15.095: Machine Learning under a Modern Optimization Lens

Lecture 1: Optimization Lenses in Machine Learning

Outline

Administration

- **Time**: Monday/Wednesday, 4pm-5:30pm
- **Place**: E51-315
- **Instructors**: Dimitris Bertsimas, E40-111 (dbertsim@mit.edu, `http://web.mit.edu/dbertsim/www/`)
  Martin Copenhaver, E40-148 (mcopen@mit.edu, `https://www.mit.edu/~mcopen`)
- **Office hours**: by appointment
- **TAs:** Colin Pawlowski: cpawlows@mit.edu, Office hours: 3-4pm Monday
  Yuchen Wang, email: yuchenw@mit.edu, Office hours: 3-4pm Monday
- **Recitation**: Friday 10:30am-11:30am, E51-335

Administration

- **Text:** Research papers and preliminary chapters from [5]; access on Canvas
- **Recitations:** `julia` and `JuMP`, computational aspects, examples, and applications.
- **Course Requirements:** 30% problem sets, 30% midterm exam, and 40% final team project.
- **Background required:** Knowledge of a class in optimization (15.081/6.251 or 15.093/6.255)

Why this class?

- Central problems in Machine Learning (ML) have been addressed using heuristic methods.
- This implies that we do not really know if we have indeed solved these problems.
- In the last two decades convex optimization (CO) methods have had increasing importance: Compressed Sensing, Matrix Completion among many others.
- Mixed integer optimization (MIO) and Robust Optimization (RO) are widely unknown in ML.
- People in ML believe that MIO problems are intractable.
- Yet MIO, RO, CO have advanced very significantly.

Objectives

- teach you the ORC brand of ML.
- take a rigorous, non-heuristic approach to ML that leads to better out of sample performance compared to heuristic approaches.
- To demonstrate that using modern optimization optimal solutions to large scale instances in ML/S
    - can be found in seconds
    - can be certified to be optimal in minutes
    - outperform classical heuristic approaches in out of sample experiments involving real and synthetic data.
- To enable you to do it using Jump and Julia.
- To link Optimization to ML/S.

MIO

$$
\begin{aligned}
\text{(MIO)} \quad \max \quad & \mathbf{c}'\mathbf{x} + \mathbf{h}'\mathbf{y} \\
\text{s.t.} \quad & \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} \leq \mathbf{b} \\
& \mathbf{x} \in Z_+^n (\mathbf{x} \geq 0, \mathbf{x} \text{ integer}) \\
& \mathbf{y} \in R_+^n (\mathbf{y} \geq 0)
\end{aligned}
$$

$$
\begin{aligned}
\text{(QMIO)} \quad \max \quad & \mathbf{x}'\mathbf{Q}\mathbf{x} + \mathbf{c}'\mathbf{x} + \mathbf{h}'\mathbf{y} \\
\text{s.t.} \quad & \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} \leq \mathbf{b} \\
& \mathbf{x} \in Z_+^n (\mathbf{x} \geq 0, \mathbf{x} \text{ integer}) \\
& \mathbf{y} \in R_+^n (\mathbf{y} \geq 0)
\end{aligned}
$$

Progress of MIO

- Speed up between CPLEX 1.2 (1991) and CPLEX 11 (2007): 29,000 times
- Gurobi 1.0 (2009) comparable to CPLEX 11
- Speed up between Gurobi 1.0 and Gurobi 6.5 (2015): 48.7 times
- Total speedup 1991-2015: 1,400,000 times
- A MIO that would have taken 16 days to solve 25 years ago can now be solved on the same 25-year-old computer in less than one second.
- Hardware speed: 93.0 PFlop/s in 2016 vs 59.7 GFlop/s in 1993 1,600,000 times
- Total Speedup: **2.2 Trillion times!**
- A MIO that would have taken 71,000 years to solve 25 years ago can now be solved in a modern computer in less than one second.

RO in Regression

- Given data $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, $(\mathbf{y} = (y_1, \ldots, y_n)$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$,

$$\text{Regression} \quad \min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\beta||^2$$

- Given errors in the data, $\mathbf{X} + \mathbf{\Delta X}$, $\mathbf{\Delta X} \in U = \{\mathbf{\Delta X} : ||\mathbf{\Delta X}|| \leq \lambda\}$.
- $$\text{Robust Regression:} \quad \min_{\boldsymbol{\beta}} \max_{\mathbf{\Delta X} \in U} ||\mathbf{y} - (\mathbf{X} + \mathbf{\Delta X})\beta||^2$$
- Progress in RO: The time to solve the RO problem is of the same order of magnitude as the nominal problem.

Convex Optimization

- Given convex functions $f(\mathbf{x})$, $g_j(\mathbf{x})$, $j = 1, \dots, m$.

- $$\begin{array}{lll} \text{(CO)} & \min & f(\mathbf{x}) \\ & \text{s.t.} & g_j(\mathbf{x}) \leq 0, \ j = 1, \dots, m. \end{array}$$

- Progress in CO: The use of first order methods enables fast running times in high dimensions.

Remarks on Complexity

- A key requirement of a theory is to be positively correlated with empirical evidence.
- Consider the Simplex method and solving the TSP.
- A 2.2 Trillion speed up forces us to reconsider what is tractable.
- A problem is practable if it can be solved for sizes and in times that are appropriate for the application.
- Online trading problems need to be solved in milliseconds.
- Regression problems used for planning need to be solved in minutes or in hours.
- Asymptotic polynomial solvability or NP-hardness is not relevant under this definition.

Lectures

#	Date	Topic	Readings
1	W, 9/05	Optimization Lenses and Machine Learning	
2	M, 9/10	Best Subset Selection in Linear Regression	[13, 26]
3	W, 9/12	Robust Linear Regression and Classification	[2, 7, 12, 23]
4	M, 9/17	Algorithmic Framework for Linear Regression	[11, 17]
5	W, 9/19	Optimal Classification and Regression Trees	[4, 5]
6	M, 9/24	Median and Convex Regression	[18, 21]
7	W, 9/26	Missing Data Imputations	[24]
8	M, 10/1	Interpretable Clustering	[22]
9	W, 10/3	Boosting	[29]
10	W, 10/10	Deep Learning	[30]
11	M, 10/15	Optimal Trees and Deep Learning	[19]
12	W, 10/17	Optimal Prescriptive Trees	[6]
13	M, 10/22	From Predictions to Prescriptions I	[9]

Lectures

#	Date	Topic	Readings
14	W, 10/24	From Predictions to Prescriptions II	[10, 20]
15	M, 10/29	Power of Optimization over Randomization	[8, 14]
16	W, 10/31	Identifying Exceptional Responders	[15]
17	M, 11/5	*Midterm*	
18	W, 11/7	Bootstrap methods	[25]
19	W, 11/14	Sparse Principal Component Analysis	[1]
20	M, 11/19	Low Rank Factor Analysis	[3]
21	W, 11/28	Sparse Inverse Covariance Estimation	[16]
22	M, 12/3	Matrix Completion	[28]
23	W, 12/5	Learning with Tensors	[27]
24	M, 12/10	*Project Presentations*	
25	W, 12/ 12	*Project Presentations*	

Regression Topics

- Best Subset Selection: $\min_{\boldsymbol{\beta}} \quad \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 \quad \text{subject to} \quad ||\boldsymbol{\beta}||_0 \leq k$
- Robust Regression: $\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta}\mathbf{X} \in U} \; ||\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta}\mathbf{X})\boldsymbol{\beta}||^2$
- Develop an algorithm based on MIO to accomodate Sparsity, Limiting multicollinearity, Categorical variables, Group sparsity, Nonlinear transformations, Robustness, Statistical significance
- Median Regression: $\min_{\boldsymbol{\beta}} \; \underset{i=1,\ldots,n}{\text{median}} \; |y_i - \mathbf{x}_i^T \boldsymbol{\beta}|$
- Convex Regression: $\min_{\boldsymbol{\beta}} \quad \min_{f:convex} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2.$

Classification Topics

- Logistic Regression: $\max_{\beta,\beta_0} \; -\sum_{i=1}^n \log\left(1 + e^{-y_i(\beta^T \mathbf{x}_i + \beta_0)}\right).$

- SVM: $\min_{\mathbf{w},b} \; \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{n} \max\{1 - y_i(\mathbf{w}^T\mathbf{x}_i - b), 0\}.$

- Optimal Trees: Partition the space with hyperplanes to minimize misclassification error.

Optimization in Design of Experiments and in ML

- Given factors $\mathbf{x}_i$, split them in k groups to minimize discrepancy. How does this approach compare to randomization, which is the gold standard in clinical trial.

- ML/S primarily focuses to make predictions. We will develop theory to extend the ML/S methods to make decisions.

Matrix Problems

- Sparse PCA: max $\mathbf{x}'\mathbf{\Sigma}\mathbf{x}, \ s.t. \ ||\mathbf{x}|| = 1, \ ||\mathbf{x}||_0 \leq k.$
- Factor Analysis

$$\begin{aligned} \min \quad & ||\mathbf{\Sigma} - (\mathbf{\Theta} + \mathbf{\Phi})|| \\ \text{subject to} \quad & \text{rank}(\mathbf{\Theta}) \leq r \\ & \mathbf{\Theta} \succeq \mathbf{0} \\ & \mathbf{\Phi} = \mathrm{diag}(\Phi_1, \ldots, \Phi_p) \succeq \mathbf{0} \\ & \mathbf{\Sigma} - \mathbf{\Phi} \succeq \mathbf{0}. \end{aligned}$$

- Estimation of Inverse Covariance Matrix:

$$\min_{\mathbf{\Theta} \succ \mathbf{0}} \quad \langle \overline{\mathbf{\Sigma}}, \mathbf{\Theta} \rangle - \log\det \mathbf{\Theta} \quad \text{s.t.} \quad ||\mathbf{\Theta}||_0 \leq k.$$

- Matrix Completion

$$\min_{\mathbf{\Theta}} \quad \sum_{(i,j)\in\Omega} (x_{ij} - \theta_{ij})^2 \quad \text{s.t.} \ \ \text{rank}(\mathbf{\Theta}) \leq r.$$

L. Berk and D. Bertsimas.

Certifiably optimal sparse Principal Component Analysis.

Mathematical Programming Computation, under review, 2017.

D. Bertsimas and M. Copenhaver.

Characterization of the equivalence of robustification and regularization in linear and matrix regression.

European Journal of Operational Research, 270:931–942, 2018.

D. Bertsimas, M. Copenhaver, and R. Mazumder.

Certifiably optimal low rank factor analysis.

Journal of Machine Learning Research, 18:1–53, 2017.

D. Bertsimas and J. Dunn.

Optimal trees.

Machine Learning, 106(7):1039–1082, 2017.

D. Bertsimas and J. Dunn.

Machine Learning under a Modern Optimization Lens.

Dynamic Ideas, 2018.

D. Bertsimas, J. Dunn, and N. Mundru.
Optimal prescriptive trees.
INFORMS Journal of Optimization, to appear, 2018.

D. Bertsimas, J. Dunn, C. Pawlowski, and Y. Zhuo.
Robust classification.
INFORMS Journal of Optimization, to appear, 2018.

D. Bertsimas, M. Johnson, and N. Kallus.
The power of optimization over randomization in designing experiments involving small samples.
Operations Research, 63 (4):868–876, 2015.

D. Bertsimas and N. Kallus.
From predictions to prescriptions.
Management Science, under review, 2015.

D. Bertsimas and N. Kallus.

Pricing from observational data.
Management Science, under review, 2017.

D. Bertsimas and A. King.
An algorithmic approach to linear regression.
Operations Research, 64(1):2–16, 2016.

D. Bertsimas and A. King.
Logistic regression: From art to science.
Statistical Science, 32(3):367–384, 2017.

D. Bertsimas, A. King, and R. Mazumder.
Best subset selection via a modern optimization lens.
Annals of Statistics, 44(2):813–852, 2016.

D. Bertsimas, N. Korolko, and A. Weinstein.
Covariate-adaptive optimization in online clinical trials.
Operations Research, under review, 2017.

D. Bertsimas, N. Korolko, and A. Weinstein.
Identifying exceptional responders in randomized trials: An optimization approach.
INFORMS Journal on Optimization, under review, 2018.

D. Bertsimas, J. Lamperski, and J. Pauphilet.
Certifiably optimal sparse inverse covariance estimation.
Mathematical Programming, under review, 2016.

D. Bertsimas and M. Li.
Accounting for significance and multicollinearity in building linear regression models.
INFORMS Journal on Optimization, under review, 2018.

D. Bertsimas and R. Mazumder.
Least quantile regression via modern optimization.
Annals of Statistics, 42 (6):2494–2525, 2014.

D. Bertsimas, R. Mazumder, and M. Sobiesk.

On the equivalence of neural networks and optimal trees.
working paper, 2018.

D. Bertsimas and C. McCord.
From predictions to prescriptions in multistage optimization problems.
Mathematical Programming, under review, 2017.

D. Bertsimas and N. Mundru.
Sparse convex regression.
INFORMS Journal on Computing, under review, 2017.

D. Bertsimas, A. Orfanoudaki, and H. Wiberg.
Interpretable clustering: An optimal trees approach.
Working paper, 2018.

D. Bertsimas, J. Pauphilet, and B. van Parys.
Sparse classification and phase transitions: a discrete optimization perspective.
Journal of Machine Learning Research, under review, 2017.

D. Bertsimas, C. Pawlowski, and Y. Zhuo.

From predictive methods to missing data imputation: An optimization approach.

Journal of Machine Learning Research, under review, 2017.

D. Bertsimas and B. Sturt.

Computation of exact bootstrap confidence intervals.

Operations Research, under review, 2017.

D. Bertsimas and B. van Parys.

Sparse high dimensional regression: Exact scalable algorithms and phase transitions.

Annals of Statistics, under review, 2016.

V. Farias and A. Li.

Learning preferences with side information.

Under review, 2017.

R. Freund, P. Grigas, and R. Mazumder.

An extended Frank-Wolfe method with "in-face" directions, and its application to low-rank matrix completion.

working paper, 2015.

Y. Freund, R. Schapire, and N. Abe.

A short introduction to boosting.

Japanese Society For Artificial Intelligence, 14(771-780):1612, 1999.

I. Goodfellow, Y. Bengio, and A. Courville.

Deep Learning.

MIT Press, 2016.

http://www.deeplearningbook.org.

Applications

1. Matrix completion
2. Treating patients
3. Factor Analysis
4. Operations management for hospitals

Million-dollar matrices

Can you predict how a person will rate a movie given a collection of his/her ratings of various movies as well as ratings of other users?

The **power** of data

Prescribing treatments

Given a patient's medical history and demographic information, how do you decide the best treatment for him?

Fundamental issue: observational nature of historical data.

Analytics does not occur in a vacuum!

Factor Analysis

Find a parsimonious representation of the covariance structure of a set of variables using a small number of *hidden factors*

Classical example:

Given measurements on a set of test questions, can you explain performance using a small set of latent factors?

Modern examples:

Identify underlying factors driving returns among a set of assets?

Understand the cross elasticities of a collection of products using sales data?

Psychometrics

Given 2,800 participants' responses on a set of 25 personality questions, can you explain performance using a small set of factors?

- "Love children"
- "Continue until everything is perfect"
- "Waste my time"

At the heart of Factor Analysis is the goal of distinguishing between variance that is *common* across all variables versus variance due to individual components.

Video surveillance

Raw frames

Static background

Healthcare operations

One of the central challenges in hospitals is *bed management* and *capacity planning*.

One of the most fundamental questions you can ask: *which patients do we expect will be discharged today?*

Primary questions:

- At 5am, predict which patients will go home today.
- Who are the patients most likely to home?
- What are the *barriers* to a patient not being discharged?
- How do you intervene for those patients?
- What does it mean to actually implement an ML model? (MLIRL)

References

1. Various references listed in syllabus
2. Netflix Prize (see e.g. Wikipedia and linked sources)
3. Discharge prediction at Mass General Hospital, work in progress by Safavi et al.