Best subset selection

Rafał Muszyński, Ryszard Szymański

1/27/2020

Agenda

1. Problem introduction
2. Solution
3. Implementation
4. Benchmarks
5. Summary

Introduction

Problem: feature selection in linear regression

$$
\begin{aligned}
y &= X\beta + \epsilon \\
\min_{\beta} &||y - X\beta||_2^2
\end{aligned}
\tag{1}
$$

Well known solution : Lasso: Tibshirani (1996)

$$
\min_{\beta} \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1 \tag{2}
$$

Testing findings of Bertsimas, King, and Mazumder (2016)

Motivation

- LASSO shortcomings e.g. larger coefficients are more penalized then smaller coefficients
- Advances in both hardware and optimization frameworks such as CPLEX and GUROBI
- Best subset selection is an NP-hard problem

Solution

Restricted formulation of a MIQP problem.

$$
\begin{aligned}
& \min \alpha^T Q \alpha + \alpha^T a \\
\text{s.t.} \quad & A\alpha \leq b \\
& \alpha_i \in \{0, 1\}, i \in \mathcal{I} \\
& \alpha_i \in I\!R, i \notin \mathcal{I}
\end{aligned}
\tag{3}
$$

Solution

Best subset selection problem formulation

$$\begin{aligned} & \min_{\beta} ||y - X\beta||_2^2 \\ \text{s.t.} \quad & ||\beta||_0 \leq k \\ & ||\beta||_0 = \sum_{i=1}^{p} 1(\beta_i \neq 0) \end{aligned} \tag{4}$$

Solution

Final problem formulation

$$
\begin{aligned}
& \min_{\beta, z} \frac{1}{2}\beta^T(X^T X)\beta - \langle X'y, \beta\rangle + \frac{1}{2}\|y\|_2^2 \\
\text{s.t.} \quad & (\beta_i, 1 - z_i) : \text{SOS-1}, \quad i = 1, \ldots, p \\
& z_i \in \{0, 1\}, \quad i = 1, \ldots, p \\
& \sum_{i=1}^{p} z_i \leq k \\
& -\mathcal{M}_U \leq \beta_i \leq \mathcal{M}_U, \quad i = 1, \ldots, p
\end{aligned} \tag{5}
$$

Solution

First order method

Data: function: $g(\beta)$, parameter: L, convergence tolerance: ϵ, parameter: k

Result: β approximation

β_1 random initialization, $\beta_1 \in I\!R, \|\beta\|_0 < k$

do

$\quad \beta_{m+1} \in H_k(\beta_m - \frac{1}{L}\nabla g(\beta_m))$

while $g(\beta_m) - g(\beta_{m+1}) \leq \epsilon$;

Algorithm 1: Discrete first-order method

Implementation

- Three different starting methods:
  - Cold
  - Mild
  - Warm
- Two different solvers used:
  - CPLEX
  - GUROBI
- Parallelization for different k values

Benchmarks

- Both synthetic and real life datasets
- Syntethic datasets generated according to the procedure described in Bertsimas, King, and Mazumder (2016)
- Performed benchmarks analyzed
  - Predictive performance
  - Speed
  - Gap values for Warm/Mild/Cold start approaches

Benchmarks - datasets

- Diabetes dataset
- Synthetic datasets
  - $x_i \sim N(0, \Sigma)$, each standardized to have unit $l2$ norm
  - $y = X\beta^0 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$
  - The choice of X, β^0, ϵ determines the Signal-To-Noise Ratio: $SNR = \frac{var(X\beta^0)}{\sigma^2}$

Benchmarks - predictive performance

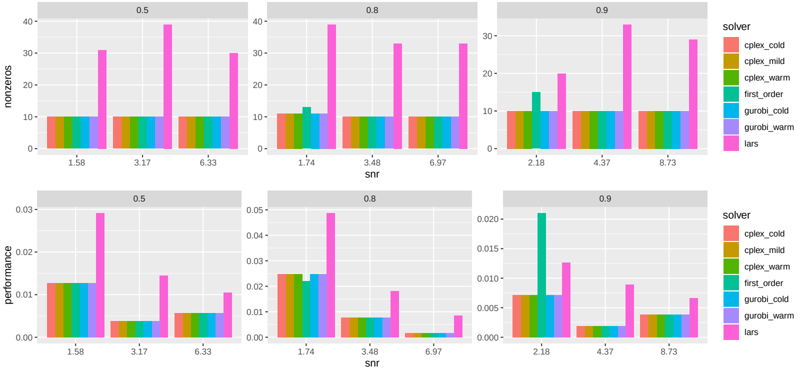

Figure 1: Predictive performance of researched methods.

Benchmarks - optimality gap for warm/cold starts

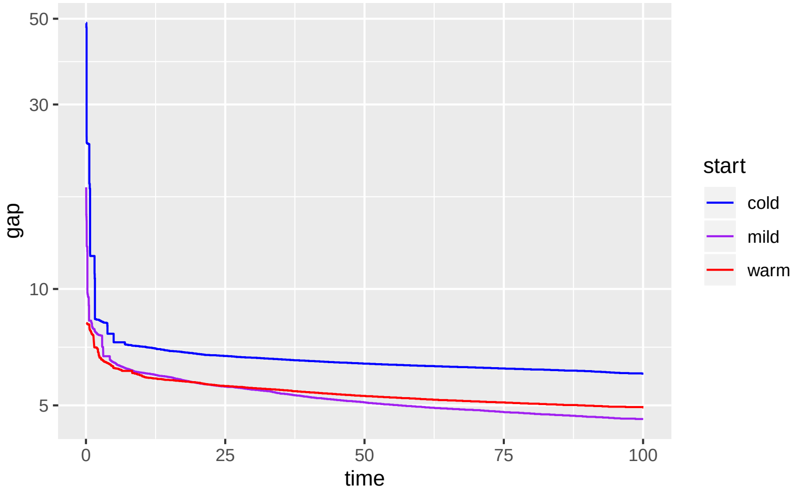

Figure 2: Optimality gap for warm, cold and mild start.

Benchmarks - speed performance

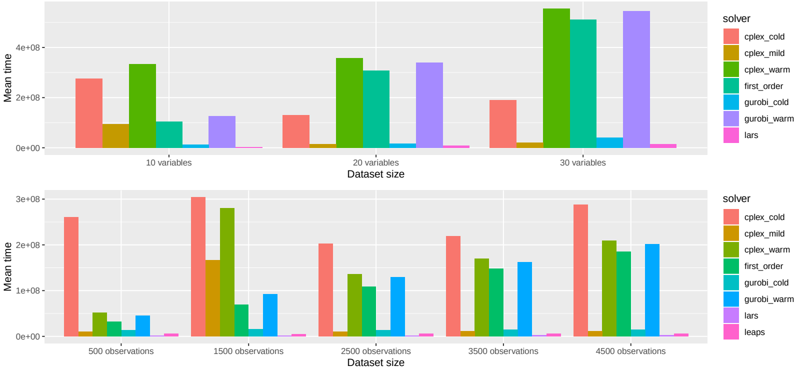

Figure 3: Speed of researched methods for datasets with a fixed number of variables (40) or observations (5000)

Summary

- The MIO approach outperforms LARS in terms of predictive results
- The proposed mild approach allows for obtaining high quality results in times similar to LARS
- The best subset approach conversely to LASSO in a single run does not aim to maximally reduce the amoun of nonzero coefficients

Bibliography

Bertsimas, Dimitris, Angela King, and Rahul Mazumder. 2016. "Best Subset Selection via a Modern Optimization Lens." *The Annals of Statistics* 44 (2): 813–52. https://doi.org/10.1214/15-AOS1388.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.