15.095 Machine Learning under a Modern Optimization Lens

Place-Time: E51-315, MW: 4:00-5:30

Instructor:

Dimitris Bertsimas, E62-560; Tel.: (617) 253-4223; Office hours: by appointment; e-mail: dbertsim@mit.edu; homepage: http://web.mit.edu/dbertsim/www/

Teaching Assistants:

Michael Li, email: mlli@mit.edu, Office hours: Mondays: 2:30pm-3-30pm, E51-242.
Ryan Cory-Wright, email: ryancw@mit.edu, Office hours: Tuesdays: 4:30pm-5:30pm, E51-242.

Recitation: Fridays: E51-335, 10:30-11:30.

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. It will be available around September 20.

Recitations: The recitations will cover software implementation in Julia, computational aspects, and examples and applications that enhance the theory developed in the lectures.

Course Requirements: Problem sets, a midterm exam and one 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 approximately as 30% problem sets, 30% midterm and 40% final team project.

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

- Pset #1: 9/18.
- Pset #2: 10/02.
- Pset #3: 10/28.
- Pset #4 (includes project proposal): 11/4.
- Pset #5: 11/18.
- Pset #6: 12/4.

Project Information The project is on a topic of the student 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 11/4 (part of Pset #4).

- Submit a poster on 12/4.
- Submit an 8 Page report due on 12/9.

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 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

Note: This course is offered concurrently with *15.071 The Analytics Edge*, taught by Prof. Alexandre Jacquillat. We have strived to align the syllabi, so that the methods covered in this course will be showcased in some of the case studies covered in 15.071. If you are taking both courses, you will first learn methodological/theoretical foundations in this course, and then work on applications in 15.071. Whenever that is the case, it is highlighted in blue in the schedule below. Still, both courses are independent from each other and can be taken separately.

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. This lecture is complemented by some case study materials covered in Lecture 4 of 15.071 entitled "Diabetes Prediction".

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

Lecture 5: Generalizes robustness and sparsity to classification problems with emphasis on logistic regression and support vector machines. This lecture is complemented by some case study materials covered in Lectures 6 and 7 of 15.071 entitled "Loan Defaults" and "Customer Retention", respectively.

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: Gives an overview of the classification and regression trees (CART) algorithm, random forests and gradient boosted decision trees, and outlines some of their limitations. This lecture is complemented by some case study materials covered in Lecture 8 and 10 of 15.071 entitled "Medical Costs" and "Click-Thru Rates in Advertising", respectively.

Lecture 8: Introduces optimal classification 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. This lecture is complemented by some case study materials covered in Lecture 9 of 15.071 entitled "Parole Decisions".

Lecture 9 Contains optimal regression trees with constant predictions, where the prediction in each leaf of the tree is the average of all the values of the dependent variable among all data points that are included in the leaf of the tree, and compares how this approach improves upon the CART methodology using real and synthetic data. It further deals with optimal regression trees with linear predictions, where the prediction in each leaf of the tree comes from a linear regression involving all the points that are included in the leaf, and presents evidence that they lead to significantly improved accuracy.

Lecture 10: Proves that a variety of neural networks (feedforward, convolutional and recurrent)

can be transformed to classification and regression trees with hyperplanes with the same accuracy in the training set, showing that such trees are at least as powerful as neural networks in modeling power. This lecture is complemented by some case study materials covered in Lecture 12 of 15.071 entitled "Radiology".

Lecture 11: 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. This lecture is complemented by some case study materials covered in Lecture 19 of 15.071 entitled "TBD".

Lecture 12: Presents optimal prescription trees that are generalizations of optimal prediction trees that lead to optimal decisions. This lecture is complemented by some case study materials covered in Lecture 19 of 15.071 entitled "TBD".

Lecture 13: The midterm examination covering the material from Lectures 1-12.

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: 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. This lecture will revisit the traditional training/test approach covered in Lectures 2–12 of 15.071.

Lecture 17: Takes a different perspective on the bootstrap, one of the most significant ideas of modern statistics. The bootstrap uses randomization, but in this Lecture we use exact counting of integer points in polyhedra to propose a method that has an edge on accuracy compared to randomization. This lecture will revisit the traditional bootstrap approach covered in Lecture 5 of 15.071.

Lecture 18: Poses the missing data problem under a general optimization framework and develops opt.impute, an algorithm for missing data imputation that significantly outperforms other heuristic

approaches. This lecture is complemented by some case study materials covered in Lecture 21 of 15.071 entitled "Customer Segmentation".

Lecture 19: It also 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. This lecture is complemented by some case study materials covered in Lecture 21 of 15.071 entitled "Customer Segmentation".

Lecture 20: Develops an approach to sparse principal component analysis using mixed integer optimization and demonstrates that it has an edge over alternative heuristic methods.

Lecture 21 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 22: Extends the framework for sparse regression developed from earlier Lectures to sparse inverse covariance estimation and demonstrates its edge over heuristic approaches.

Lecture 23: Develops algorithms for matrix completion with and without side information. It places particular attention to interpretability. This lecture is complemented by some case study materials covered in Lecture 24 of 15.071 entitled "Online movie recommendations".

Lecture 24: Introduces algorithms for tensor completion with and without side information and leads to superior predictions for anti-cancer drug response.

Lecture 25: 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 26: Project presentations. Note that the duration of this lecture is is 4-7:30pm. Teams will present posters.

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 6 and to the development of optimal trees that are treated in Lectures 8, 9 as well as in prescriptive trees (Lecture 12), stable regression (Lecture 16), interpretable clustering (Lecture 18), interpretable matrix completion (Lecture 22) and interpretable optimization (Lecture 24).

(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 24 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, 5, 6, 16 and 21.

(d) Randomization versus optimization. In Lectures 14-17, we show that optimization has a performance edge over randomization in many ML problems. Furthermore, in Lectures 8, 9 we demonstrate empirically 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 that the motivation of this class is 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 11, 12.

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	Median and Convex Regression	Ch. 4
5	W, 9/18	Robust and Sparse Classification	Ch. 6
6	M, 9/23	Holistic Regression	Ch. 5
7	W, 9/25	CART, Random Forest and Boosted Trees	Ch. 7
8	M, 9/30	Optimal Classification Trees	Ch. 8, 9
9	W, 10/02	Optimal Regression Trees	Ch. 10, 11
10	M, 10/07	Deep Learning and Optimal Trees	Ch. 12
11	W, 10/09	From Predictions to Prescriptions	Ch. 13
12	W, 10/16	Optimal Prescriptive Trees	Ch. 14
13	M, 10/21	Midterm	
14	W, 10/23	Optimal Design of Experiments	Ch. 15
15	M, 10/28	Identifying Exceptional Responders	Ch. 16
16	W, 10/30	Stable Regression	Ch. 17
17	M, 11/4	Exact Bootstrap	Ch. 18
18	W, 11/6	Missing Data Imputations	Ch. 19
19	W, 11/13	Interpretable Clustering	Ch. 20
20	M, 11/18	Sparse Principal Component Analysis	Ch. 21
21	W, 11/20	Factor Analysis	Ch. 22
22	M, 11/25	Sparse Inverse Covariance Estimation	Ch. 23
23	M, 12/2	Matrix Completion	Ch. 24
24	W, 12/4	Tensor Learning	Ch. 25
25	M, 12/9	Interpretable Optimization	Ch. 26
	M, 12/9	Poster presentations (5:30pm-7:00pm)	
26	W, 12/11	Poster presentations, (4:00-7:00pm)	