Georgia Tech College of Engineering
H. Milton Stewart School of Industrial and Systems Engineering

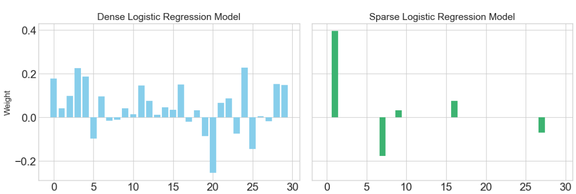
A Branch and Bound Algorithm for Sparse Logistic Regression

James Jones | Advisors: Dr. Juba Ziani and Dr. Weijun Xie

Motivation

Logistic Regression is a statistical method that aims to determine the probability of an event occurring based on a set of one or more variables.

Applications span fields such as medicine, finance, engineering, data science, machine learning, and AI.



Sparsity helps provide clarity on the underlying nature of the relationships.

Goal: provide a stable and efficient framework for sparse logistic regression using an ℓ_0 constraint

Formulation

Given feature vectors $X \in \mathbb{R}^{m \times n}$ and responses $y \in \{0,1\}^m$, select coefficient vector $\theta \in \mathbb{R}^n$ with objective:

$$\min_{\theta} \sum_{i=1}^{n} \left[-y_i \ln\left(\frac{1}{1+e^{-X_i^\top \theta}}\right) - (1-y_i)\ln\left(\frac{1}{1+e^{-X_i^\top \theta}}\right)\right]$$

such that $\|\theta\|_0 \leq k$.

The ℓ_0 constraint makes this a **nonconvex combinatorial problem** scaling as $\binom{n}{k}$.

Branch and Bound Algorithm

At each node on the tree, f^+ = the set of included variables and f^- = the set of excluded variables. A node is feasible when $|f^+| \leq k$ or $|f^-| \geq n-k$. Each node is evaluated by a relaxed objective where it is given all features that are not in f^-.

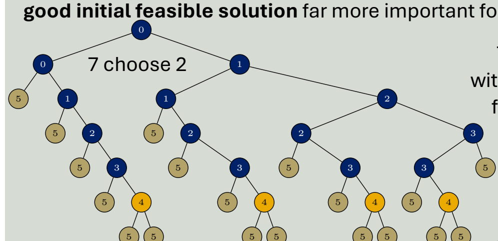
1) Maintain a set of feasible nodes F, a set of unexplored infeasible nodes U, an upper bound $UB = min(F)$, and a lower bound $LB = min(U)$.
2) Initialize a root node with $f^+ = f^- = \emptyset$, add it to U
3) Optionally use a heuristic to find an initial feasible solution, add it to F
4) while $UB > LB$:
    1) Remove the node with the min relaxed objective from U
    2) Select a variable $j \notin f^+ \cup f^-$ for this node
    3) Create 2 new nodes, one with $f_L^+ = f^+ \cup j$, the other with $f_R^- = f^- \cup j$
    4) Evaluate each of these new nodes, placing them in U or F
    5) Prune any nodes with greater relaxed objective than UB from U

On Left and Right

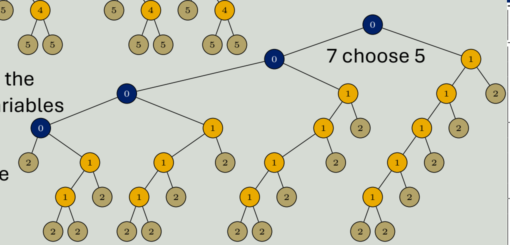
"Left" child nodes are defined as the node with the additional variable included, "right" child nodes are defined as the node with the additional variable excluded

To maintain monotonicity, relaxed objective is fully defined by the variables excluded, independent of the variables included

For $n \gg k$ the tree becomes right dominant, and the number of internal nodes with much better bounds is large, making **early performance** and a **good initial feasible solution** far more important for pruning

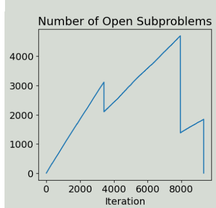


There are only 4 nodes with 1 less excluded than feasible **(good pruning opportunities)**, and they are far down the tree



Node values represent the number of excluded variables

14 nodes have 1 less excluded than feasible **(good pruning opportunities)**



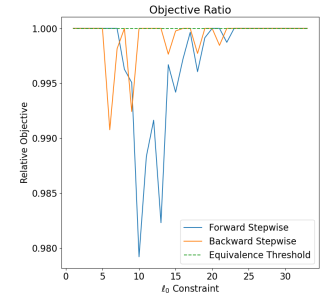
Pruning makes the largest advancements towards completion

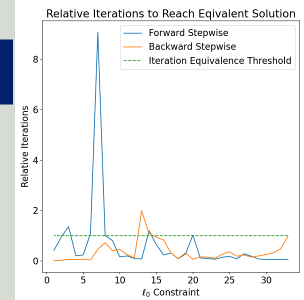
Comparison with Existing Methods

The most comparable existing heuristic is Stepwise Regression, where variables are greedily added (forward) or removed (backward). Using a fractional variable selector on smaller datasets, we found

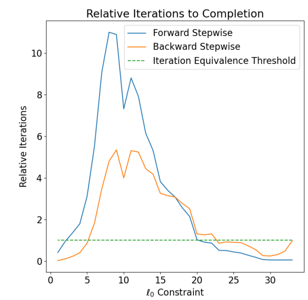
- **better** runtime to equivalence
- reasonable completion time
- **guaranteed improvement** on final objective.

Comparison of final objective value ratios between B&B solver and heuristic solutions. Ratios below 1.0 indicate that the solver found better solutions than the heuristic.





Number of iterations required by the B&B solver to reach an objective value equivalent to that of the heuristic solution, normalized by the heuristic's iteration count.



Number of iterations required by the B&B solver to complete, normalized by the heuristic's iteration count.

Test Results

Dataset	Instances (m)	Features (n)	l_0 Constraint (k)	Iterations	Bound Gap Percent	Time (seconds)	Conclusive
SPECT Heart	267	22	2	81	0.00%	5.73	True
			6	336	0.00%	45.35	True
			10	127	0.00%	13.26	True
			15	28	0.01%	2.66	True
			19	21	0.00%	1.76	True
Ionosphere	351	34	2	63	0.00%	2.28	True
			6	1042	0.01%	36.14	True
			10	2158	0.01%	80.93	True
			14	2048	0.00%	82.34	True
			18	1175	0.00%	49.82	True
Myocardial Infarction Complications	1700	111	2	4417	0.01%	1591.74	True
			6	24781	10.52%	16790.21	False
			10	24645	8.15%	14175.19	False
			14	23668	6.55%	13535.43	False
			18	24181	5.60%	13946.05	False

Test datasets are from the UCI Machine Learning Repository

Acknowledgements

I would like to extend a huge thank you to Dr. Xie and Dr. Ziani for their mentorship. I would also like to thank Dr. Xie, Yongchi Li, and Quill Healy for providing the initial framework for B&B.