

Learning Interpretable Decision Rule Sets: A Submodular Optimization Approach

Fan Yang, Kai He, Linxiao Yang, Hongxia Du, Jingbang Yang, Bo Yang, Liang Sun

Decision Intelligence Lab, Alibaba DAMO Academy

Hangzhou, China

NeurIPS 2021

Explainable/Interpretable Machine Learning

- Allow human users to interpret predictions made by ML algorithms
- Why interpretability is a concern
 - An essential property requested by trustworthy & human-centered AI
 - Enable decision-makers to determine when to trust or distrust the predictions
 - Nowadays interpretability is becoming one of the key considerations when
 - Deploying ML models to high-stake decision-making scenarios
 - Fitting ML models to understand the data
- Two paradigms
 - Build inherently interpretable ML models
 - E.g., rule models, sparse linear models, generalized additive models
 - Provide post-hoc explanations for black-box models
 - E.g., Shapley values, integrated gradients, counterfactual explanation

Interpretable Rule Models

- Rule models: longstanding attempt towards interpretable ML
 - Making predictions with human-understandable logical rules
 - Particularly suited for tabular data
 - Contain mixed-type features and exhibit complex high-order feature interactions
 - To be interpretable, a rule model should be **simple**
 - Model complexity is not explicitly optimized by traditional rule learning algorithms
 - E.g., it is not easy to understand a deep CART

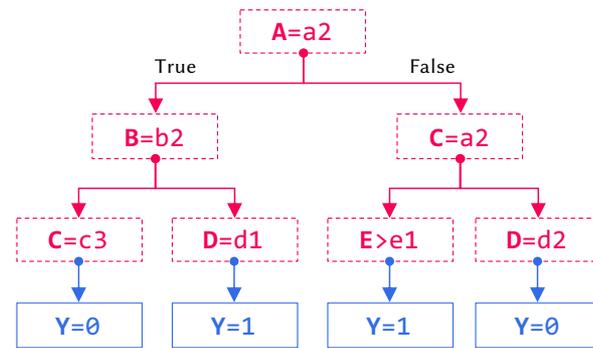
rule set

```
IF ( A=a2 AND B=b2 )
OR ( A=a2 AND C=c3 )
OR ( D=d1 AND e1<=E<=e3 )
OR ( ... )
THEN Y=1
ELSE Y=0
```

rule list

```
IF ( A=a2 AND B=b2 )
THEN Y=1
ELIF ( A=a4 AND C=c2 )
THEN Y=0
ELIF ( D=d1 AND e1<=E<=e3 )
THEN Y=1
ELSE Y=0
```

decision tree



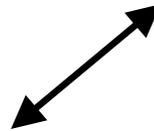
Rule Sets

A	B	C	D	E	Y
a1	b3	c1	d2	e1	0
a2	b2	c3	d1	e4	1
...

rule set
learning



```
IF ( A=a1 AND B=b2 )  
OR ( A=a2 AND C=c3 )  
OR ( D=d1 AND e1<=E<=e3 )  
OR ( ... )  
THEN Y=1 ELSE Y=0
```



```
IF ( A=a1 AND B=b2 ) THEN Y=1  
IF ( A=a2 AND C=c3 ) THEN Y=1  
IF ( D=d1 AND e1<=E<=e3 ) THEN Y=1  
IF ( ... ) THEN Y=1
```

- Have a simpler combinatorial structure than rule lists and decision trees
 - Easier to interpret and to learn from data

Rule Sets in Disjunctive Normal Form

A	B	C	D	E	Y
a1	b3	c1	d2	e1	0
a2	b2	c3	d1	e4	1
...

rule set
learning



```

IF  ( A=a1 AND B=b2 )
OR  ( A=a2 AND C=c3 )
OR  ( D=d1 AND e1<=E<=e3 )
OR  ( ... )
THEN Y=1 ELSE Y=0
    
```

one-hot
encoding



A= a1	A= a2	B= b1	B= b2	B= b3	...	Y
1	0	0	0	1	...	0
0	1	0	1	0	...	1
...

DNF
learning



```

Y = ( A=a1 ∧ B=b2 )
∨ ( A=a2 ∧ C=c3 )
∨ ( D=d1 ∧ E>=e1 ∧ E<=e3 )
∨ ( ... )
    
```

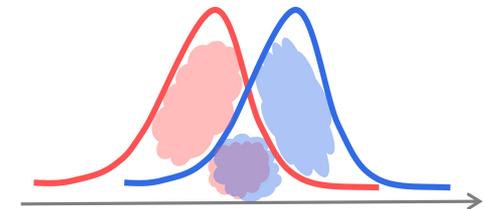
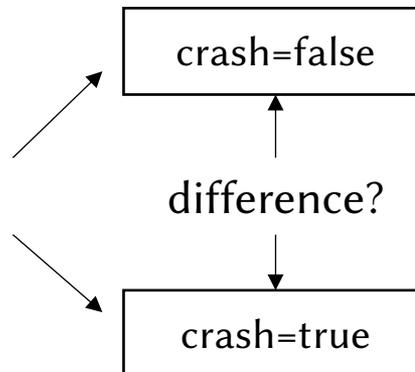


Applications

- White-box predictive model for high-stake decision-making
 - E.g., loan approval, crime prediction

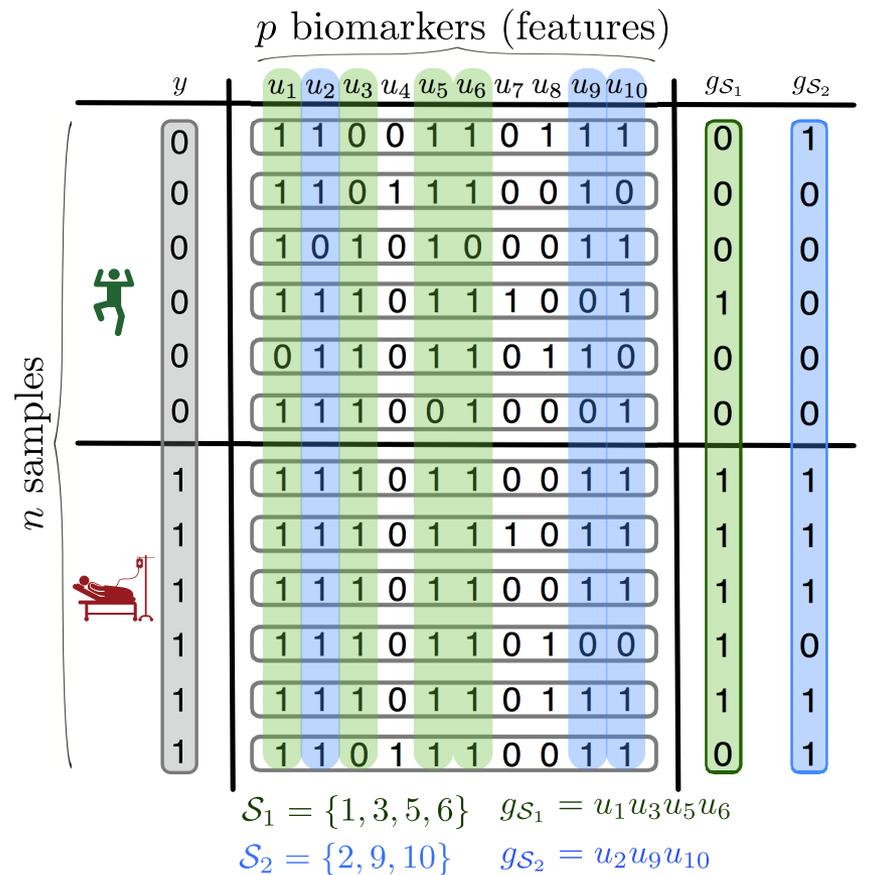
- Explaining data differences

app_version	device_type	os	crash
v1	iPhone X	11.0	false
...
v2	Galaxy S9	8.0	true
...
v3	HTC One	8.0	false



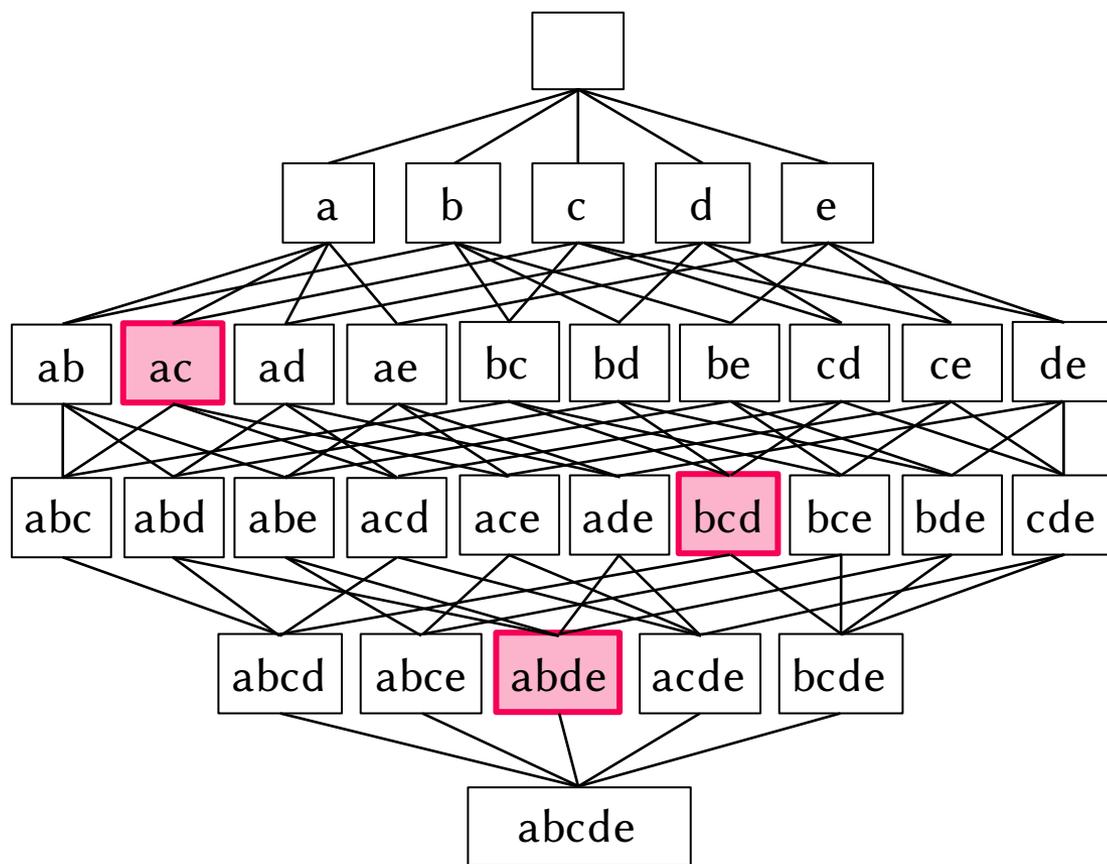
Applications (Cont.)

➤ Interaction detection



Rule Set Learning

- Major challenge: exponentially sized search space



d literals (i.e.,
binary features)

↓
 2^d conjunctions

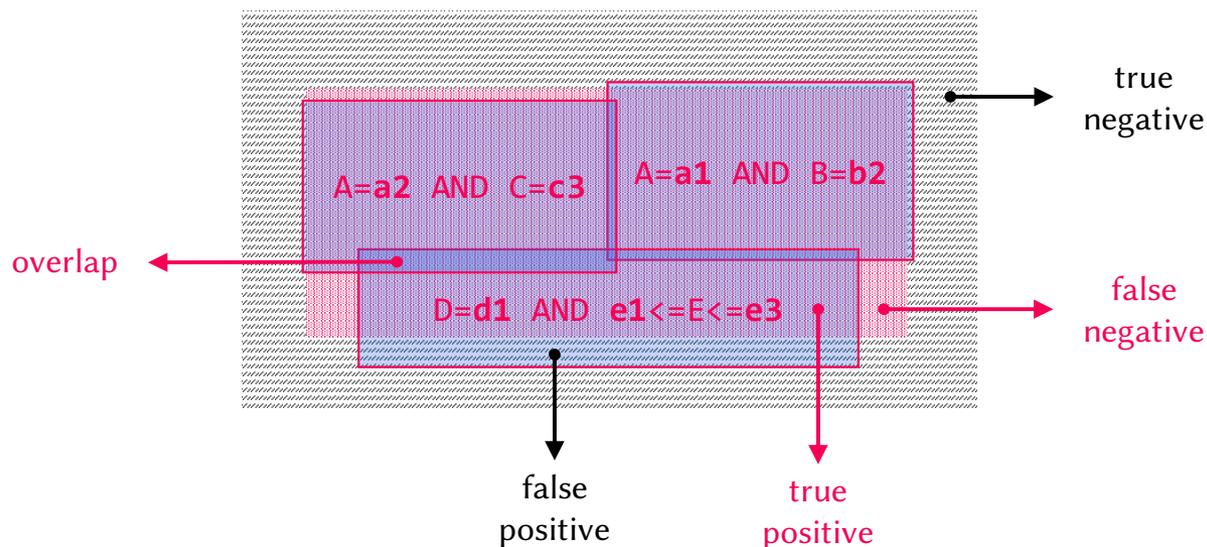
↓
 2^{2^d} DNFs

Rule Set Learning

- Goal: Learn an **interpretable** and **accurate** rule set $S \subseteq 2^{[d]}$
 - In which each rule $R \in S$ is a subset of $[d]$

$$L(S) = \beta_0 \sum_{R \in S} |\mathcal{X}_{\{R\}}^-| + \beta_1 |\mathcal{X}^+ \setminus \mathcal{X}_S^+| + \beta_2 \left(\sum_{R \in S} |\mathcal{X}_{\{R\}}^+| - |\mathcal{X}_S^+| \right) + \lambda \sum_{R \in S} |\mathcal{R}|$$

The equation is presented in four boxes: FP, FN, Overlap, and Complexity.



Submodularity

Minimize

$$L(\mathcal{S}) = \beta_0 \sum_{\mathcal{R} \in \mathcal{S}} |\mathcal{X}_{\{\mathcal{R}\}}^-| + \beta_1 |\mathcal{X}^+ \setminus \mathcal{X}_{\mathcal{S}}^+| + \beta_2 \left(\sum_{\mathcal{R} \in \mathcal{S}} |\mathcal{X}_{\{\mathcal{R}\}}^+| - |\mathcal{X}_{\mathcal{S}}^+| \right) + \lambda \sum_{\mathcal{R} \in \mathcal{S}} |\mathcal{R}|$$

Reorganize

$$L(\mathcal{S}) = \beta_1 |\mathcal{X}^+| - (\beta_1 + \beta_2) |\mathcal{X}_{\mathcal{S}}^+| + \sum_{\mathcal{R} \in \mathcal{S}} \beta_0 |\mathcal{X}_{\{\mathcal{R}\}}^-| + \beta_2 |\mathcal{X}_{\{\mathcal{R}\}}^+| + \lambda |\mathcal{R}|$$

Maximize

$$V(\mathcal{S}) = g(\mathcal{S}) - \sum_{\mathcal{R} \in \mathcal{S}} c(\mathcal{R})$$

Regularized Submodular Maximization

- Cardinality constrained submodular maximization

$$\max_{\mathcal{S} \subseteq 2^{[d]}, |\mathcal{S}| \leq K} V(\mathcal{S}) \quad (4)$$

- Distorted greedy algorithm

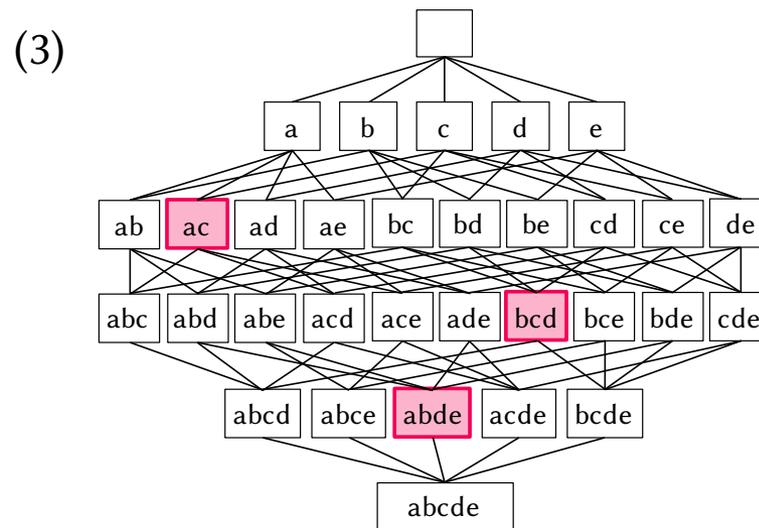
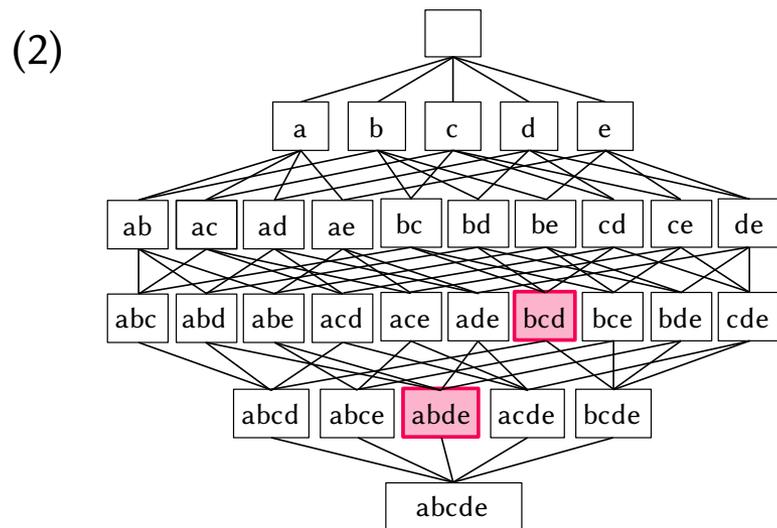
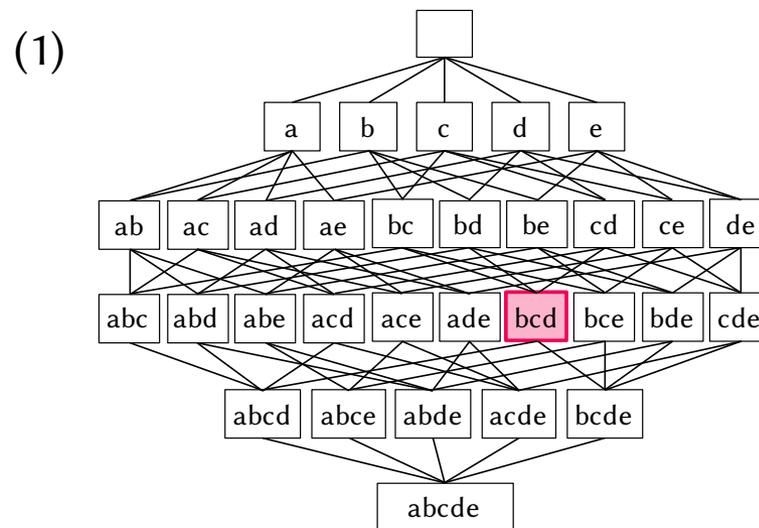
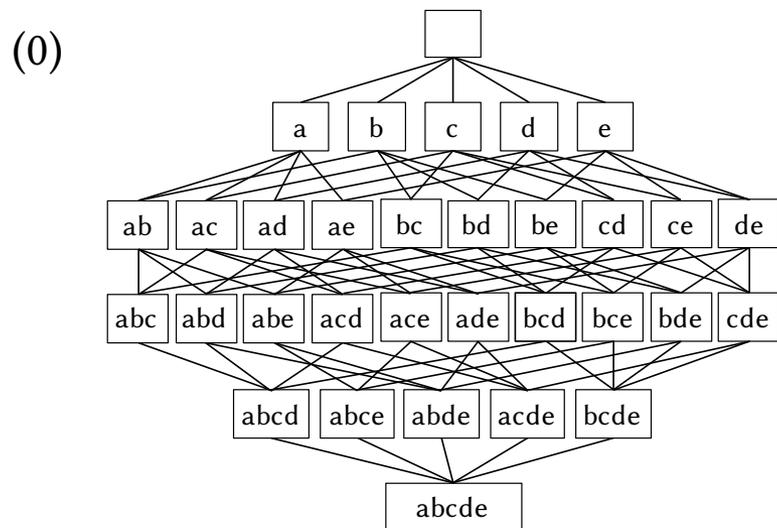
Algorithm 1 Rule set learning

```
1 Input: Training data  $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ , hyperparameters  $(\beta, \lambda)$ , cardinality  $K$ 
2 Initialize  $\mathcal{S} \leftarrow \emptyset$ 
3 for  $k = 1$  to  $K$  do
4   Define  $v_k(\mathcal{R}) = (1 - 1/K)^{K-k} g(\mathcal{R}|\mathcal{S}) - c(\mathcal{R})$            /*  $g(\mathcal{R}|\mathcal{S}) := g(\mathcal{S} \cup \{\mathcal{R}\}) - g(\mathcal{S})$  */
5   Solve  $\mathcal{R}^* \leftarrow \arg \max_{\mathcal{R} \subseteq [d]} v_k(\mathcal{R})$ 
6   if  $v_k(\mathcal{R}^*) > 0$  then  $\mathcal{S} \leftarrow \mathcal{S} \cup \{\mathcal{R}^*\}$  end if
7 end for
8 Output:  $\mathcal{S}$ 
```

- Approximation guarantee

$$V(\mathcal{S}) = g(\mathcal{S}) - \sum_{\mathcal{R} \in \mathcal{S}} c(\mathcal{R}) \geq (1 - 1/e)g(OPT) - \sum_{\mathcal{R} \in OPT} c(\mathcal{R})$$

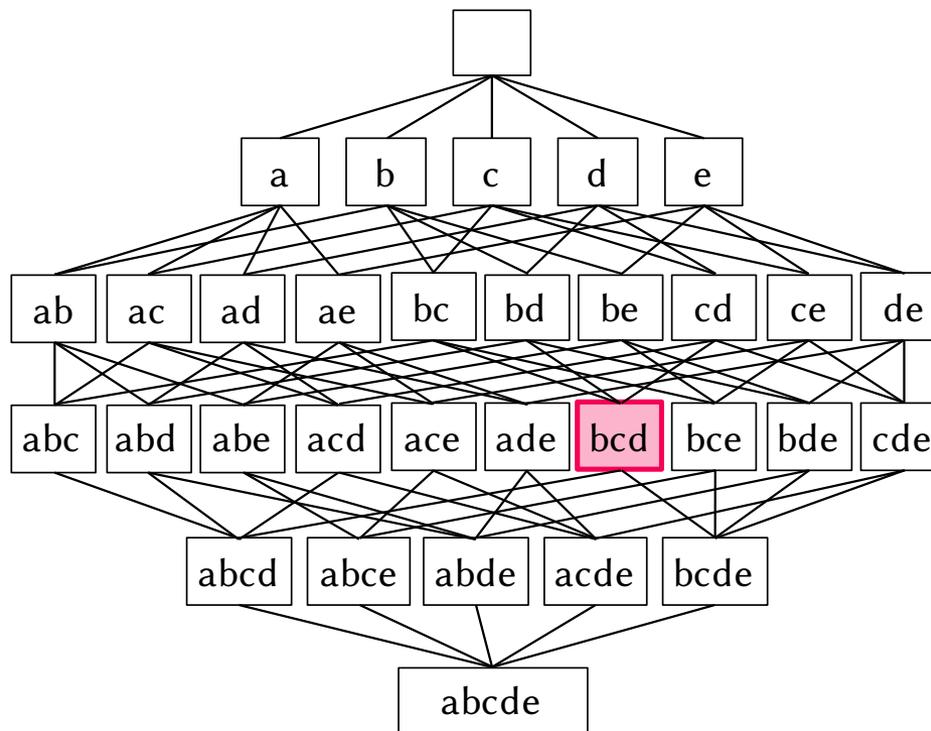
Distorted Greedy



Marginal Gain Maximization

- 4 Define $v_k(\mathcal{R}) = (1 - 1/K)^{K-k} g(\mathcal{R}|\mathcal{S}) - c(\mathcal{R})$
- 5 Solve $\mathcal{R}^* \leftarrow \arg \max_{\mathcal{R} \subseteq [d]} v_k(\mathcal{R})$

➤ Exhaustive enumeration: $O(2^d)$ ☹️

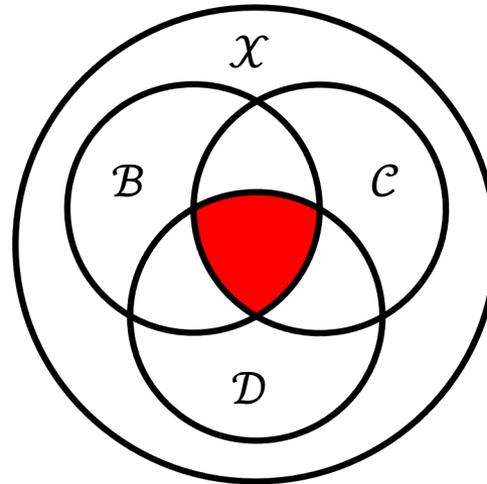
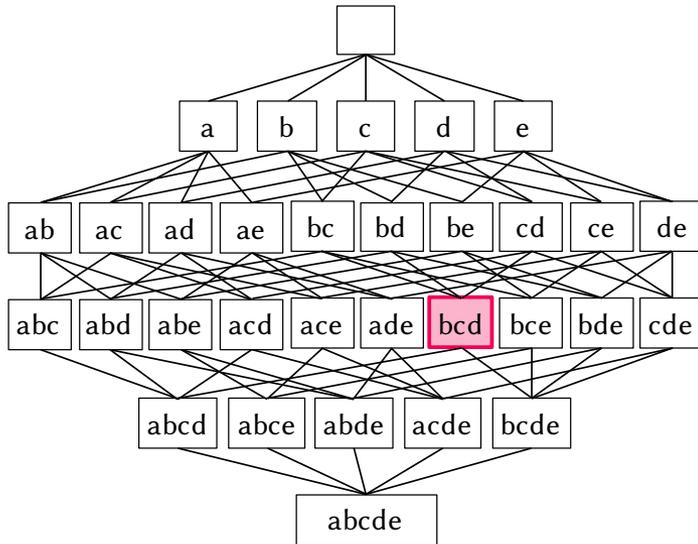


Can We Exploit Submodularity One More Time?

➤ Sadly, $v(\mathcal{R})$ is not a submodular set function

$$v(\mathcal{R}) = \sum_{i=1}^n \omega_i \mathbb{1}_{\mathcal{R} \subseteq \mathbf{x}_i} - \lambda |\mathcal{R}|$$

➤ Examples covered by a rule: common examples covered by its features



set intersection
(non-submodular)

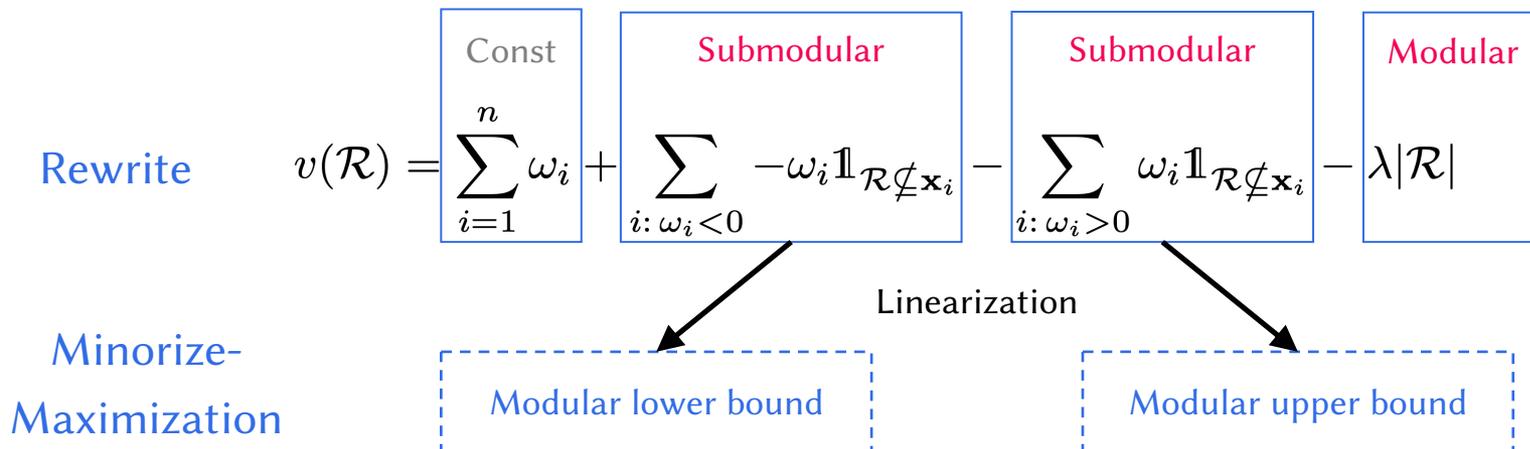
$$B \cap C \cap D \Leftrightarrow X \setminus (\bar{B} \cup \bar{C} \cup \bar{D})$$

set union
(submodular)

Approximate Subproblem Solving

- We rewrite the subobjective as a difference of submodular (DS) functions
- Based on this DS decomposition, an iterative refinement algorithm is proposed to solve the subproblem approximately

Maximize $v(\mathcal{R}) = \sum_{i=1}^n \omega_i \mathbb{1}_{\mathcal{R} \subseteq \mathbf{x}_i} - \lambda |\mathcal{R}|$



Experiments

➤ Predictive performance

Table 1: Predictive performance measured by average test accuracy (%).

Dataset	#samples	#features	Ours	RIPPER	BRS	CG	CART	RF
tic-tac-toe	958	54	100.0 (0.0)	99.7 (0.7)	100.0 (0.0)	100.0 (0.0)	94.2 (1.9)	99.1 (0.9)
liver	345	104	69.5 (5.1)	66.0 (5.8)	60.6 (8.3)	68.7 (5.4)	68.6 (6.3)	73.9 (9.3)
heart	303	118	82.2 (7.7)	76.2 (7.7)	79.7 (7.5)	78.0 (6.8)	82.2 (6.1)	82.8 (7.1)
ionosphere	351	566	91.4 (5.4)	87.2 (7.5)	85.0 (4.2)	90.6 (4.4)	89.5 (3.3)	94.0 (3.4)
ILPD	583	160	71.4 (0.8)	57.8 (7.7)	69.0 (5.3)	71.7 (3.4)	69.4 (6.4)	71.2 (4.0)
WDBC	569	540	94.0 (4.8)	94.7 (1.6)	93.9 (1.2)	94.7 (3.4)	93.5 (3.8)	97.0 (3.6)
pima	768	134	75.4 (4.3)	75.9 (3.3)	72.2 (3.3)	74.0 (3.4)	75.4 (5.5)	76.9 (3.3)
transfusion	748	64	78.1 (3.2)	78.2 (2.7)	77.1 (5.1)	78.2 (3.6)	78.7 (2.8)	79.7 (2.8)
banknote	1372	72	98.7 (1.0)	92.8 (2.4)	91.1 (2.5)	98.8 (0.9)	99.1 (1.2)	99.6 (0.6)
mushroom	8124	224	100.0 (0.0)	100.0 (0.0)	99.7 (0.2)	99.9 (0.1)	100.0 (0.0)	100.0 (0.0)
COMPAS-2016	5020	30	66.5 (2.3)	57.7 (1.0)	63.4 (1.7)	66.7 (2.2)	66.2 (2.2)	66.6 (2.5)
COMPAS-binary	6907	24	67.0 (1.5)	56.0 (0.6)	65.5 (1.7)	66.4 (1.9)	67.3 (1.5)	67.3 (1.6)
FICO-binary	10459	34	71.2 (1.1)	60.1 (1.2)	70.5 (1.1)	71.1 (1.2)	71.9 (1.4)	72.3 (1.4)
COMPAS	12381	180	73.3 (1.3)	72.3 (1.5)	70.7 (1.1)	N/A	72.2 (1.4)	73.8 (1.1)
FICO	10459	312	70.4 (1.2)	69.1 (1.9)	70.1 (0.9)	71.0 (0.7)	70.9 (1.1)	72.3 (0.8)
adult	48842	262	84.4 (0.6)	83.3 (0.9)	80.3 (1.4)	82.8 (0.4)	83.7 (0.4)	84.7 (0.5)
bank-market	11162	174	84.4 (0.8)	82.9 (1.1)	76.9 (1.2)	82.3 (0.9)	83.0 (1.0)	85.2 (0.9)
magic	19020	180	84.6 (0.8)	82.2 (1.3)	N/A	80.8 (1.0)	84.7 (0.5)	86.7 (0.5)
musk	6598	2922	97.3 (0.8)	96.1 (0.8)	90.2 (2.0)	95.0 (0.7)	96.0 (0.9)	97.7 (0.6)
gas	13910	2304	98.2 (0.4)	99.0 (0.4)	N/A	95.9 (0.7)	99.0 (0.3)	99.8 (0.1)

Experiments

➤ Interpretability

Table 3: Examples of learned rule sets.

Dataset
mushroom
odor != a AND odor != l AND odor != n spore_print color = r
gill_size = n AND stalk_surface_below_ring = y
cap_color = w AND population = c
tic-tac-toe
top-right = x AND middle-middle = x AND bottom-left = x top-left = x AND middle-middle = x AND bottom-right = x middle-left = x AND middle-middle = x AND middle-right = x bottom-left = x AND bottom-middle = x AND bottom-right = x top-left = x AND top-middle = x AND top-right = x top-left = x AND middle-left = x AND bottom-left = x top-right = x AND middle-right = x AND bottom-right = x top-middle = x AND middle-middle = x AND bottom-middle = x

even simpler than the rule
given in dataset description:

odor=n AND stalk-surface-below-ring=y
AND stalk-color-above-ring != n

Table 2: Interpretability measured by number of rules, number of literals, and overlap among rules.

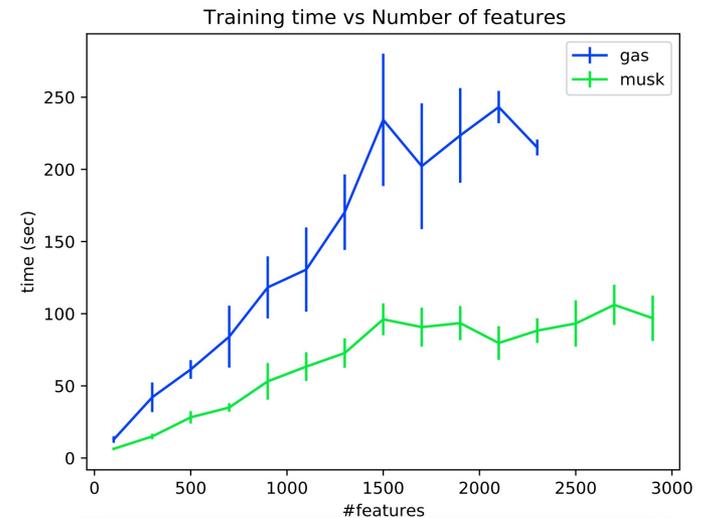
Dataset	#Rules				#Literals				Overlap (%)		
	Ours	RIPPER	CG	CART	Ours	RIPPER	CG	CART	Ours	RIPPER	CG
tic-tac-toe	8.0 (0.0)	9.5 (1.4)	8.0 (0.0)	69.9 (3.6)	24.0 (0.0)	31.1 (5.8)	24.3 (0.5)	138.8 (7.1)	2.3 (1.2)	52.8 (8.1)	23.3 (0.5)
liver	18.0 (2.4)	2.1 (0.7)	14.5 (1.2)	5.0 (0.0)	83.8 (10.5)	7.1 (3.3)	58.5 (4.9)	9.0 (0.0)	7.5 (4.9)	28.0 (17.7)	9.7 (1.7)
heart	2.1 (0.3)	4.0 (1.1)	10.3 (0.8)	11.4 (1.1)	4.4 (1.3)	11.0 (3.8)	41.5 (3.2)	21.8 (2.1)	16.8 (7.7)	48.4 (4.9)	27.4 (2.4)
ionosphere	2.0 (0.7)	3.6 (0.8)	4.3 (0.8)	24.7 (2.1)	8.0 (2.4)	12.5 (3.1)	20.3 (3.8)	48.4 (4.2)	3.4 (5.0)	57.2 (7.9)	32.1 (7.1)
ILPD	1.1 (0.3)	2.6 (0.5)	2.0 (0.0)	4.3 (0.48)	0.2 (0.6)	7.0 (1.5)	3.0 (0.0)	7.6 (1.0)	0.0 (0.0)	31.7 (6.7)	0.0 (0.1)
WDBC	8.0 (1.1)	5.0 (1.1)	5.3 (0.6)	7.9 (1.0)	27.7 (3.0)	10.6 (3.0)	13.4 (1.6)	14.8 (2.0)	2.4 (6.0)	35.0 (5.3)	26.8 (1.0)
pima	3.2 (2.2)	3.6 (1.3)	6.7 (1.7)	10.1 (0.6)	10.0 (10.7)	13.0 (6.5)	20.0 (6.5)	19.2 (1.1)	2.6 (3.1)	27.0 (8.2)	5.4 (1.2)
transfusion	1.4 (0.8)	2.1 (0.7)	2.7 (0.5)	11.0 (0.5)	4.3 (2.8)	9.0 (3.0)	7.9 (1.6)	21.0 (0.9)	0.0 (0.0)	21.3 (13.1)	0.8 (0.5)
banknote	8.7 (1.3)	7.4 (1.3)	4.0 (0.0)	28.2 (0.7)	32.8 (5.8)	21.0 (4.2)	10.9 (1.4)	55.4 (5.4)	0.1 (0.3)	41.0 (3.2)	9.5 (0.7)
mushroom	3.9 (0.3)	6.1 (1.1)	5.0 (0.0)	14.1 (0.6)	8.4 (1.3)	10.9 (1.1)	7.0 (0.0)	27.2 (1.1)	0.0 (0.0)	41.6 (15.8)	21.0 (0.2)
COMPAS-2016	11.8 (4.6)	9.0 (1.6)	3.0 (0.0)	31.0 (0.7)	41.9 (20.1)	31.4 (7.0)	6.0 (0.0)	61.0 (1.3)	0.0 (0.1)	30.9 (0.2)	0.5 (0.2)
COMPAS-binary	11.4 (1.3)	12.0 (1.7)	2.9 (0.3)	78.0 (1.1)	42.2 (6.3)	48.1 (8.8)	5.4 (0.9)	155.0 (2.1)	0.1 (0.2)	32.0 (0.7)	0.0 (0.0)
FICO-binary	21.6 (3.3)	16.7 (2.1)	2.0 (0.0)	158.5 (2.7)	134.0 (23.8)	106.8 (14.3)	4.0 (0.0)	316.0 (5.4)	0.3 (0.3)	37.8 (0.9)	0.0 (0.0)
COMPAS	5.5 (2.7)	12.0 (2.7)	N/A	85.9 (2.3)	24.0 (15.8)	66.0 (15.3)	N/A	170.8 (4.7)	0.2 (0.3)	21.2 (3.5)	N/A
FICO	16.0 (5.5)	16.7 (3.9)	1.1 (0.3)	69.5 (1.4)	118.6 (39.3)	99.7 (26.6)	1.4 (1.2)	138.0 (2.7)	3.5 (1.3)	39.8 (3.8)	0.0 (0.0)
adult	9.1 (3.1)	42.7 (15.2)	2.0 (0.0)	398.3 (4.9)	83.4 (30.7)	337.0 (128.9)	5.7 (0.5)	795.6 (9.9)	0.8 (0.6)	25.8 (7.0)	1.8 (0.4)
bank-market	17.6 (3.2)	43.8 (7.1)	11.0 (0.0)	289.0 (2.8)	118.5 (15.1)	269.1 (49.2)	18.7 (0.8)	577.0 (5.7)	3.6 (1.9)	43.8 (3.6)	7.8 (0.4)
magic	19.1 (6.6)	52.3 (10.6)	2.7 (0.5)	398.9 (4.7)	136.1 (49.2)	391.3 (75.0)	6.2 (0.4)	796.8 (9.3)	2.8 (1.7)	50.0 (0.0)	7.6 (2.8)
musk	8.9 (1.8)	19.6 (1.5)	5.0 (0.4)	180.0 (7.1)	61.2 (14.8)	101.4 (7.6)	21.0 (1.9)	359.0 (14.2)	0.8 (0.5)	36.9 (3.9)	2.7 (1.4)
gas	13.1 (1.0)	24.2 (1.8)	4.0 (0.0)	172.2 (2.3)	74.2 (4.2)	106.8 (11.6)	14.7 (1.5)	343.4 (4.7)	15.2 (1.7)	50.0 (0.0)	22.1 (3.0)

Experiments

➤ Scalability

Table 5: Average running time in seconds.

Dataset	Ours	CG	RIPPER	BRS	CART	RF
tic-tac-toe	0.794	12.815	0.204	14.833	0.002	0.097
liver	4.113	62.482	0.232	19.513	0.002	0.083
heart	0.853	62.840	0.227	14.858	0.001	0.077
ionosphere	6.064	51.475	0.914	17.304	0.008	0.090
ILPD	0.909	81.869	0.325	23.254	0.004	0.097
WDBC	8.209	23.009	1.042	26.592	0.008	0.091
pima	1.580	66.515	0.471	54.542	0.005	0.105
transfusion	0.679	8.246	0.208	21.857	0.001	0.095
banknote	2.142	13.043	0.274	659.874	0.002	0.093
mushroom	1.637	16.369	2.083	48.763	0.031	0.252
COMPAS-2016	2.860	14.914	1.243	33.815	0.003	0.159
COMPAS-binary	3.380	16.151	2.178	41.120	0.003	0.174
FICO-binary	7.705	11.199	6.890	72.515	0.016	0.432
COMPAS	16.534	N/A	10.359	237.615	0.083	0.897
FICO	33.935	159.838	18.826	695.484	0.215	1.121
adult	15.952	288.338	202.279	39787.330	0.815	4.802
bank-market	34.185	107.736	30.563	8956.680	0.124	0.842
magic	39.432	222.451	65.904	N/A	0.197	1.459
musk	88.215	659.791	371.562	864.823	1.388	1.644
gas	192.125	5353.880	582.690	N/A	2.331	2.772



Experiments

➤ Approximation quality

■ Exact versus approximate subproblem solving

Table 6: Approximation quality measured by relative gaps.

Dataset	#features	$V(\mathcal{S}_{\text{approx}})$	$V(\mathcal{S}_{\text{bnb}})$	Relative Gap
COMPAS-binary	24	871.00	875.00	0.0046
COMPAS-2016	30	594.40	590.00	-0.0075
FICO-binary	34	1977.00	1919.00	-0.0302
tic-tac-toe	54	433.78	433.78	0.0000
transfusion	64	12.00	12.00	0.0000
banknote	72	599.40	602.40	0.0050
heart	118	99.48	99.48	0.0000
ILPD	160	217.00	217.00	0.0000
mushroom	224	3908.00	3908.00	0.0000
liver	104	127.68	124.69	-0.0240
pima	134	74.84	76.00	0.0153
bank-market	174	3329.07	3323.59	-0.0016
magic	180	9251.09	9193.73	-0.0062
COMPAS	180	563.00	642.57	0.1238
adult	262	3690.00	3665.10	-0.0068
FICO	312	1936.30	1927.00	-0.0048
WDBC	540	209.00	207.01	-0.0096
ionosphere	566	198.80	199.20	0.0020
musk	2922	565.50	609.90	0.0728
gas	2304	6234.64	6181.82	-0.0085