IMLI: An Incremental Framework for MaxSAT-Based Learning of Interpretable Classification Rules

Bishwamittra Ghosh

Joint work with
Kuldeep S. Meel

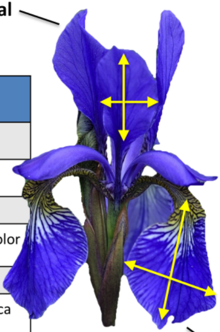
NUS National University of Singapore

Applications of Machine Learning

Example Dataset

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica



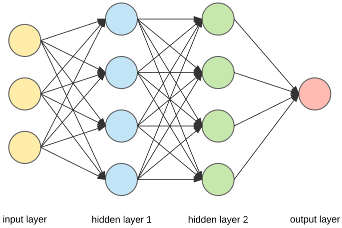
Features
(attributes, measurements, dimensions)

Class labels
(targets)

Representation of an interpretable model and a black box model

A sample is Iris Versicolor if
(sepal length > 6.3 **OR** sepal width > 3
OR petal width ≤ 1.5)
AND
(sepal width ≤ 2.7 **OR** petal length > 4
OR petal width > 1.2)
AND
(petal length ≤ 5)

Interpretable Model



Black Box Model

Formula

- A CNF (Conjunctive Normal Form) formula is a conjunction of clauses where each clause is a disjunction of literals
- A DNF (Disjunctive Normal Form) formula is a disjunction of clauses where each clause is a conjunction of literals
- Example
  - CNF: $(a \vee b \vee c) \wedge (d \vee e)$
  - DNF: $(a \wedge b \wedge c) \vee (d \wedge e)$
- Decision rules in CNF and DNF are highly interpretable [Malioutov'18; Lakkaraju'19]

Expectation from a ML model

- Model needs to be interpretable
- End users should understand the reasoning behind decision-making
- Examples of interpretable models:
  - Decision tree
  - Decision rules (If-Else rules)
  - ...

Definition of Interpretability in Rule-based Classification

- There exists different notions of interpretability of rules
- Rules with fewer terms are considered interpretable in medical domains [Letham'15]
- We consider rule size as a proxy of interpretability for rule-based classifiers
- Rule size = number of literals

Outline

Motivation

- Recently a MaxSAT-based interpretable rule learning framework MLIC has been [Malioutov'18]
- MLIC learns interpretable rules expressed as **CNF**
- The number of clauses in the query is linear with the number of samples in the dataset
- Suffers from poor scalability for large datasets

Can we design?

A sound framework-

- takes benefit of success of MaxSAT solving
- scales to large dataset
- provides interpretability
- achieves competitive prediction accuracy

IMLI: **I**ncremental approach to **M**axSAT-based **L**earning of **I**nterpretable Rules

- p is the number of partition
- n is the number of samples
- The number of clauses in MaxSAT query is $\mathcal{O}(\frac{n}{p})$

Continued...

- consider binary variables b_i for feature i
- $b_i = \mathbb{1}\{\text{feature } i \text{ is selected in } \mathcal{R}\}$
- Consider assignment $b_1 = 1, b_2 = 0, b_3 = 0, b_4 = 1$

$$\mathcal{R} = (1^{\text{st}} \text{ feature } \textbf{OR } 4^{\text{th}} \text{ feature})$$

Continued...

In MaxSAT

- **Hard Clause:** always satisfied, weight $= \infty$
- **Soft Clause:** can be falsified, weight $= \mathbb{R}^+$

MaxSAT finds an assignment that satisfies all hard clauses and most soft clauses such that the weight of satisfied soft clauses is maximize

Continued...

$(i-1)$-th partition

we learn assignment

- $b_1 = 0$
- $b_2 = 1$
- $b_3 = 0$
- $b_4 = 1$

i-th partition

we construct soft unit clause

- $\neg b_1$
- b_2
- $\neg b_3$
- b_4

Experimental Results

Accuracy and training time of different classifiers

Dataset	Size	Features	RF	SVC	RIPPER	MLIC	IMLI
PIMA	768	134	76.62 (1.99)	75.32 (0.37)	75.32 (2.58)	**75.97** Timeout	73.38 (**0.74**)
Tom's HW	28179	844	97.11 (27.11)	96.83 (354.15)	96.75 (37.81)	96.61 Timeout	**96.86** (**23.67**)
Adult	32561	262	84.31 (36.64)	84.39 (918.26)	**83.72** (37.66)	79.72 Timeout	80.84 (**25.07**)
Credit-default	30000	334	80.87 (37.72)	80.69 (847.93)	**80.97** (**20.37**)	80.72 Timeout	79.41 (32.58)
Twitter	49999	1050	95.16 (67.83)	Timeout	**95.56** (98.21)	94.78 Timeout	94.69 (**59.67**)

Table: For every cell in the last seven columns the top value represents the test accuracy (%) on unseen data and the bottom value surrounded by parenthesis represents the average training time (seconds).

Size of interpretable rules of different classifiers

Dataset	RIPPER	MLIC	IMLI
Parkinsons	2.6	**2**	8
Ionosphere	9.6	13	**5**
WDBC	7.6	14.5	**2**
Adult	107.55	44.5	**28**
PIMA	8.25	16	**3.5**
Tom's HW	30.33	**2**	2.5
Twitter	21.6	20.5	**6**
Credit	14.25	6	**3**

Table: Size of the rule of interpretable classifiers.

Rule for WDBC Dataset

Tumor is diagnosed as malignant if
standard area of tumor > 38.43 **OR**
largest perimeter of tumor > 115.9 **OR**
largest number of concave points of tumor > 0.1508

Conclusion

- We propose IMLI: an incremental approach to MaxSAT-based framework for learning interpretable classification rules
- IMLI achieves up to three orders of magnitude runtime improvement without loss of accuracy and interpretability
- The generated rules appear to be reasonable, intuitive, and more interpretable

Thank You !!

MaxSAT

- MaxSAT is an optimization problem of general SAT problem
- Try to maximize the number of satisfied clauses in the formula

MaxSAT

- MaxSAT is an optimization problem of general SAT problem
- Try to maximize the number of satisfied clauses in the formula
- A variant of general MaxSAT is weighted partial MaxSAT
  - Maximize the weight of satisfied clauses
  - Consider two types of clause
    1. Hard clause: weight is infinity, hence always satisfied
    2. Soft clause: priority is set based on positive real valued weight
  - Cost of the solution is the total weight of unsatisfied clauses

Example of MaxSAT

$$\begin{aligned} 1 &: \quad x \\ 2 &: \quad y \\ 3 &: \quad z \\ \infty &: \neg x \vee \neg y \\ \infty &: \quad x \vee \neg z \\ \infty &: \quad y \vee \neg z \end{aligned}$$

Example of MaxSAT

$$\begin{aligned} 1 &: \quad x \\ 2 &: \quad y \\ 3 &: \quad z \\ \infty &: \neg x \vee \neg y \\ \infty &: \quad x \vee \neg z \\ \infty &: \quad y \vee \neg z \end{aligned}$$

$$\begin{aligned} {\color{red}1} &{\color{red}:} \quad {\color{red}x} \\ {\color{blue}2} &{\color{blue}:} \quad {\color{blue}y} \\ {\color{red}3} &{\color{red}:} \quad {\color{red}z} \\ {\color{blue}\infty} &{\color{blue}: \neg x \vee \neg y} \\ {\color{blue}\infty} &{\color{blue}: \quad x \vee \neg z} \\ {\color{blue}\infty} &{\color{blue}: \quad y \vee \neg z} \end{aligned}$$

Example of MaxSAT

$$\begin{aligned} 1 &: \; x \\ 2 &: \; y \\ 3 &: \; z \\ \infty &: \neg x \vee \neg y \\ \infty &: \; x \vee \neg z \\ \infty &: \; y \vee \neg z \end{aligned} \qquad\qquad \begin{aligned} 1 &: \; x \\ 2 &: \; y \\ 3 &: \; z \\ \infty &: \neg x \vee \neg y \\ \infty &: \; x \vee \neg z \\ \infty &: \; y \vee \neg z \end{aligned}$$

Optimal Assignment : $\neg x, y, \neg z$

Cost of the solution is $1 + 3 = 4$

Solution Outline

- Reduce the learning problem as an optimization problem
- Define the objective function
- Define decision variables
- Define constraints
- Choose a proper solver to find the assignment of the decision variables
- Construct the rule

Input Specification

- Discrete optimization problem requires dataset to be in binary
- Categorical and real-valued datasets can be converted to binary by applying standard techniques, e.g., one hot encoding and comparison of feature value with predefined threshold.
- Input instance $\{\mathbf{X}, \mathbf{y}\}$ where $\mathbf{X} \in \{0,1\}^{n \times m}$, and $\mathbf{y} \in \{0,1\}^n$
- $\mathbf{x} = \{x_1, \ldots, x_m\}$ is the boolean feature vector
- Learn a k-clause CNF rule

Objective Function

- Let $|\mathcal{R}|$ = number of literals in the rule
- $\mathcal{E}_{\mathcal{R}}$ = set of samples which are misclassified by $\mathcal{R}$
- λ be data fidelity parameter
- We find a classifier $\mathcal{R}$ as follows:

$$\min_{\mathcal{R}} |\mathcal{R}| + \lambda|\mathcal{E}_{\mathcal{R}}| \text{ such that } \forall \mathbf{X}_i \notin \mathcal{E}_{\mathcal{R}}, y_i = \mathcal{R}(\mathbf{X}_i)$$

- $|\mathcal{R}|$ defines interpretability or sparsity
- $|\mathcal{E}_{\mathcal{R}}|$ defines classification error

Decision Variables

Two types of decision variables-

1. Feature variable b_j^l
   - Feature x_j can participate in each of the l-th clause of CNF rule $\mathcal{R}$
   - If b_j^l is assigned *true*, feature x_j is present in the l-th clause of $\mathcal{R}$
     - Let $\mathcal{R} = (x_1 \vee x_2 \vee x_3) \wedge (x_1 \vee x_4)$
     - For feature x_1, decision variable b_1^1 and b_1^2 are assigned *true*

Decision Variables

Two types of decision variables-

1. Feature variable b_j^l
   - Feature x_j can participate in each of the l-th clause of CNF rule $\mathcal{R}$
   - If b_j^l is assigned *true*, feature x_j is present in the l-th clause of $\mathcal{R}$
     - Let $\mathcal{R} = (x_1 \vee x_2 \vee x_3) \wedge (x_1 \vee x_4)$
     - For feature x_1, decision variable b_1^1 and b_1^2 are assigned *true*
2. Noise variable (classification error) η_q
   - If η_q is assigned *true*, the q-th sample is misclassified by $\mathcal{R}$

MaxSAT Constraints Q_i

- MaxSAT constraint is a CNF formula where each clause has a weight
- Q_i is the MaxSAT constraints for the i-th partition.
- Q_i consists of three set of clauses.

1. Soft Clause for Feature Variable

- IMLI tries to *falsify* each feature variable b_j^l for sparsity

1. Soft Clause for Feature Variable

- IMLI tries to *falsify* each feature variable b_j^l for sparsity
- If a feature variable is assigned *true* in $\mathcal{R}_{i-1}$, IMLI keeps previous assignment

1. Soft Clause for Feature Variable

- IMLI tries to *falsify* each feature variable b_j^l for sparsity
- If a feature variable is assigned *true* in $\mathcal{R}_{i-1}$, IMLI keeps previous assignment

$$V_j^l := \begin{cases} b_j^l & \text{if } x_j \in \mathit{clause}(\mathcal{R}_{i-1}, l) \\ \neg b_j^l & \text{otherwise} \end{cases}; \quad W(V_j^l) = 1$$

Example

$$\mathbf{X}_i = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}; \qquad \mathbf{y}_i = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

- #samples $n = 2$, #features $m = 3$
- We learn a 2-clause rule, i.e. $k = 2$

Let

- $\mathcal{R}_{i-1} = (b_1^1 \vee b_2^1) \wedge (b_1^2)$

Now

$$V_1^1 = (b_1^1); \quad V_2^1 = (b_2^1); \quad V_3^1 = (\neg b_3^1);$$
$$V_1^2 = (b_1^2); \quad V_2^2 = (\neg b_2^2); \quad V_3^2 = (\neg b_3^2);$$

2. Soft Clause for Noise Variable

- IMLI tries to *falsify* as many noise variables as possible
- As data fidelity parameter λ is proportionate to accuracy, IMLI puts λ weight to following soft clause

$$N_q := (\neg \eta_q); \qquad W(N_q) = \lambda$$

Example

$$\mathbf{X}_i = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}; \qquad \mathbf{y}_i = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$N_1 := (\neg \eta_1)$$
$$N_2 := (\neg \eta_2)$$

3. Hard Clause

- Hard clause is always *true*
- If a sample is predicted *correctly*, the *class label is equal to the prediction* of the generated rule and noise variable is assigned *false*
- Otherwise, the noise variable is assigned *true*

3. Hard Clause

- "$\circ$" operator returns the dot product between two vectors
- $\mathbf{u}$ is a vector of constant
- $\mathbf{v}$ is a vector of feature variable
- $\mathbf{u} \circ \mathbf{v} = \bigvee_i (u_i \wedge v_i)$, where u_i and v_i denote a variable/constant at the i-th index of vector $\mathbf{u}$ and $\mathbf{v}$ respectively
- Here "$\wedge$" has standard interpretation, i.e., $a \wedge 1 = a, a \wedge 0 = 0$

3. Hard Clause

- "$\circ$" operator returns the dot product between two vectors
- $\mathbf{u}$ is a vector of constant
- $\mathbf{v}$ is a vector of feature variable
- $\mathbf{u} \circ \mathbf{v} = \bigvee_i (u_i \wedge v_i)$, where u_i and v_i denote a variable/constant at the i-th index of vector $\mathbf{u}$ and $\mathbf{v}$ respectively
- Here "$\wedge$" has standard interpretation, i.e., $a \wedge 1 = a, a \wedge 0 = 0$
- Let $\mathbf{B}_l = \{b_j^l | j \in [1, m]\}$ be the vector of feature variables for the l-th clause

$$D_q := (\neg \eta_q \rightarrow (y_q \leftrightarrow \bigwedge_{l=1}^{k} (\mathbf{X}_q \circ \mathbf{B}_l))); \qquad W(D_q) = \infty$$

Example

$$\mathbf{X}_i = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}; \qquad \mathbf{y}_i = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$D_q := (\neg \eta_q \rightarrow (y_q \leftrightarrow \bigwedge_{l=1}^{k} (\mathbf{X}_q \circ \mathbf{B}_l))); W(D_q) = \infty$$

$$\begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \circ \begin{bmatrix} b_1^1 & b_2^1 & b_3^1 \end{bmatrix} = b_2^1 \vee b_3^1$$

$$\begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \circ \begin{bmatrix} b_1^2 & b_2^2 & b_3^2 \end{bmatrix} = b_2^2 \vee b_3^2$$

$$D_1 := (\neg \eta_1 \rightarrow ((b_2^1 \vee b_3^1) \wedge (b_1^2 \vee b_3^2))$$

$$\begin{bmatrix} 1 & 0 & 1 \end{bmatrix} \circ \begin{bmatrix} b_1^1 & b_2^1 & b_3^1 \end{bmatrix} = b_1^1 \vee b_3^1$$

$$\begin{bmatrix} 1 & 0 & 1 \end{bmatrix} \circ \begin{bmatrix} b_1^2 & b_2^2 & b_3^2 \end{bmatrix} = b_1^2 \vee b_3^2$$

$$D_2 := (\neg \eta_2 \rightarrow (\neg (b_2^1 \vee b_3^1) \vee \neg (b_1^2 \vee b_3^2))$$

MaxSAT constraint Q_i

Q_i is the conjunction of all soft and hard clauses

$$Q_i := V_j^I \wedge N_q \wedge D_q$$

MaxSAT Constraint Q_i

$$
\begin{aligned}
1 &: \quad b_1^1 \\
1 &: \quad b_2^1 \\
1 &: \quad \neg b_3^1 \\
1 &: \quad b_1^2 \\
1 &: \quad \neg b_2^2 \\
1 &: \quad \neg b_3^2 \\
\lambda &: \quad \neg \eta_1 \\
\lambda &: \quad \neg \eta_2 \\
\infty &: \quad \neg \eta_1 \rightarrow ((b_2^1 \vee b_3^1) \wedge (b_2^2 \vee b_3^2)) \\
\infty &: \quad \neg \eta_2 \rightarrow (\neg(b_1^1 \vee b_3^1) \vee \neg(b_1^2 \vee b_3^2))
\end{aligned}
$$

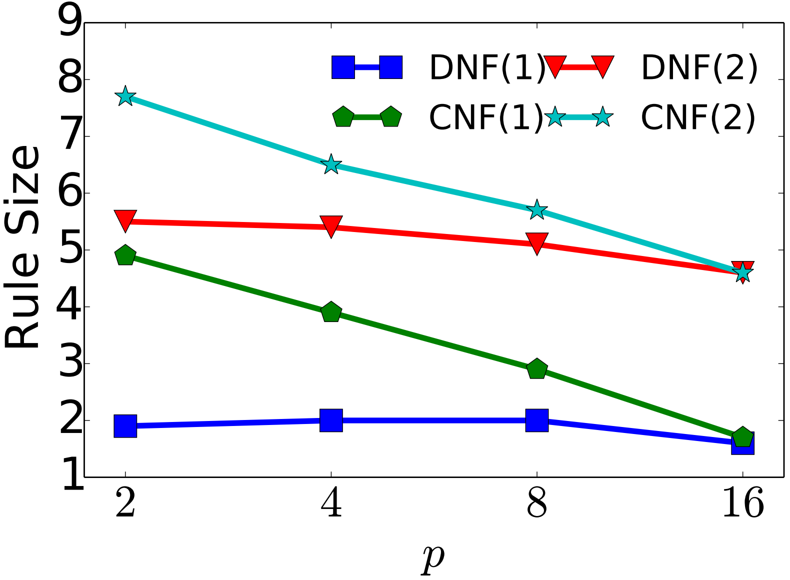
Construction of Rule $\mathcal{R}$

$\mathcal{R}$ consists of features which are assigned *true*

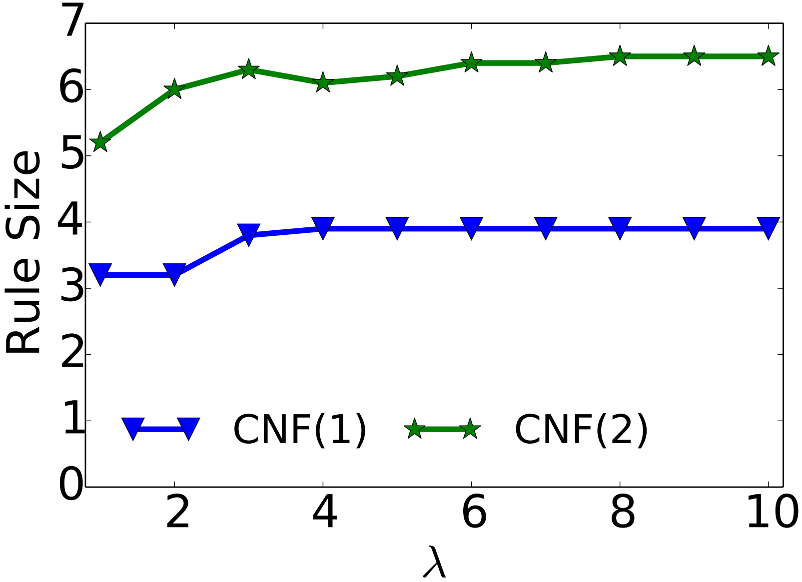
Construction

Let $\sigma^* = \text{MaxSAT}(Q_i, W)$*, then* $x_j \in$ *clause*$(\mathcal{R}_i, l)$ *iff* $\sigma^*(b_j^l) =$ *true.*

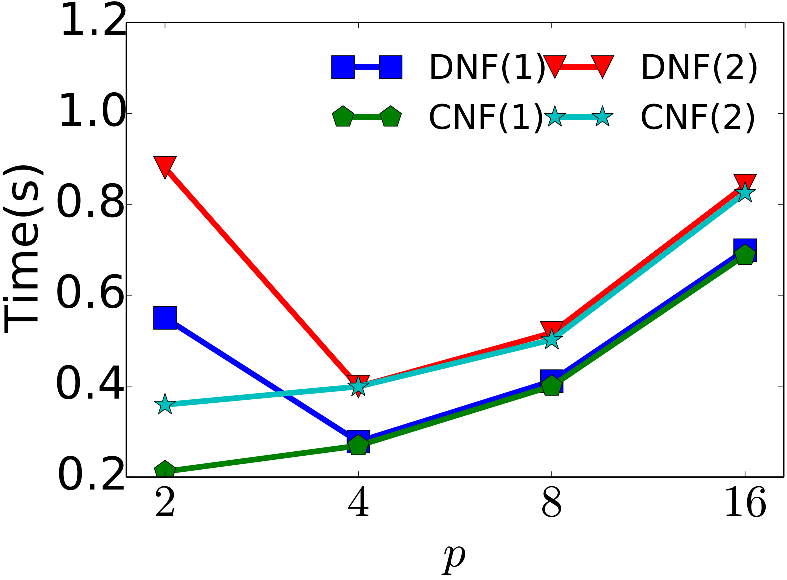
Effect of #partition on rule size



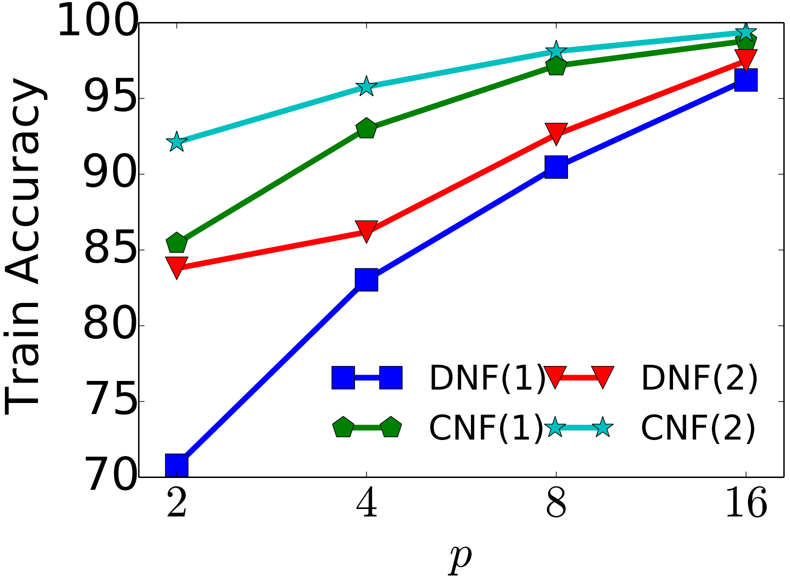
Effect of data fidelity on rule size



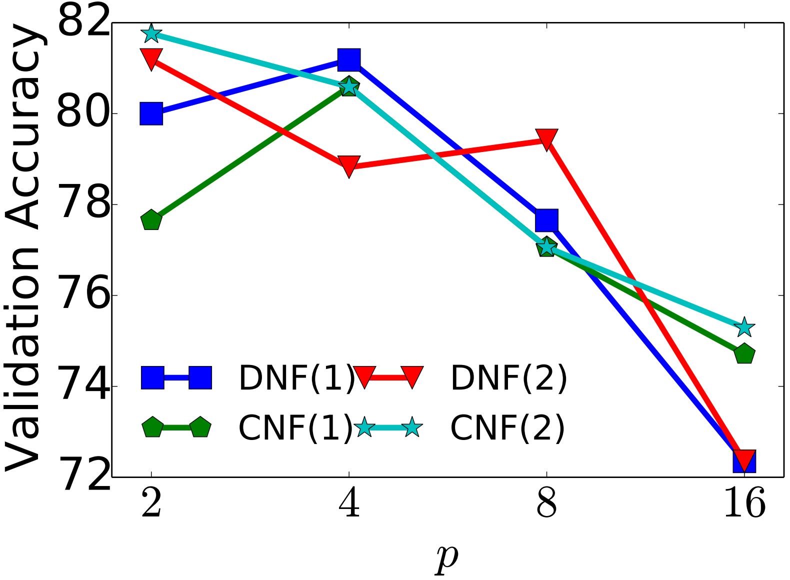
Effect of #partition on training time



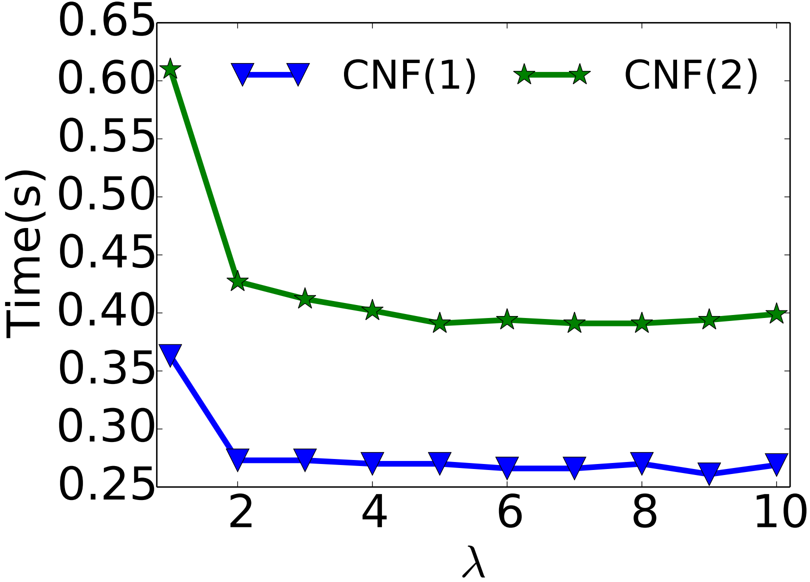
Effect of #partition on training accuracy



Effect of #partition on validation accuracy



Effect of data fidelity on training time



Interpretable Rule: Twitter Dataset

A topic is popular if
Number of Created Discussions at time 1 $>$ 78 OR
Attention Level measured with number of authors at time 6 $>$ 0.000365 OR
Attention Level measured with number of contributions at time 0 $>$ 0.00014 OR
Attention Level measured with number of contributions at time 1 $>$ 0.000136 OR
Number of Authors at time 0 $>$ 147 OR
Average Discussions Length at time 3 $>$ 205.4 OR
Average Discussions Length at time 5 $>$ 654.0

Interpretable Rule: Parkinson's Disease Dataset

A person has Parkinson's disease if
(minimum vocal fundamental frequency $\leq$ 87.57 Hz OR
minimum vocal fundamental frequency $>$ 121.38 Hz OR
Shimmer:APQ3 $\leq$ 0.01 OR
MDVP:APQ $>$ 0.02 OR
D2 $\leq$ 1.93 OR
NHR $>$ 0.01 OR
HNR $>$ 26.5 OR
spread2 $>$ 0.3)
AND
(Maximum vocal fundamental frequency $\leq$ 200.41 Hz OR
HNR $\leq$ 18.8 OR
spread2 $>$ 0.18 OR
D2 $>$ 2.92)

Rule for Pima Indians Diabetes Database

Tested positive for diabetes if
Plasma glucose concentration > 125 AND
Triceps skin fold thickness ≤ 35 mm AND
Diabetes pedigree function > 0.259 AND
Age > 25 years

Rule for Blood Transfusion Service Center Dataset

A person will donate blood if
Months since last donation ≤ 4 AND
total number of donations > 3 AND
total donated blood ≤ 750.0 c.c. AND
months since first donation ≤ 45

Rule for WDBC Dataset

Tumor is diagnosed as malignant if
standard area of tumor > 38.43 OR
largest perimeter of tumor > 115.9 OR
largest number of concave points of tumor > 0.1508