Optimization and Machine Learning

Lecture 1: IP models to learn rules I

Sanjeeb Dash
IBM Research
(co-lecturer Parikshit Ram)

JPOC 13 Summer School, June 26-28, 2023
University of Clermont-Ferrand

IP models to learn rules I: Classification

- ► ML models for Binary Classification
  - Boolean (decision) rules
- ► Interpretable Machine Learning
- ► Integer Programming Formulation
  - Column Generation Technique
- ► Results
  - Winning entry in FICO Challenge
- ► Cardinality constrained Multilinear set
  - Polyhedral results
- ► Variants/applications of basic model
  - Fairness/Model diagnostics

Goal of lecture 1

- Present MIP models for classification problems

Let $x \in \mathbb{R}^n$ and $0 \leq k \leq n$

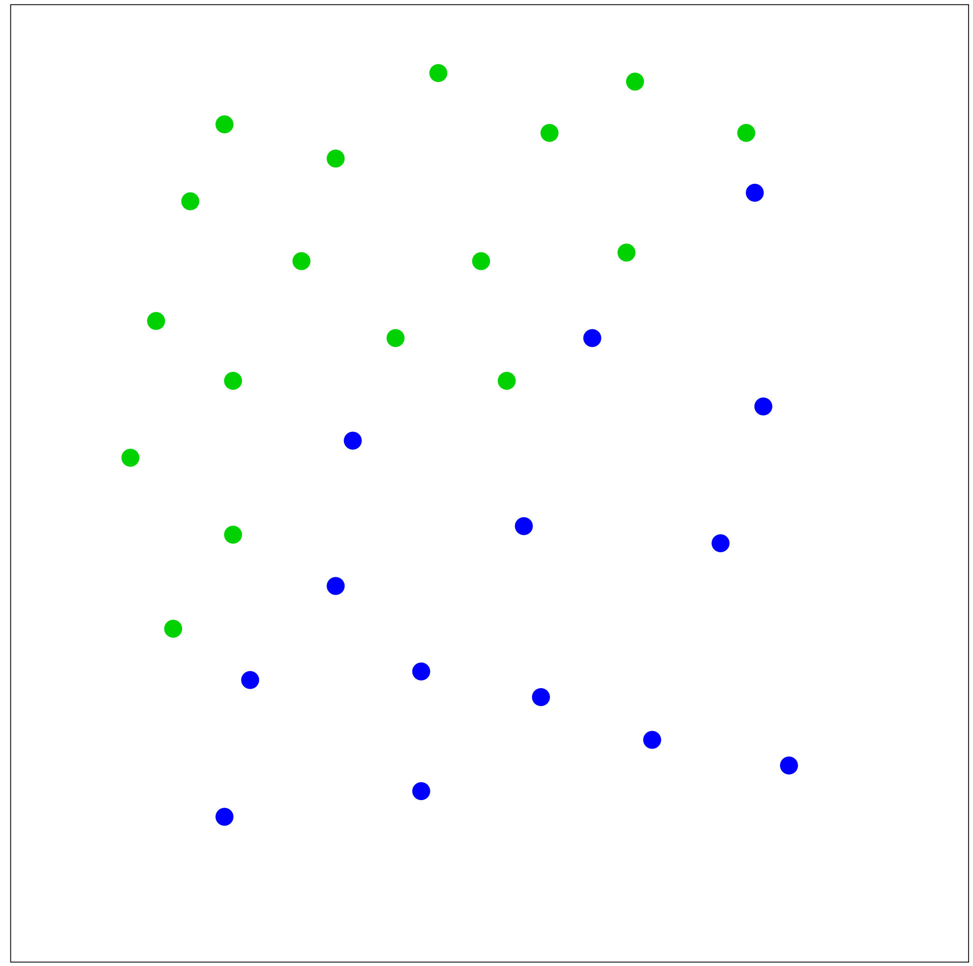
$$\text{MIP} \equiv \min \; c \cdot x \quad Ax \leq b \quad x_1, \ldots, x_k \in \mathbb{Z}$$

Supervised Binary Classification

- Features: $X_1, \ldots, X_m$
- Data: $\{(x_i; y_i) : i \in 1, 2, \ldots, n\}$ where $x_i \in \mathbb{R}^m$.
- Label $y_i \in \{0, 1\}$
- Feature X_j is either numeric or categorical.
- Goal: Separate 0s from 1s or find function f such that $y_i \approx f(x_i)$.

Supervised Binary Classification

Blood Pressure	Choles- -terol	Heart Disease
X_1	X_2	Y
100	75	0
120	175	1
80	250	1
110	150	0
90	190	1
⋮	⋮	⋮

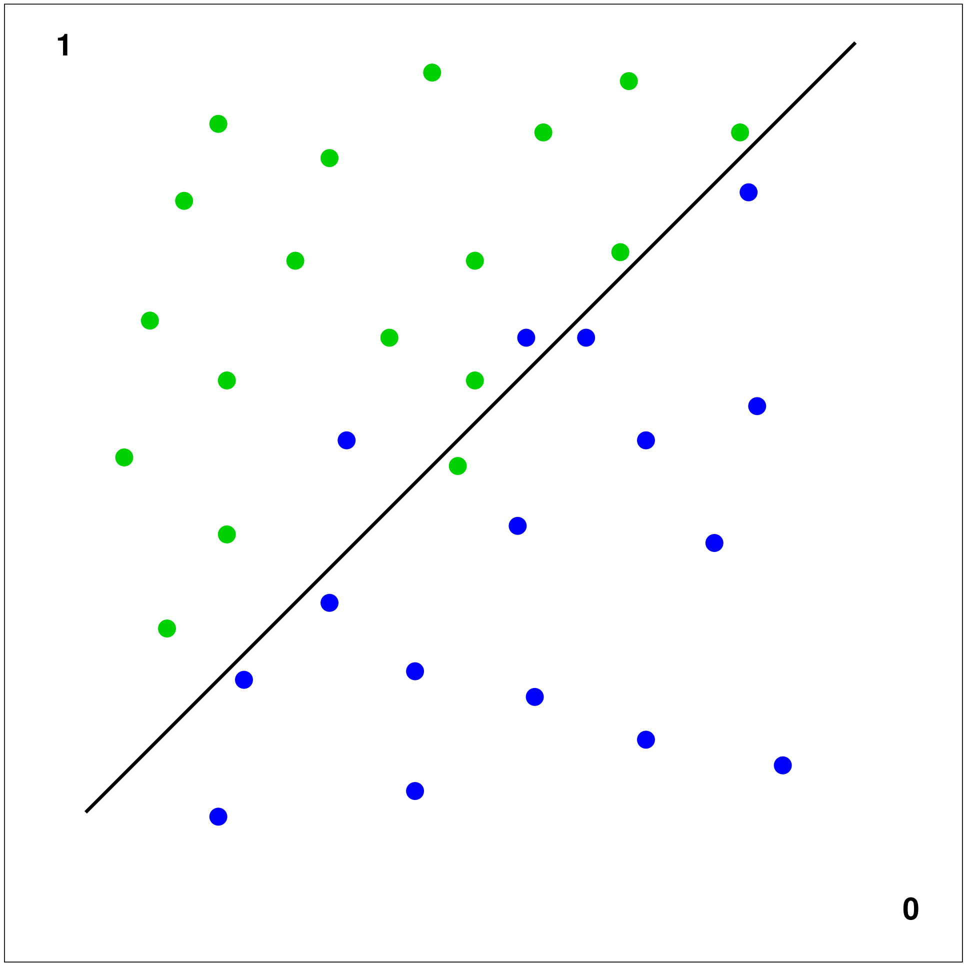


- ▷ Linear support vector machines
- ▷ Decision Trees
- ▷ Neural networks

Linear support vector machines

Find hyperplane that separates points labelled 1 from points labelled 0

Few nonzero coefficients $\rightarrow$ more interpretable



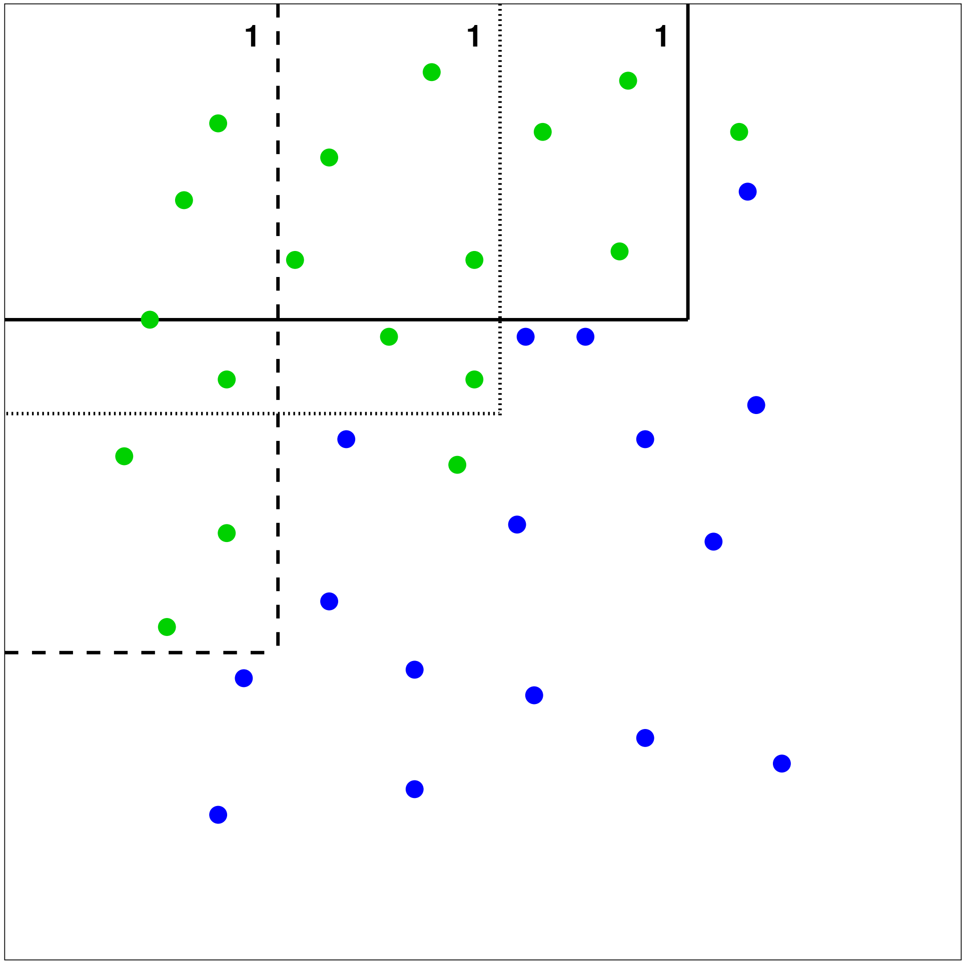
▷ Vapnik, Chervonenkis '63 - SVM
▷ Boser, Guyon, Vapnik '92 - Kernel "trick"

Learn boolean rule sets for binary classification

$(X_1 \leq 150$ AND $X_2 \geq 170)$ OR
$(X_1 \leq 100$ AND $X_2 \geq 130)$ OR
$(X_1 \leq 80$ AND $X_2 \geq 70)$

Boolean rule set $\equiv$ Boolean formulae in Disjunctive normal form

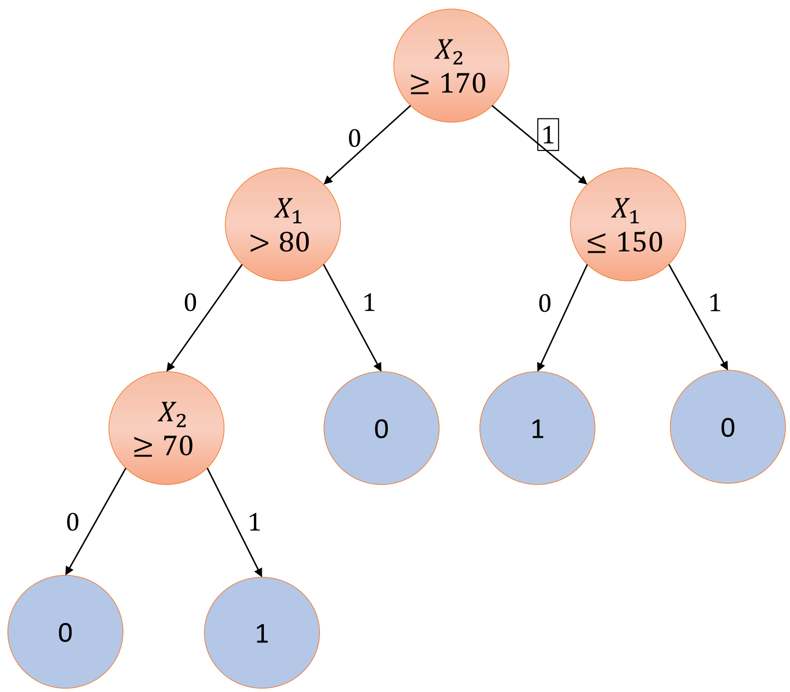
A data point is classified as 1 if it satisfies at least one rule

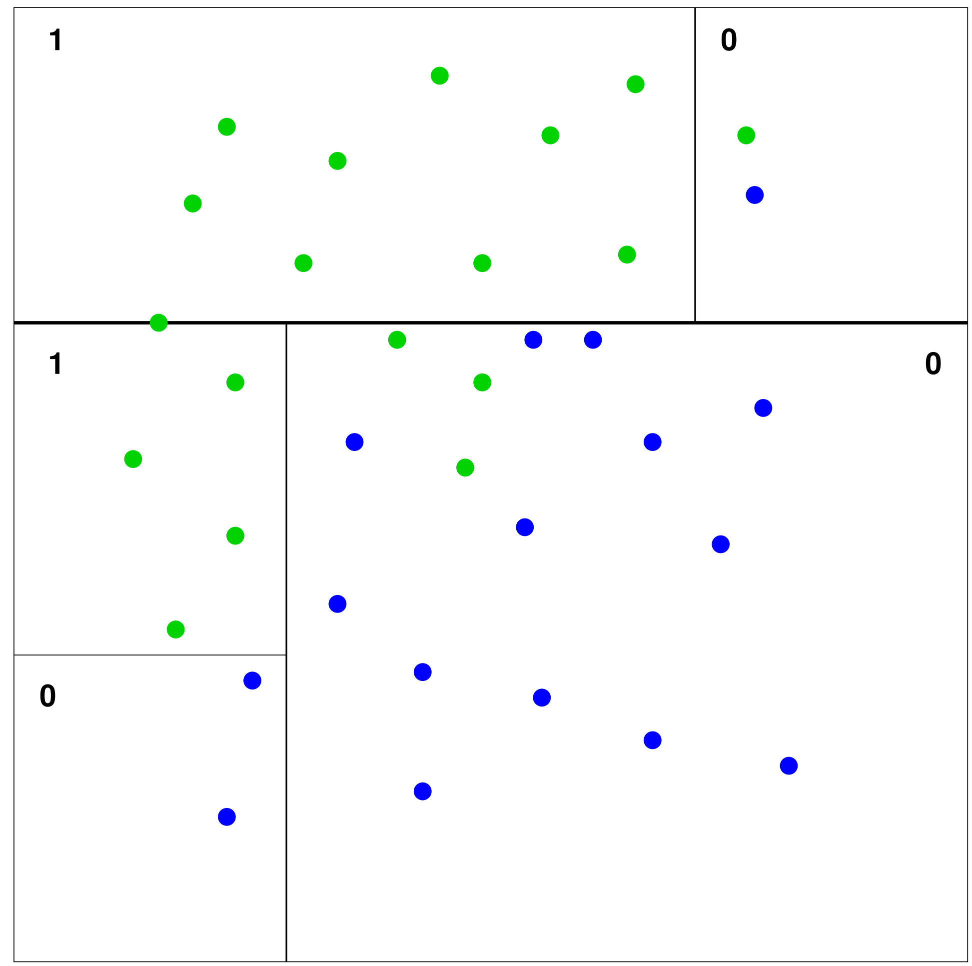


- ▷ Dawes ’79
- ▷ Cohen ’95 - RIPPER
- ▷ Hongyu, Rudin, Seltzer ’17 - Scalable Bayesian Rule Sets
- ▷ Boros, Hammer, Ibaraki, Kogan, Mayoraz, Muchnik ’00 - L.A.D.

Decision trees

Recursively partition space by axis-parallel hyperplanes





▷ Hunt, Marin, Stone ’66

▷ Quinlan ’86 - ID3, C4.5

▷ Breiman, Friedman, Olshen, Stone ’84 - CART

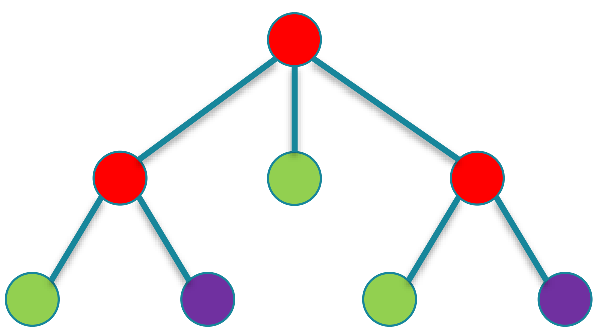
▷ Bertsimas, Dunn ’17 - IP formulations for decision trees

Related classifiers

DNF Boolean rule = Decision rule set

IF A THEN Y=1
IF B AND C THEN Y=1
IF D AND E THEN Y=1
ELSE Y=0

Decision tree



Decision list

IF A THEN Y=1
ELSE IF B AND C THEN Y=1
ELSE IF D AND E THEN Y=1
ELSE Y=0

Rivest ’87: Learning decision lists

Transforming one classifier to another one can lead to exponential blowup in “size”/”complexity”.

Interpretable Machine Learning

Explainable AI (XAI), Interpretable AI/ML, or Transparent AI refer to models that can be easily understood by humans unlike “black box” models where it is hard to explain why the AI took a specific decision.

Interpretable models	Noninterpretable models
Sparse linear models (e.g., SVM)	Dense linear models
Decision trees	Neural Networks
Decision rule sets	Random Forests
Decision lists	

▷ Interpretable models can be examined for:

Safety/Reliability, Fairness/Lack of Bias, Causality, Robustness

Doshi-Velez, Kim ’17: “Towards a rigorous science of interpretable ML”
Schmidt et. al. ’17, Muggleton et. al. ’18: Higher inspection time $\rightarrow$ lower interpretability

FICO Search... SEARCH LOGIN

USER FORUMS GROUPS HELP TRIALS BLOGS SIGN UP

Submissions will be accepted from April 18- August 31, 2018.

OVERVIEW DATASET DETAILS CHALLENGE RULES EXAMPLE EXPLANATIONS CHALLENGE FORUM ENTER YOUR SUBMISSION FAQ

Introduction

Complex machine learning models have recently achieved great predictive successes for many applications. While these models excel at capturing complex, non-linear relationships between variables, it is often the case that neither the trained model nor its individual predictions are readily explainable. In settings where regulators or consumers demand explanations, understanding the structure and predictions of these models will pave the way for their wide adoption in practice. Explainability will also help data scientists understand their datasets and the models' predictions, uncover and correct for biases, and ultimately create better models.

Motivation: Why the financial services industry?

Advanced machine learning methods are quickly finding applications throughout the financial services industry, transforming the handling of large and complex datasets, but there is a huge gap between our ability to construct effective predictive models and our ability to understand and control these models. In order to drive forward research in this area, FICO and a number of academic partners have collaborated to design a challenge based on a real financial dataset. The challenge is not necessarily focused on accuracy, rather, it is focused on evaluating the explanations generated by the participants.

Every year, credit scoring methodologies provide millions of scores that evaluate the risk in billions of dollars in loans; in fact, the FICO Score is used in more

Prior work on Boolean rule sets

► Using heuristics and/or multiple criteria
- Covering e.g. RIPPER (Cohen ’95)
- Bottom-up combining
- Associative classification

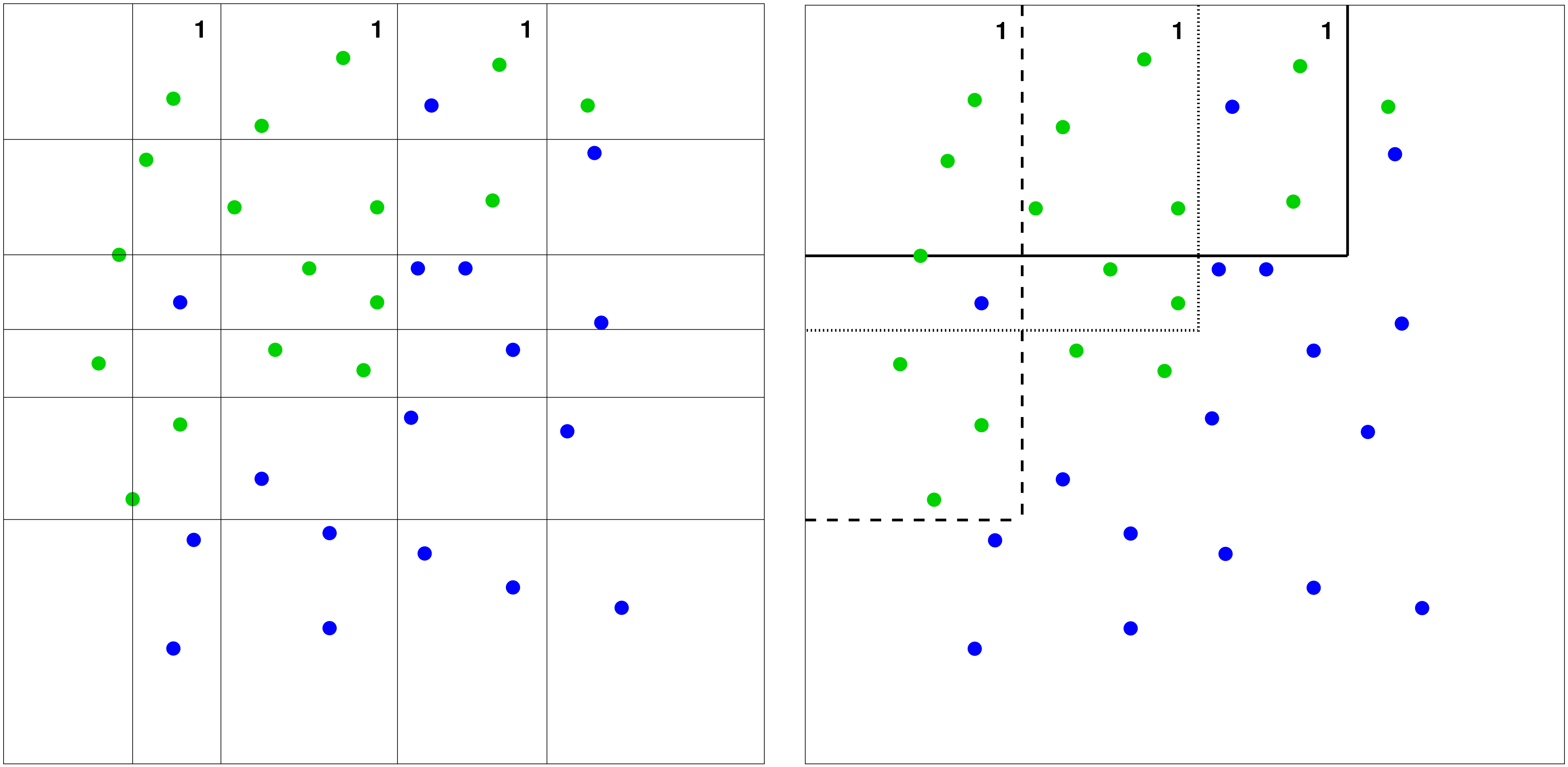
Interpretable Boolean rule sets: Few rules/few conditions per rule

► Accuracy-simplicity optimization
- Interpretable decision sets (IDS): Lakkaraju, Bach, Leskovec ’16
- Bayesian rule sets (BRS): Wang, Rudin, Doshi-Velez, Liu, Klampfl, MacNeille ’17
- Optimized ORs of ANDs: Wang Rudin ’15
- Disjunctions of conjunctions: Hauser et. al. ’10

These methods use rule mining to generate candidate clauses

- IP formulation with fixed # clauses, solved approximately using LP: Su, Wei, Varshney, Malioutov ’16

Learn boolean rule sets for binary classification



1) Choose rectangle boundaries from fixed gridlines.
2) Penalty for misclassication - 1 or # of times misclassified?

Binarization

X_1	X_2	Y
100	75	0
120	175	1
80	250	1
110	150	0
90	190	1
⋮	⋮	⋮

$\rightarrow$

$X_1 \le 80$	$X_1 \le 100$	$X_2 \le 150$	$X_2 \le 200$	Y
0	1	1	1	0
0	0	0	1	1
1	1	0	0	1
1	1	1	1	0
0	1	0	1	1
⋮	⋮	⋮	⋮	⋮

Goal: Learn boolean functions (assume X_i is binary) in DNF form as classifiers

$$(X_1 \wedge X_3 \wedge X_4) \vee (X_2 \wedge X_5)$$

Notation

Inputs to model

$x_i \in \mathbb{R}^m, y_i \in \{0,1\}$ - data point i and it's label

C - upper bound on complexity of chosen rule set

Model information

$\mathcal{K}$ - set of all possible rules

$a_{ik} \in \{0,1\}$ - $a_{ik} = 1$ iff data point i satisfies rule k.

c_k - 'complexity' of rule k = 1 + number of conditions in rule

Variables

$w_k \in \{0,1\}$ - binary variable which is 1 iff rule k is selected

$\xi_i \geq 0$ - variable which is 1 iff chosen rules do not 'cover' data point i

- All features are assumed to be binary at this point
- $a_{ik} = x_{ik}$ if k is index of a rule containing a single binary feature
- $\sum_{i \in k} a_{ik} w_k \geq 1$ iff $\vee_{k:w_k=1}$rule$_k(x_i) = 1$
- Here we assume rule$_k$ is a function from $\mathbb{R}^m \to \{0,1\}$.

IP to select "best" subset of rules

Minimize 0-1 loss subject to complexity bound:

loss on positive instances loss on negative instances

$$\min_{w,\xi} \sum_{i:y_i=1} \xi_i + \sum_{i:y_i=0} \xi_i$$

cover positives ⟶ $\xi_i + \sum_{k \in K} a_{ik} w_k \geq 1, \quad \xi_i \geq 0, \qquad i: y_i = 1$

cover negatives ⟶ $\xi_i \geq a_{ik} w_k, \quad k \in K, \quad \xi_i \geq 0, \qquad i: y_i = 0$

complexity bound ⟶ $\sum_{k \in K} c_k w_k \leq C$

select clause k or not $w_k \in \{0,1\}, \qquad k \in K$

MIP has exponentially many inequalities/variables and is hard to solve

"Master" IP with Hamming Loss objective

Minimize Hamming loss subject to complexity bound:

loss on positive instances | loss on negative instances

$$\min_{w,\xi} \sum_{i:y_i=1} \xi_i + \sum_{i:y_i=0} \sum_{k \in K} a_{ik} w_k$$

cover positives $\longrightarrow$ $$\xi_i + \sum_{k \in K} a_{ik} w_k \geq 1, \quad \xi_i \geq 0, \qquad i: y_i = 1$$

complexity bound $\longrightarrow$ $$\sum_{k \in K} c_k w_k \leq C$$

select clause k or not $$w_k \in \{0,1\}, \qquad k \in K$$

Dash, Günlük, Wei (NIPS 2018): Search over exponential list of clauses using column generation.

Gap between 0-1 loss and Hamming Loss

Thm (Lawless, Dash, Günlük, Wei ’22) The 0-1 loss of the Hamming IP solution can be ‘arbitrarily’ worse than that of the of 0-1 IP solution.

(No constant ratio between the two losses)

m

$m/2$

$$\begin{array}{cccccccc|cc}
0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
\dots & & & & & & & & & \\
0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\
\hline
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1
\end{array}$$

$m-2$

Positive labels

n

negative labels

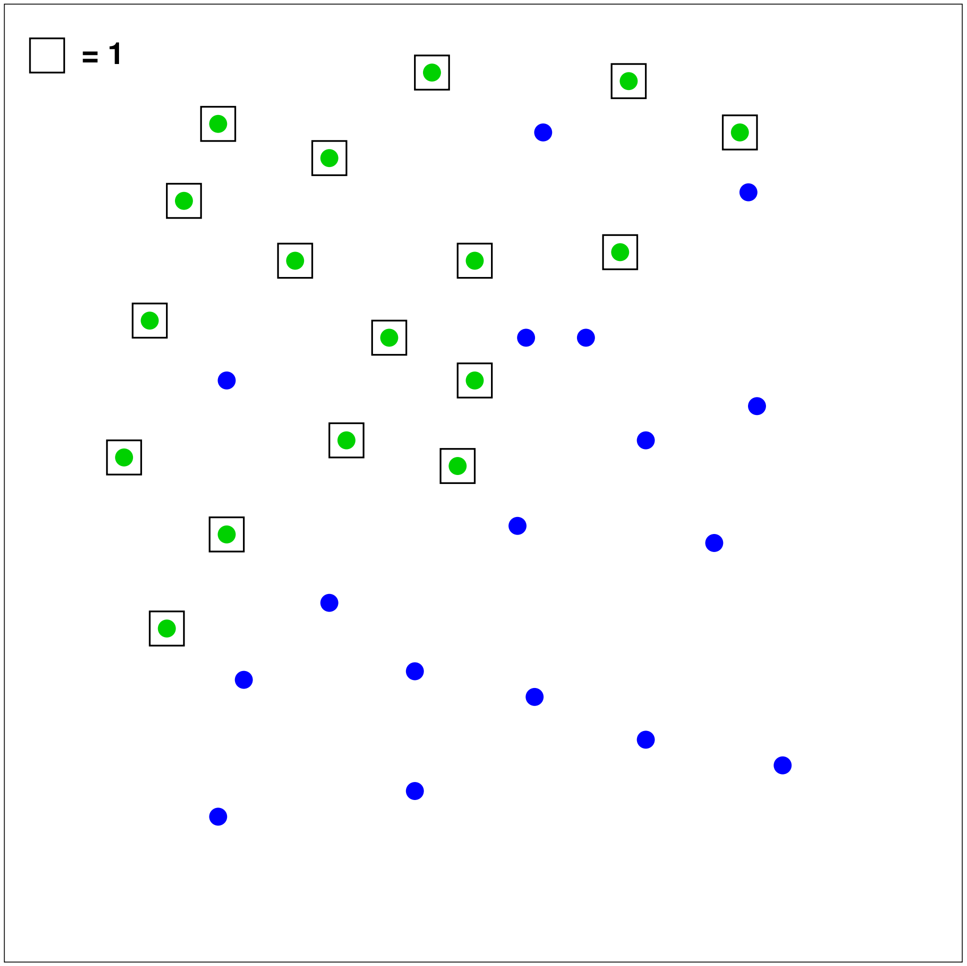
$$n = \binom{m-2}{m/2} + 2$$

K = set of all rules that are conjunctions of $m/2$ features

Optimal rule set for 0-1 loss = disjunction of all rules which has 0-1 loss of 2/n

Optimal rule set for Hamming loss = empty disjunction which has 0-1 loss of (n-2)/n

Overfitting



"Master" LP with Hamming Loss objective

Minimize Hamming loss subject to complexity bound:

loss on positive instances — loss on negative instances

$$\min_{w,\xi} \sum_{i:y_i=1} \xi_i + \sum_{i:y_i=0} \sum_{k \in K} a_{ik} w_k$$

cover positives $\longrightarrow$ $$\xi_i + \sum_{k \in K} a_{ik} w_k \geq 1, \quad \xi_i \geq 0, \qquad i: y_i = 1$$

complexity bound $\longrightarrow$ $$\sum_{k \in K} c_k w_k \leq C$$

select clause k or not $$w_k \in [0,1], \qquad k \in K$$

---

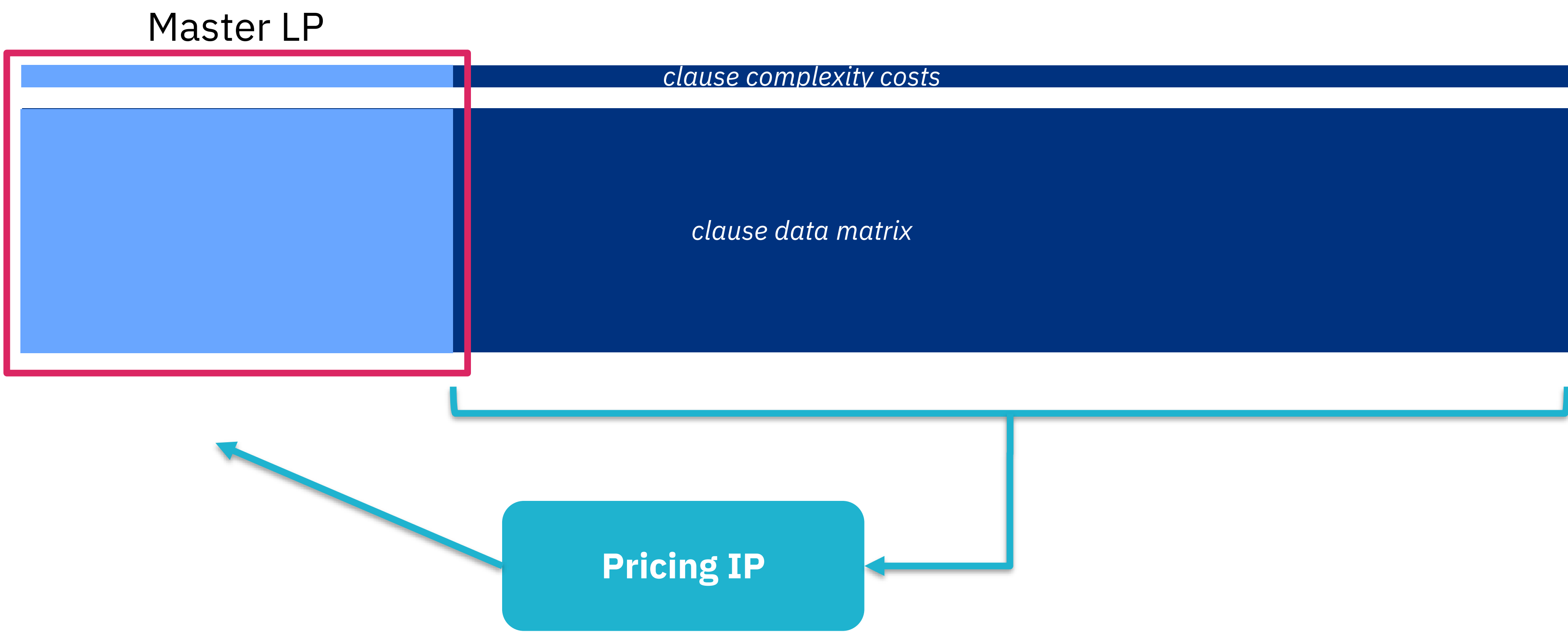
reduced cost of rule *k* $$\sum_{i:y_i=0} a_{ik} - \sum_{i:y_i=1} \mu_i a_{ik} + \lambda c_k$$

Related work

► "Boosting" rule-based classifiers

- Demirez, Bennett, Shawe-Taylor '02: LP-Boost
- Goldberg, Eckstein '10: L_0-RBoost
- Eckstein, Kagawa, Goldberg '17, '19: Rule-enhanced penalized regression

Column generation subproblem



► Almost the same as the **Maximum monomial agreement problem**

- Kearns, Shapire, Sellie ’94
- Goldberg, Shan ’07
- Eckstein, Goldberg ’10, ’12: branch-and-bound method

Column generation subproblem..

X_1	X_2	Y
100	75	0
120	175	1
80	250	1
110	150	0
90	190	1
$\vdots$	$\vdots$	$\vdots$

$\rightarrow$

$X_1 \leq 80$	$X_1 \leq 100$	$X_2 \leq 150$	$X_2 \leq 200$	Y
0	1	1	1	0
0	0	0	1	1
1	1	0	0	1
1	1	1	1	0
0	1	0	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

$\downarrow$

$X_1 \leq 80$	$X_1 \leq 100$	$X_2 \leq 150$	$X_2 \leq 200$	$X_1 \leq 80 \wedge X_2 \leq 150$	$X_1 \leq 100 \wedge X_2 \leq 200$
0	1	1	1	0	1
0	0	0	1	0	0
1	1	0	0	0	0
1	1	1	1	1	1
0	1	0	1	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Pricing Problem

reduced cost of clause incl. complexity penalty

$$\min_{z,a} \quad \sum_{i:y_i=0} a_i - \sum_{i:y_i=1} \mu_i a_i + \lambda\left(1 + \sum_{j=1}^{d} z_j\right)$$

clause
acts as
conjunction
of features

$$a_i = \prod_{j:x_{ij}=0} (1 - z_j) \qquad \forall i$$

at most U
features

$$1 \leq \sum_{j=1}^{d} z_j \leq U, \qquad z_j \in \{0,1\}, j = 1, \dots, d$$

whether to select feature j

Pricing IP

reduced cost of clause incl. complexity penalty

$$\min_{z,a} \sum_{i:y_i=0} a_i - \sum_{i:y_i=1} \mu_i a_i + \lambda\left(1 + \sum\nolimits_{j=1}^{d} z_j\right)$$

clause acts as conjunction of features

$$a_i + z_j \leq 1, \qquad i: y_i = 1, \qquad j: x_{ij} = 0$$

$$a_i + \sum_{j:x_{ij}=0} z_j \geq 1, \qquad a_i \geq 0, \qquad i: y_i = 0$$

at most U features

$$1 \leq \sum\nolimits_{j=1}^{d} z_j \leq U, \qquad z_j \in \{0,1\}, j = 1, \ldots, d$$

whether to select feature j

Solving the pricing IP

IP

▷ Pricing IP is hard to solve, as it has a poor LP relaxation bound, e.g., for the ILPD data set with $U = 5$:

# Binary Features	# data points	Opt. Obj Value	LP Bound
155	520	-9 (after 10 min)	-98

▷ Limit clause size, time limit
▷ Sample data points for large data sets

Clause generation heuristic

For $k = 1, 2, \ldots$

1) Extend previously generated $k-1$-literal clauses to k-literals, choose the best
2) Use bounds to eliminate some of the generated clauses.

Numerical Evaluation

► Main competitors
- Bayesian Rule Sets (BRS): Wang et al. '17]
- Alternating Minimization: (AM) Su, Wei, Varshney, Malioutov '16
- Block Coordinate Descent: (BCD) Su, Wei, Varshney, Malioutov '16
- IDS: Lakkaraju et al. '17 [code was too slow]

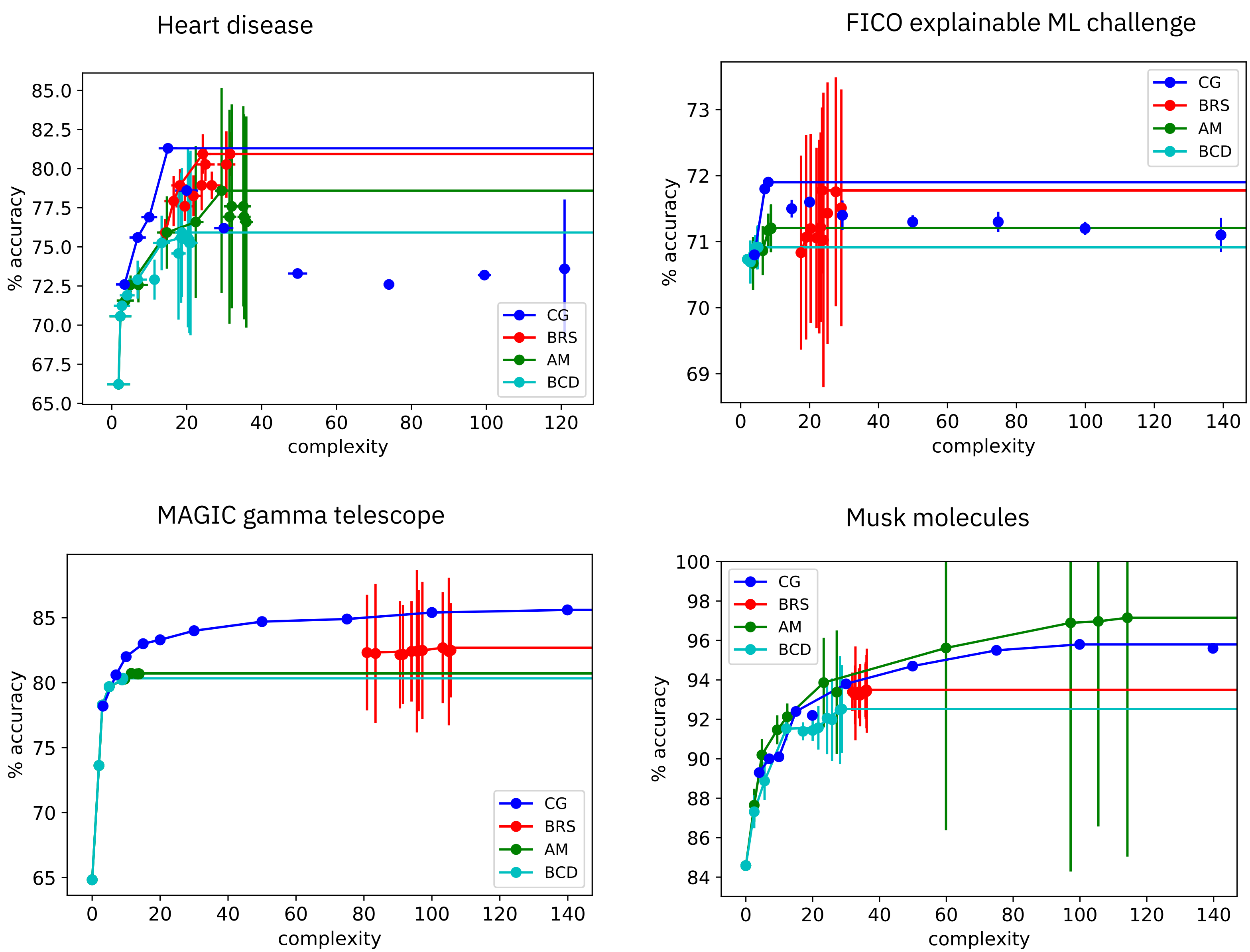
▷ Complexity: # clauses + total # conditions

▷ Accuracy: 10-fold Cross Validation

▷ Binarization
- Sample decile thresholds for numerical features

▷ Time limits for our code: $\leq$ 5 min overall
- Master LP solves fast with CPLEX Barrier
- After column generation, we fix columns and solve IP with CPLEX

Complexity versus Accuracy



Comparison with other methods

Complexity = # clauses + total # conditions

accuracy

dataset	CG	BRS	AM	BCD	RIPPER	CART	RF
adult	83.5	81.7	83.0	82.4	83.6	83.1	84.7
bank	90.0	87.4	90.0	89.7	89.9	89.1	88.7
gas	98.0	92.2	97.6	97.0	99.0	95.4	99.7
magic	85.3	82.5	80.7	80.3	84.5	82.8	86.6
mushroom	100.0	99.7	99.9	99.9	100.0	96.2	99.9
musk	95.6	93.3	96.9	92.1	95.9	90.1	86.2
FICO	71.7	71.2	71.2	70.9	71.8	70.9	73.1

complexity

							size
adult	88.0	39.1	15.0	13.2	133.3	95.9	29,304
bank	9.9	13.2	6.8	2.1	56.4	3.0	37,609
gas	123.9	22.4	62.4	27.8	145.3	104.7	12,518
magic	93.0	97.2	11.5	9.0	177.3	125.5	17,117
mushroom	17.8	17.5	15.4	14.6	17.0	9.3	8,124
musk	123.9	33.9	101.3	24.4	143.4	17.0	5,937
FICO	13.3	23.2	8.7	4.8	88.1	155.0	9,871

FICO 2018 XML Challenge

Predict repayment risk (good/bad) from credit history: roughly 10,000 data points, 23 numerical features, 0/1 labels

- Winning entry:

$$(\text{NumSatTrades} \geq 23 \ \text{ AND } \ \text{ExtRiskEstimate} \geq 71 \ \text{ AND } \ \text{NetFracRevolBurden} \leq 64)$$
$$\text{OR}$$
$$(\text{NumSatTrades} \leq 22 \ \text{ AND } \ \text{ExtRiskEstimate} \geq 76 \ \text{ AND } \ \text{NetFracRevolBurden} \leq 79)$$

Pricing Problem

reduced cost of clause incl. complexity penalty

$$\min_{z,a} \quad \sum_{i:y_i=0} a_i - \sum_{i:y_i=1} \mu_i a_i + \lambda\left(1 + \sum_{j=1}^{d} z_j\right)$$

clause acts as conjunction of features

$$a_i = \prod_{j:x_{ij}=0} (1 - z_j) \qquad \forall i$$

at most U features

$$1 \leq \sum_{j=1}^{d} z_j \leq U, \qquad z_j \in \{0,1\}, j = 1, \ldots, d$$

whether to select feature j

Multilinear optimization

Let $S_1, \ldots, S_m$ be subsets of $\{1, \ldots, n\}$. The pricing problem is $\equiv$

$$\begin{array}{ll} \min & \displaystyle\sum_{i=1}^{m} c_i \delta_i + \sum_{i=1}^{n} f_i z_i \\ \text{s.t.} & \displaystyle\delta_i = \prod_{j \in S_i} z_j, \quad i = 1, \ldots, m \\ & \displaystyle l \leq \sum_{j=1}^{n} z_j \leq u, \;\; z_j \in \{0,1\}, \;\; \delta_i \in \{0,1\} \end{array}$$

An integer linear programming formulation of this problem is given by the “standard linearization inequalities”:

$$0 \leq \delta_i \leq z_j \leq 1$$

$$\delta_i \geq \sum_{j \in S_i} z_j - (|S_i| - 1)$$

Multilinear sets

The **multilinear set**:

$$X = \{(z, \delta) \in \{0,1\}^n \times \{0,1\}^m : \;\; \delta_i = \prod_{j \in S_i} z_j, \quad i = 1, \dots, m\}$$

Del Pia, Khajavirad '16, '18, Del Pia, Khajavirad, Sahinidis '18
Crama, Rodriguez-Heck '17

The **cardinality constrained multilinear set**:

$$X^{l,u} = \{(z, \delta) \in X : \; l \; \leq \sum_{j=1}^{n} z_j \leq \; u\}$$

Mehrotra '97, Fischer, Fischer, McCormick '18

Pricing problem $\equiv$ optimizing a linear function over $X^{l,u}$ $\equiv$ optimizing a linear function over conv$(X^{l,u})$.

Binary polynomial optimization

Fortet '60: Binary polynomial optimization $\equiv$ binary MIP

- Unconstrained binary polynomial optimization $\equiv$ optimizing a linear function over the multilinear set X

Let $\beta, \gamma_i \in \mathbb{R}$, $\alpha_{ij} \in \mathbb{Z}_+$, $S_i \subseteq \{1, \ldots, n\}$

$$f(x) = \beta + \sum_{i=1}^{m} \gamma_i \prod_{j \in S_i} x_j^{\alpha_{ij}} =$$

$$\beta + \sum_{i=1}^{m} \gamma_i \prod_{j \in S_i} x_j \text{ if } x \in 0, 1^n.$$

Convex hull of $X^{l,u}$

$$X = \{(z, \delta) \in \{0,1\}^n \times \{0,1\}^m : \ \delta_i = \prod_{j \in S_i} z_j, \quad i = 1, \ldots, m\}$$

$$X^{l,u} = \{(z, \delta) \in X : \ l \ \leq \sum_{j=1}^{n} z_j \leq \ u\}$$

▷ Fischer, Fischer, McCormick '18: Polyhedral characterization of $X^{l,u}$ when $l = 0$ and nested S_i

Nested S_i: $S_1 \subset S_2 \subset \cdots \subset S_m \subset \{1, \ldots, n\}$

▷ Dash, Günlük, Chen '21: Polyhedral characterization of $X^{l,u}$ for any $0 \leq l < u \leq n$ for nested S_i.

▷ Dash, Günlük, Chen '23: Polyhedral characterization of $X^{l,u}$ for any $0 \leq l < u \leq n$ when $m = 2$.

Notation

I - index set of data points
$P \subseteq I$ - set of data points with label 1
$N \subseteq I$ - set of data points with label 0
neg_k = number of data points in N to which a rule assigns value 1

Assume P and N are partitioned into P_1, P_2 and N_1, N_2.
Interpretation is (P_1, N_1) correspond to one group, (P_2, N_2) to another.
neg_k^l = number of data points in N_l to which a rule assigns value 1

Fairness

Achieve classification parity across multiple groups
Assume we wish to add constraints to 0-1 loss MIP

Equality of opportunity
Difference in rate of loss for positive
instances of group 1 and 2 is bounded

$$\frac{1}{|P_1|}\sum_{i\in P_1}\xi_i - \frac{1}{|P_2|}\sum_{i\in P_2}\xi_i \le \varepsilon$$

$$\frac{1}{|P_2|}\sum_{i\in P_2}\xi_i - \frac{1}{|P_2|}\sum_{i\in P_1}\xi_i \le \varepsilon$$

Equalized odds
Former condition +
Difference in rate of loss for negative
Instances of group 1 and 2 is bounded

$$\frac{1}{|N_1|}\sum_{i\in N_1}\xi_i - \frac{1}{|N_2|}\sum_{i\in N_2}\xi_i \le \varepsilon$$

$$\frac{1}{|N_2|}\sum_{i\in N_2}\xi_i - \frac{1}{|N_1|}\sum_{i\in N_1}\xi_i \le \varepsilon$$

Add more constraints to ensure ξ variables are correctly constrained

Fairness

Achieve classification parity across multiple groups
Lawless, Dash, Gunluk, Wei ’21: add constraints to Hamming loss MIP

Equality of opportunity
Difference in rate of loss for positive instances of group 1 and 2 is bounded

$$\frac{1}{|P_1|}\sum_{i\in P_1}\xi_i - \frac{1}{|P_2|}\sum_{i\in P_2}\xi_i \le \varepsilon$$

$$\frac{1}{|P_2|}\sum_{i\in P_2}\xi_i - \frac{1}{|P_2|}\sum_{i\in P_1}\xi_i \le \varepsilon$$

Equalized odds
Former condition +
Difference in rate of loss for negative
Instances of group 1 and 2 is bounded

$$\frac{1}{|N_1|}\sum_{k}\mathrm{neg}_k^1 w_k - \frac{1}{|N_2|}\sum_{k}\mathrm{neg}_k^2 w_k \le \varepsilon$$

$$\frac{1}{|N_2|}\sum_{k}\mathrm{neg}_k^2 w_k - \frac{1}{|N_1|}\sum_{k}\mathrm{neg}_k^1 w_k \le \varepsilon$$

Add more constraints to ensure ξ variables are correctly constrained

Heavy sets

Let $\mathcal{R}$ be a set of *simple* regions $S \subset \mathbb{R}^N$
Let $|S| = |\{i \mid x_i \in S\}|$, and c_i be weight of datapoint x_i
We focus on sparse AND-rule regions (Malioutov, Dash, Wei '23)

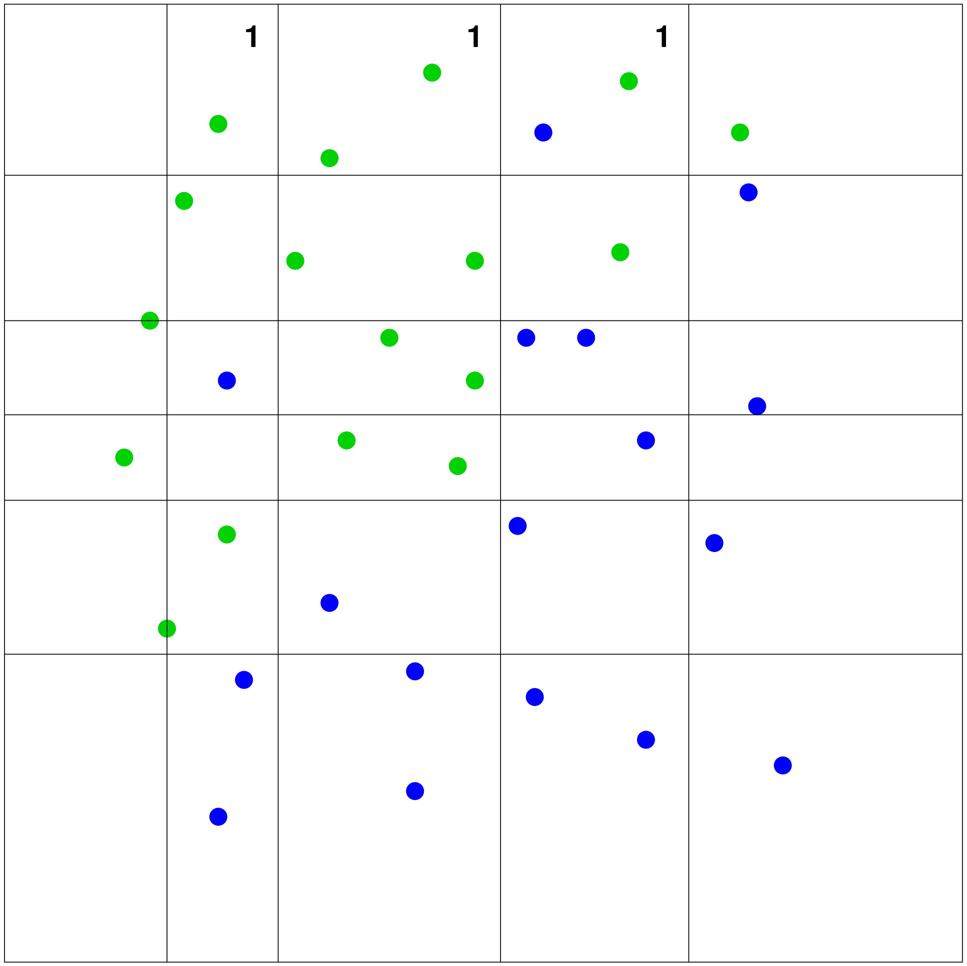
Goal: (1) Find the heaviest-weight simple region from $\mathcal{R}$ subject to region-size constraints

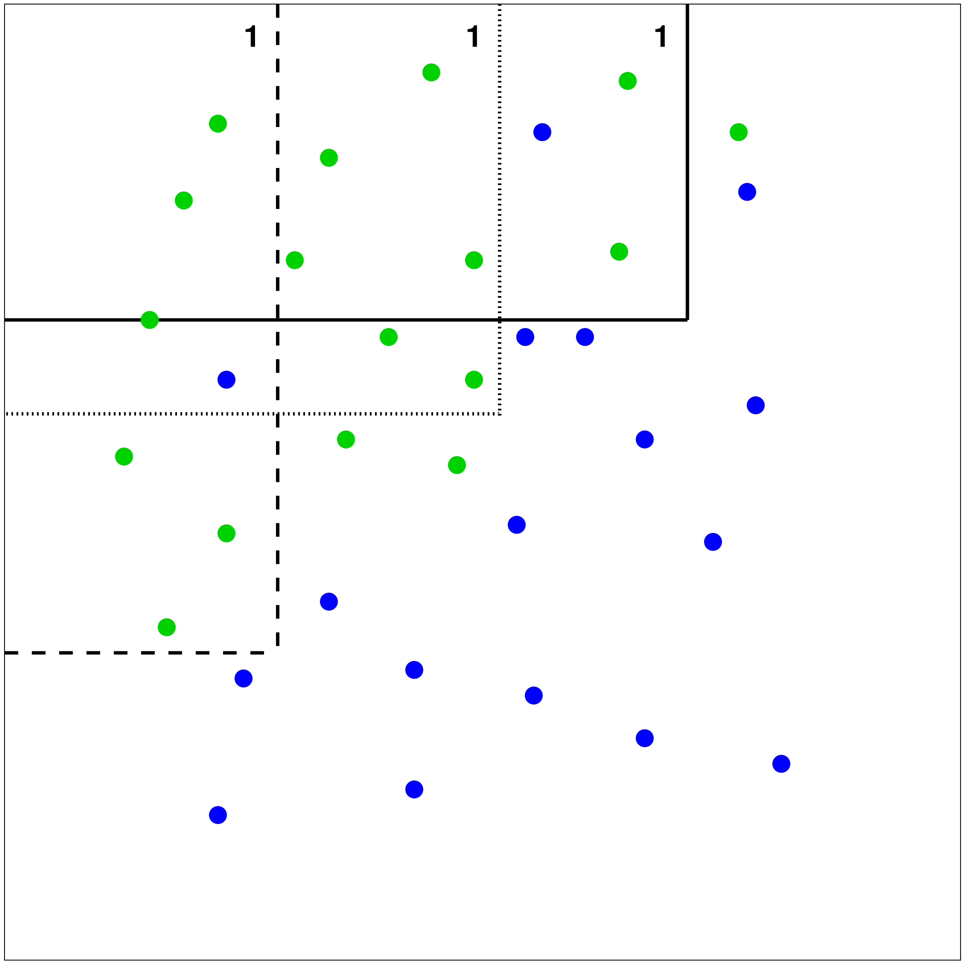
$$S^* = \arg\max_{S \in \mathcal{R}} \sum_{x_i \in S} c_i \ , \quad \text{such that} \quad |S| \leq K,$$

(2) Max average-of-weights with upper and lower bounds on region size:

$$S^* = \arg\max_{S \in \mathcal{R}} \frac{1}{|S|} \sum_{x_i \in S} c_i \ , \quad K_{\min} \leq |S| \leq K_{\max}.$$

Simple regions





Model changes (Let m_A, m_B be two models) $c_i = |m_B(x_i) - m_A(x_i)|^2$
High-error regions $c_i = |m_A(x_i) - y_i|^p$, $p = 1, 2$ for regression
High-variance regions. $c_i = m_A(x_i)^2$, using the max-avg formulation

Heavy sets IP

Let $I = \{1, \ldots, n\}$ be index set of datapoints
$a \in \{0,1\}^I$ is a vector of binary variables; $a_i = 1$ iff datapoint i is in chosen region
J - index set of region boundaries

$$
\begin{aligned}
\text{max} \; & \sum_{i \in I} c_i a_i \\
\text{s.t.} \; & a_i + z_j \leq 1, && \forall i \in I, j \in J : x_{ij} = 0 \\
& a_i + \sum_{j : x_{ij} = 0} z_j \geq 1 && \forall i \in I \\
& 1 \leq \sum_{j \in J} z_j \\
& a_i \in \{0,1\} && \forall i \in I \\
& z_j \in \{0,1\} && \forall j \in J
\end{aligned}
$$

References

1. S. Dash, O. Gunluk, D. Wei, Boolean decision rules via column generation, NeurIPS 2018.
2. R. Chen, S. Dash, O. Gunluk, Multilinear Sets with Two Monomials and Cardinality Constraints, Discrete Applied Mathematics **324**, 67–79 (2023), arXiv:2105.10771
3. C. Lawless, S. Dash, O. Gunluk, D. Wei, Interpretable and fair boolean rule sets via column generation, JMLR 2023, arXiv:2111.08466
4. D. M. Malioutov, S. Dash, D. Wei, Heavy Sets with Applications to Interpretable Machine Learning Diagnostics, AISTATS 2023.