15.095 Machine Learning via a Modern Optimization Lens
Fall 2018

Time and place: E51-315, Monday/Wednesday, 4:00pm-5:30pm

Instructors:

Dimitris Bertsimas, E40-111; Tel.: (617) 253-4223; Office hours: by appointment;
e-mail: dbertsim@mit.edu; homepage: http://web.mit.edu/dbertsim/www/

Martin Copenhaver, E40-148; Office hours: by appointment; e-mail: mcopen@mit.edu;
homepage: http://www.mit.edu/~mcopen/

Teaching Assistants:

Colin Pawlowski, email: cpawlows@mit.edu, Office hours: 3-4pm Monday, E51-242

Yuchen Wang, email: yuchenw@mit.edu, Office hours: 3-4pm Monday, E51-242

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

Course Content and Objectives: The majority of the central problems of regression, classification, and estimation have been addressed using heuristic methods even though they can be formulated as formal optimization problems. While continuous optimization approaches have had a significant impact in Machine Learning (ML)/Statistics (S), mixed integer optimization (MIO) has played a very limited role, primarily based on the belief that MIO models are computationally intractable. The last three decades have witnessed (a) algorithmic advances in MIO, which coupled with hardware improvements have resulted in an astonishing over 2 trillion factor speedup in solving MIO problems, (b) significant advances in our ability to model and solve very high dimensional robust and convex optimization models.

Our objective in this course is to revisit some of the classical problems in ML/S and demonstrate that they can greatly benefit from a modern optimization treatment. The optimization lenses we use in this course include convex, robust, and mixed integer optimization. In all cases we demonstrate that optimal solutions to large scale instances (a) can be found in seconds, (b) can be certified to be optimal/near-optimal in minutes and (c) outperform classical heuristic approaches in out of sample experiments involving real and synthetic data.

The problems we address in this course include:

- variable selection in linear and logistic regression,
- convex, robust, and median regression,
- an algorithmic framework to construct linear and logistic regression models that satisfy properties of sparsity, robustness, significance, absence of multi-collinearity in an optimal way,
- clustering,
- deep learning,
- how to transform predictive algorithms to prescriptive algorithms,
- optimal prescriptive trees,
- the design of experiments via optimization ,
- missing data imputation,
- bootstrap methods, and
- matrix estimation including principal component analysis, factor analysis, inverse covariance matrix estimation, matrix completion, and tensor methods.

Text: Research papers and preliminary chapters from [5]. All handouts can be downloaded from the course website: `https://stellar.mit.edu/S/course/15/fa18/15.095/`

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, an in-class midterm, 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 exam, and 40% final team project.

Lecture Schedule

Lecture	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]
	M, 10/8	*No class (Columbus Day holiday)*	
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]
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]
	M, 11/12	*No class (Veterans Day holiday)*	
19	W, 11/14	Sparse Principal Component Analysis	[1]
20	M, 11/19	Low Rank Factor Analysis	[3]
	W, 11/21	*No class*	
	M, 11/26	*No class (ORC conference)*	
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	

References

[1] L. Berk and D. Bertsimas. Certifiably optimal sparse Principal Component Analysis. *Mathematical Programming Computation*, under review, 2017.

[2] 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.

[3] D. Bertsimas, M. Copenhaver, and R. Mazumder. Certifiably optimal low rank factor analysis. *Journal of Machine Learning Research*, 18:1–53, 2017.

[4] D. Bertsimas and J. Dunn. Optimal trees. *Machine Learning*, 106(7):1039–1082, 2017.

[5] D. Bertsimas and J. Dunn. *Machine Learning under a Modern Optimization Lens*. Dynamic Ideas, 2018.

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

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

[8] 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.

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

[10] D. Bertsimas and N. Kallus. Pricing from observational data. *Management Science*, under review, 2017.

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

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

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

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

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

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

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

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

[19] D. Bertsimas, R. Mazumder, and M. Sobiesk. On the equivalence of neural networks and optimal trees. working paper, 2018.

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

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

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

[23] 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.

[24] 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.

[25] D. Bertsimas and B. Sturt. Computation of exact bootstrap confidence intervals. *Operations Research*, under review, 2017.

[26] D. Bertsimas and B. van Parys. Sparse high dimensional regression: Exact scalable algorithms and phase transitions. *Annals of Statistics*, under review, 2016.

[27] V. Farias and A. Li. Learning preferences with side information. Under review, 2017.

[28] 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.

[29] Y. Freund, R. Schapire, and N. Abe. A short introduction to boosting. *Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[30] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.