learning (ML) has experienced tremendous growth in the twenty-first century that has influenced society in a variety of areas of human activity. We attribute this growth to its adaptation of a data driven perspective. The majority of the central ML problems, however, have been addressed using heuristic methods even though they can be formulated as formal optimization problems. Examples include Lasso for sparse regression, and classification and regression trees (CART) for classification and regression, among many others.

The purpose of the class is to provide a unified, insightful, and modern treatment of ML using three modern optimization lenses: convex, robust and mixed integer optimization. We de-emphasize the use of probability models, start with data and then revisit the central problems of ML using formal optimization methods and demonstrate that they can greatly benefit from a modern optimization treatment. We take a rigorous, non-heuristic, optimization-based approach to ML that leads to better out of sample performance compared to heuristic approaches. Specifically, throughout the class we demonstrate that using modern optimization we can find solutions to large scale instances of central ML problems that

(a) can be found in seconds/minutes,

(b) can be certified to be optimal in minutes/hours,

(c) outperform classical heuristic approaches in out of sample experiments involving real world and synthetic datasets.

Text: Dimitris Bertsimas and Jack Dunn, Machine Learning under a Modern Optimization Lens, Dynamic Ideas Press, 2019. You can find the book at the MIT Coop or order it here.

Recitations: The recitations will be an integral part of the class. They cover software implementation in Julia, computational aspects, and examples and applications that enhance the theory developed in the lectures.

Course Requirements: Problem sets, two midterms and a final team project. A project will need to involve up to two students per project. Grades will be determined by performance on the above requirements weighted as 20% problem sets, 25% Midterm I, 25% Midterm II, and 30% final team project.

Homework Due Dates: The due dates of the problems sets are as follows:

- Pset #1: 9/18 (covers Lec. 1-3).
- Pset #2: 9/30 (covers Lec. 4-6).
- Pset #3: 10/14 (covers Lec. 7-9).
- Pset #4: 11/04 (covers Lec. 10-13, includes project proposal).
- Pset #5: 11/18 (covers Lec. 14-17).
- Pset #6: 11/29 (covers Lec. 18-20).
- Pset #7: Final report 12/06.

Please submit your work on time. 10 points (out of 100) will be taken out for each half-day late, unless you obtain an extension before the due date or there are extenuating circumstances (e.g. needing to isolate due to COVID-19).

Project Information: The project is on a topic of the students' choice. It is expected that it involves two students. It is possible to do it alone or in teams of three, but you need to talk with the instructor. The requirements involve:

- Submit a project proposal, due on 11/04 (part of Pset #4).
- Present on 12/04.
- Submit an 8-page report due on 12/06.

Policy on Individual Work: Students can discuss problem sets with other students, but their answers must be their own. Students should list people they have talked to about each problem at the top of each problem set. Copying a solution in a problem set, in an exam, or in the project, violates the policy in individual work. Violations on this policy will result in lowering one's grade, taking an F in the class, among others.

A lecture by lecture preview

Lecture 1: Introduces the optimization lenses we use in this class: convex, robust and mixed integer optimization; describes the astonishing progress of mixed integer optimization and discusses what tractability means from a practical perspective.

Lecture 2: Develops robust linear regression under the lens of robust optimization, characterizes precisely its relationship with regularized regression and suggests that the remarkable success Lasso has experienced since the mid-1990s can be attributed to its robustness rather than its sparsity properties.

Lecture 3: Proposes both primal and dual methods to solve sparse linear regression under the lens of mixed integer and convex integer optimization, solves sparse linear regression in dimensions and samples in the 100,000s in seconds, observes new phase transition phenomena and argues that the dual sparse regression approach presents a superior alternative over heuristic methods available at present.

Lecture 4: Proposes a robust optimization framework for optimally selecting training and validation sets for regression problems and shows it leads to lower prediction error and lower standard deviation for both the prediction and the coefficients compared to the randomization approach.

Lecture 5: Contains extensions to nonlinear and median regression under the lenses of mixed integer and convex optimization.

Lecture 6: Outlines holistic regression, a framework based on mixed integer optimization, which develops a linear regression with a variety of desirable properties simultaneously, such as robustness, sparsity, significance, absence of multicollinearity, and others.

Lecture 7: Generalizes robustness and sparsity to classification problems with emphasis on logistic regression and support vector machines.

Lecture 8: Gives an overview of the classification and regression trees (CART) algorithm, random forests and gradient boosted decision trees, and outlines some of their limitations.

Lecture 9 Introduces optimal classification and regression trees (OCT) using parallel splits, provides solutions derived both using MIO and local improvement methods and presents results on accuracy in both synthetic and real world datasets. The Lecture includes two examples of the application of OCT in medicine: **(a)** redesigning the system of liver transplantation in the United States that promises to avert 400 deaths annually, **(b)** estimating the mortality and morbidity risk for emergency surgery patients. It further discusses optimal classification trees using hyperplane splits and emphasizes how the method compares with random forests and boosted trees using both real and synthetic datasets.

Lecture 10: Introduces Deep Learning models for supervised and unsupervised machine learning tasks. We discuss standard model architectures such as convolutional and recurrent neural networks, as well as more modern architectures such as transformers.

Lecture 11: Presents Multimodal Machine Learning, where we combine structured and unstructured data to make predictions. We cover applications in multiple areas such as Medicine and Metereology.

Lecture 12: Proposes a framework for extending predictive ML methods to prescriptive ones, and demonstrates that such methods provide an edge in decision-making directly from data. It also includes a demonstration of the power of prescriptive methods in a real world inventory management problem faced by the distribution arm of an international media conglomerate.

Lecture 13: Presents optimal prescription trees and optimal policy trees that are generalizations of optimal prediction trees that lead to optimal decisions.

Lecture 14: Provides theoretical and computational evidence that, in the context of design of experiments, groups created by optimization have exponentially lower discrepancy in pre-treatment covariates than those created by randomization or by existing matching methods.

Lecture 15: Identifies a subgroup in a clinical trial for which the average treatment effect is exceptionally strong or exceptionally weak and which can be defined by a small pre-specified number of covariates under the lens of mixed integer optimization.

Lecture 16: Presents a new way for clustering that is interpretable and provides insights on the nature of the clusters by utilizing the methodology from optimal classification trees.

Lecture 17: Presents Multi-Task Learning, where we create models by solving multiple Machine Learning tasks simultaneously.

Lecture 18: Develops an approach to sparse principal component analysis using mixed integer optimization and demonstrates that it has an edge over alternative heuristic methods.

Lecture 19: Provides a rigorous framework for factor analysis under the lenses of convex and mixed integer optimization that leads to provably optimal solutions in high dimensions.

Lecture 20: Develops algorithms for matrix completion with and without side information. It places particular attention to interpretability.

Lecture 21: Introduces algorithms for tensor completion with and without side information and leads to superior predictions for anti-cancer drug response.

Lecture 22: Explores the application of optimal classification trees and neural networks to predict the optimal solution in an optimization problem as parameters of the problem vary.

Lecture 23: Project presentations.

Lecture 24: Midterm Review.

Philosophy

Some of the key philosophical principles that characterize the class are:

(a) **Interpretability.** We believe that interpretability in ML matters. In an accident involving a driverless car that uses ML for its vision that leads to loss of life, we feel that society will not tolerate not knowing whether the algorithm made a mistake. Especially in critical applications involving decisions of significant magnitude, it has been our experience that decision-makers need to understand the logic of the algorithm. We have placed particular emphasis on the ideas of sparsity that lead to interpretable regression and classification models in Lectures 3 and 7 and to the development of optimal trees that are treated in Lecture 9 as well as in prescriptive and policy trees (Lecture 13), stable regression (Lecture 4), interpretable clustering (Lecture 16), interpretable matrix completion (Lecture 20) and interpretable optimization (Lecture 22).

(b) **The link between ML and optimization.** Historically, statistics has been linked to probability theory. One of our objectives is to reveal that the link of ML/statistics to optimization leads to significant advances in our ability to solve ML/statistics problems and to provide a fresh perspective that enhances our understanding of ML/statistics. Furthermore, in Lecture 25 we explore the reverse direction: using ML to give interpretability to optimization problems.

(c) **Robustness is more important than optimality.** In our experience a robust solution is preferable to a brittle optimal one. This is the reason we have placed significant emphasis on deriving robust solutions in Lectures 2, 3, 6, and 7.

(d) **Randomization versus optimization.** In Lectures 4, 14, and 15, we show that optimization has a performance edge over randomization in many ML problems. Furthermore, in Lecture 9, we empirically demonstrate that optimal classification and regression trees are as powerful in terms of performance compared to random forests.

(e) **Practability.** In contrast to complexity theory, we judge methods based on their ability to solve problems in times and for sizes that are appropriate for the application that motivated the problem. In our view, polynomial solvability or $\mathcal{NP}$-hardness of a problem does not give relevant information for our ability to solve the problem in the real world. Given the motivation of this class to solve real world problems, we use the notion of practical tractability alongside theoretical tractability when evaluating algorithms.

(f) **Prescriptive methods.** The majority of ML has focused on prediction. It is our belief that the ultimate objective should be the ability to make high-quality decisions. We present prescriptive methods in Lectures 12–14.

(g) **Deep Learning and Multimodal, Multitask ML.** We introduce deep learning and discuss multimodal ML for making predictions that involve both structured and unstructured data in Lectures 10,11,17.

Course Syllabus

Lecture	Time	Topic	Chapter in Book
1	W, 9/04	Optimization Lenses and ML	Ch. 1
2	M, 9/9	Robust Linear Regression	Ch. 2
3	W, 9/11	Sparse Linear Regression	Ch. 3
4	M, 9/16	Stable Regression	Ch. 17
5	W, 9/18	Median and Convex Regression	Ch. 4
6	M, 9/23	Holistic Regression	Ch. 5
7	W, 9/25	Robust Classification	Ch. 6
8	M, 9/30	CART, Random Forest and Boosted Trees	Ch. 7
9	W, 10/02	Optimal Classification and Regression Trees	Ch. 8-11
	M, 10/07	Midterm I	
10	W, 10/09	Deep Learning Models	Handout
11	W, 10/16	Multimodal Machine Learning	Handout
12	W, 10/23	From Predictions to Prescriptions	Ch. 13
13	M, 10/28	Optimal Prescriptive and Policy Trees	Ch. 14
14	W, 10/30	Optimal Design of Experiments	Ch. 15
15	M, 11/04	Identifying Exceptional Responders	Ch. 16
16	W, 11/06	Interpretable Clustering	Ch. 20
17	W, 11/13	Multi-Task Learning	Handout
18	M, 11/18	Sparse Principal Component Analysis	Ch. 21
19	W, 11/20	Factor Analysis	Ch. 22
20	M, 11/25	Matrix Completion	Ch. 24
21	W, 11/27	Tensor Learning	Ch. 25
22	M, 12/02	Interpretable Optimization	Ch. 26
23	W, 12/04	Project presentations	
24	M, 12/09	Review	
	W, 12/11	Midterm II	