LECTURE #1

LECTURER: DIMITRIS BERTSIMAS   SCRIBE: KIRAN VODRAHALLI

Contents

1 Introduction

Professor Bertsimas from MIT in the Operations Research Center. His interests include optimization. He has made great contributions, many of his students are also leading figures.

2 Outline

I have three optimization lenses:

1. Mixed-Integer Optimization (very little used)

2. Robust Optimization (little used)

3. Convex Optimization (heavily used)

I believe the intersection of these methods can get us great steps forward.

1. Best Subset Selection:
$$\min_{\beta}(1/2)\|y - X\beta\|_2^2$$

   subject to $\|\beta\|_0 \leq k$. This is a problem of sparse regression. Furnival and Wilson (1974) solve it via implicit enumeration, which cannot scale beyond $p = 30$. Lasso was proposed which scales to very large problems via convex quadratic optimization:

$$\min_{\beta}(1/2)\|y - X\beta\|_2^2 + \lambda \sum_i |\beta_i|$$

   Under regularity conditions on X, Lasso leads to sparse models and good predictive performance. Lasso proposed in Tibshirani (1996) has $14,255$ citations, so it's fair to say this is one of the most cited papers in **science**, not just statistics!

   From my view, Lasso is a robustification method, not sparsity inducing. There is an equivalence between robustification and regularization:

$$\min_{\beta} \max_{\Delta \in U_{q,p}} \|y - (X + \Delta)\beta\|_p = \min_{\beta} \|y - X\beta\|_p + \lambda\|\beta\|_q$$

   With $p = 2, q = 1$, you get Lasso. I claim that Lasso is not particularly sparse, but it is robust (from a mathematical viewpoint).

There is a natural MIO formulation:

$$\min_{\beta,z}(1/2)\|y - X\beta\|_2^2$$

subject to $|\beta_i| \le M \cdot z_i$, $i \in [p]$, $\sum_i z_i \le k$, $z_i \in \{0,1\}$. We can solve this.

Typically what happens is the correct answer is found extremely quickly, and then the rest of the time takes a long time to prove optimality. Here, we propose to replace the proving part with MIO.

You can use a convex first order method to first of all find feasible solutions with no claim of optimality.

For the case $n > p$, MIO and warm-starts find provably optimal solutions for $n = 1000s$, $p = 100s$ in minutes. For $n < p$, MIO and warm-starts finds solutions with better prediction accuracy than Lasso.

This is a guaranteed method to prove optimality for the given optimization problem.

2. Least Median of Squares Regression

3. An algorithmic approach to linear regression

Here is what we tell them in our regression class. We put all the possible variables we think on the data we have, we run it, some of the coefficients are not significant, we take them out, we iterate, maybe we want sparsity (add Lasso), etc. It is not a scientific approach. It is not like you give me data, and here goes the model: there is a lot of human input instead.

If something is nonlinear, we are only saying what it is not. The art of building regression models in current practice:

  (a) Transform variables.

  (b) Pairwise scatterplots, correlation matrix.

  (c) Delete redundant variables.

  (d) Fit model, delete variables with insignificant t-tests, examine residuals.

  (e) See if additional variables can be dropped, new variables brought in.

So this is very nonscientific, you'll get a bunch of things.

So our aspiration is to go from art to science: We propose an algorithm automated process to build regression models. The approach is to express all desirable properties as MIO constraints.

2.1 Motivation

Continuous optimization methods have historically played a significant role in statistics. In the last two decades convex optimization methods have had increased importance in compressed sensing, matrix completion and others. Most problems in statistics and machine learning can naturally be expressed as **mixed integer optimization (MIO)** problems. MIO in statistics are considered impractical and corresponding problems intractable, and thus heuristic approaches are used instead: For instance, Lasso for best subset regression or CART for optimal classification.

   However, MIO has progressed quite a bit. A MIO that would have taken 7 years to solve 20 years ago can now be solved on the same 20-year-old computer in less than one second. Adding in hardware speedup gives a total speedup of 450 billion times from 20 years ago. In the early nineties I was looking at integer programming methods for control. It took a day to solve a certain problem in 1994, today it can be solved in a second provably optimally.

   Given the dramatically increased power of MIO, is MIO able to solve key multivariate statistics problems considered intractable a decade ago? How do MIO solutions compete with state-of-the-art? Can we algorithmize the process of building regression models? There is no well-defined algorithmic process for regression

currently, yet this is one of the most often used techniques. We want an algorithm to iterate through models rather than ad-hoc do it - we are not there yet. But given that humans have decided what specific model to use, it is feasible to do the best you can.

Most statistics algorithms that I know only give feasible solutions with no guarantees on optimality. The main thing that happens is that you get the optimal solution really really quickly, and after you prove optimality. This is the tradeoff.