Optimal Randomized Classification Trees

Cristina Molero-Río

Joint work with Rafael Blanquero, Emilio Carrizosa and Dolores Romero Morales.

Fréjus, June 28th 2018

Contents

1. Classic Classification Trees
   - CARTs
2. Optimal Classification Trees
   - ORCTs
3. Current and future research
   - Sparsity on ORCTs at depth 1
   - Sparsity on ORCTs at any depth

Contents

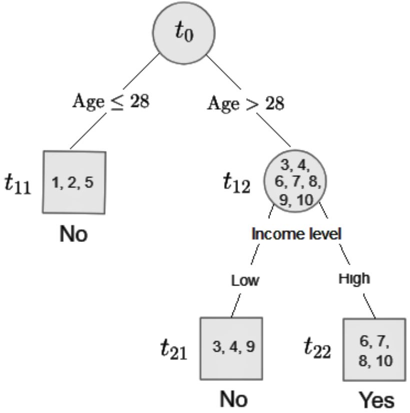
1 Classic Classification Trees
- CARTs

2 Optimal Classification Trees
- ORCTs

3 Current and future research
- Sparsity on ORCTs at depth 1
- Sparsity on ORCTs at any depth

CARTs (Breiman et al. 1984)

Applicant	Age	Income level	Loan granted
1	22	Low	No
2	26	High	No
3	30	Low	Yes
4	32	Low	No
5	20	High	No
6	45	High	Yes
7	60	High	No
8	54	High	Yes
9	50	Low	No
10	48	High	Yes



Motivation

Pros

- They are rule-based and, when they are not very deep, deemed to be easy-to-interpret.
- Low computational times.

Cons

- Classification Trees is a GREEDY procedure, not OPTIMAL.

+ Advances in both computer performance and Mathematical Optimization solvers

Contents

1 Classic Classification Trees
- CARTs

2 Optimal Classification Trees
- ORCTs

3 Current and future research
- Sparsity on ORCTs at depth 1
- Sparsity on ORCTs at any depth

Recent literature

- Integer Programming-based strategies:
  - Bertsimas and Dunn 2017.
  - Günlük et al. 2018.
  - Verwer and Zhang 2017, Verwer et al. 2017.
- It is commonly assumed that training sets are small.
- A CPU time limit is imposed to the solver.

Recent literature

- Integer Programming-based strategies:
  - Bertsimas and Dunn 2017.
  - Günlük et al. 2018.
  - Verwer and Zhang 2017, Verwer et al. 2017.
- It is commonly assumed that training sets are small.
- A CPU time limit is imposed to the solver.

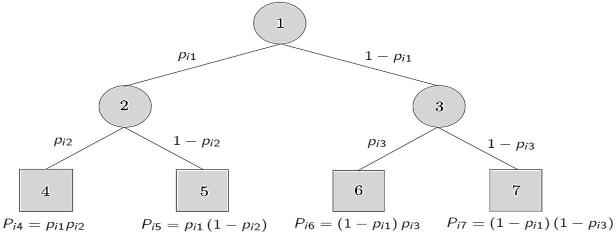
Our proposal: a **continuous** optimization-based method which yields **better results** by performing several local searches in relatively **short time**.

Optimal Randomized Classification Trees

We have a sample $I = \{(\boldsymbol{x}_i, y_i)\}_{1 \leq i \leq n}$, where $\boldsymbol{x}_i \in [0,1]^p$ and $y_i \in \{1, \ldots, K\}$.

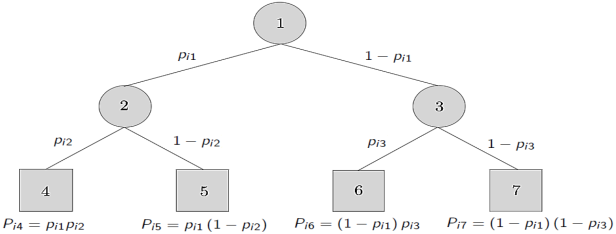
Optimal Randomized Classification Trees

We have a sample $I = \{(\boldsymbol{x}_i, y_i)\}_{1 \leq i \leq n}$, where $\boldsymbol{x}_i \in [0,1]^p$ and $y_i \in \{1, \ldots, K\}$.
A maximal binary tree of depth D. Nodes: Branch $t \in \tau_B$, Leaf $t \in \tau_L$.



Optimal Randomized Classification Trees

We have a sample $I = \{(\boldsymbol{x}_i, y_i)\}_{1\leq i \leq n}$, where $\boldsymbol{x}_i \in [0,1]^p$ and $y_i \in \{1, \ldots, K\}$.
A maximal binary tree of depth D. Nodes: Branch $t \in \tau_B$, Leaf $t \in \tau_L$.



- Orthogonal splits:

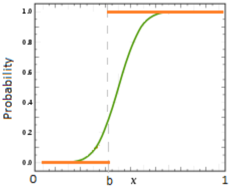
$$a_{jt} = \begin{cases} 1, & \text{variable } j \text{ splits } t \\ 0, & \text{otherwise} \end{cases}, \; j = 1, \ldots, p, \; t \in \tau_B.$$

$$\sum_{j=1}^{p} a_{jt} = 1, \; t \in \tau_B.$$

Optimal Randomized Classification Trees

- Probabilities

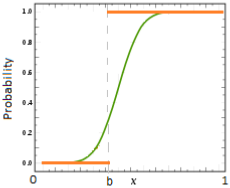
$$\textbf{CDF } F(\cdot;\boldsymbol{\alpha}),\ \boldsymbol{\alpha} \in A.$$



Optimal Randomized Classification Trees

- Probabilities

$$\textbf{CDF}\ F(\cdot;\boldsymbol{\alpha}),\ \boldsymbol{\alpha}\in A.$$

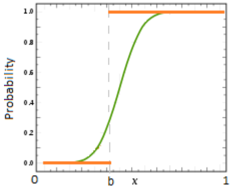


$$p_{it} = F\left(\sum_{j=1}^{p} a_{jt}x_{ij};\boldsymbol{\alpha}_t\right),\ i=1,\ldots,n,\ t\in\tau_B.$$

Optimal Randomized Classification Trees

- Probabilities

CDF $F(\cdot;\boldsymbol{\alpha})$, $\boldsymbol{\alpha} \in A$.



$$p_{it} = F\left(\sum_{j=1}^{p} a_{jt} x_{ij}; \boldsymbol{\alpha}_t\right), \; i = 1, \ldots, n, \; t \in \tau_B.$$

$$P_{it} \equiv \mathbb{P}(\boldsymbol{x}_i \in t) = \prod_{t_l \in N_L(t)} p_{it_l} \prod_{t_r \in N_R(t)} (1 - p_{it_r}), \; i = 1, \ldots, n, \; t \in \tau_L.$$

Optimal Randomized Classification Trees

- Each $t \in \tau_L$ is labeled with one class:

$$C_{kt} = \begin{cases} 1, & \text{node } t \text{ is labeled with class } k \\ 0, & \text{otherwise} \end{cases}, k = 1, \dots, K, \ t \in \tau_L$$

$$\sum_{k=1}^{K} C_{kt} = 1, \ t \in \tau_L.$$

Optimal Randomized Classification Trees

- Each $t \in \tau_L$ is labeled with one class:

$$C_{kt} = \begin{cases} 1, & \text{node } t \text{ is labeled with class } k \\ 0, & \text{otherwise} \end{cases}, k = 1, \dots, K, \ t \in \tau_L$$

$$\sum_{k=1}^{K} C_{kt} = 1, \ t \in \tau_L.$$

- Each class $k = 1, \dots, K$ is identified by, at least, one terminal node:

$$\sum_{t \in \tau_L} C_{kt} \geq 1, \ k = 1, \dots, K.$$

Optimal Randomized Classification Trees

- We now introduce a misclassification cost for classifying an individual from class k in class k':

$$W_{kk'} \geq 0,\ \ k, k' = 1, \ldots, K,\ \ k \neq k'.$$

Optimal Randomized Classification Trees

- We now introduce a misclassification cost for classifying an individual from class k in class k':

$$W_{kk'} \geq 0, \ \ k, k' = 1, \ldots, K, \ \ k \neq k'.$$

- **Objective**

$$\min \quad \sum_{k=1}^{K} \sum_{i \in I_k} \sum_{t \in \tau_L} P_{it} \sum_{k' \neq k} C_{k't} W_{kk'}$$

Optimal Randomized Classification Trees

(Mixed-Integer Non-Linear Optimization Problem)

$$\begin{aligned}
\min \quad & \sum_{k=1}^{K} \sum_{i \in I_k} \sum_{t \in \tau_L} P_{it} \sum_{k' \neq k} C_{k't} W_{kk'} \\
\text{s.t.} \quad & \sum_{j=1}^{p} a_{jt} = 1, \; t \in \tau_B, \\
& \sum_{k=1}^{K} C_{kt} = 1, \; t \in \tau_L, \\
& \sum_{t \in \tau_L} C_{kt} \geq 1, \; k = 1, \ldots, K, \\
& a_{jt} \in \{0,1\}, \; j = 1, \ldots, p, \; t \in \tau_B, \\
& C_{kt} \in \{0,1\}, \; k = 1, \ldots, K, \; t \in \tau_L, \\
& \boldsymbol{\alpha}_t \in A, \; t \in \tau_B.
\end{aligned}$$

Optimal Randomized Classification Trees

(Continuous Non-Linear Optimization Problem) **OBLIQUE** splits

$$\min \quad \sum_{k=1}^{K} \sum_{i \in I_k} \sum_{t \in \tau_L} P_{it} \sum_{k' \neq k} C_{k't} W_{kk'}$$

s.t.

$$\begin{aligned} & \sum_{k=1}^{K} C_{kt} = 1, \; t \in \tau_L, \\ & \sum_{t \in \tau_L} C_{kt} \geq 1, \; k = 1, \dots, K, \\ & a_{jt} \in [-1, 1], \; j = 1, \dots, p, \; t \in \tau_B, \\ & C_{kt} \in [0, 1], \; k = 1, \dots, K, \; t \in \tau_L, \\ & \boldsymbol{\alpha}_t \in A, \; t \in \tau_B. \end{aligned} \qquad \text{(ORCT)}$$

Optimal Randomized Classification Trees

Theorem

There exists an optimal solution to ORCT such that $C_{kt} \in \{0, 1\}$, $k = 1, \ldots, K, t \in \tau_L$.

ORCT's prediction

A new unlabeled observation $\boldsymbol{x}$

↓

Once the optimization problem has been solved


,

the decision variables are used for predicting its class:

$$m_n(\boldsymbol{x}) = \arg\max_k \left\{ \sum_{t \in \tau_L} \mathbb{P}\left(\boldsymbol{x} \in k | \boldsymbol{x} \in t\right) \mathbb{P}\left(\boldsymbol{x} \in t\right) \right\} = \arg\max_k \left\{ \sum_{t \in \tau_L} C_{kt} \cdot P_{\boldsymbol{x}t} \right\}$$

Computational experience

UCI Machine Learning Repository

Data set	n	p	K	Class distribution
Connectionist-bench-sonar	208	60	2	55% - 45%
Wisconsin	569	30	2	63% - 37%
Credit-approval	653	37	2	55% - 45%
Pima-indians-diabetes	768	8	2	65% - 35%
Statlog-project-German-credit	1000	48	2	70% - 30%
Ozone-level-detection-one	1848	72	2	97% - 3%
Spambase	4601	57	2	61% - 39%
Iris	150	4	3	33.3%-33.3%-33.3%
Wine	178	13	3	40%-33%-27%
Seeds	210	7	3	33.3%-33.3%-33.3%
Thyroid-disease-ann-thyroid	3772	21	3	92.5%-5%-2.5%
Car-evaluation	1728	15	4	70%-22%-4%-4%

Computational experience

- Logistic CDF:

$$F(\cdot;\mu,\gamma) = \frac{1}{1+\exp(-(\cdot-\mu)\gamma)},\ \mu \in \mathbb{R},\ \gamma > 0.$$

$\mu_t \in [-1,1],\ t \in \tau_L,\ \gamma_t = \gamma = 512,\ t \in \tau_L.$

Computational experience

- Logistic CDF:

$$F(\cdot;\mu,\gamma) = \frac{1}{1+\exp\left(-\left(\cdot-\mu\right)\gamma\right)},\ \mu \in \mathbb{R},\ \gamma > 0.$$

$\mu_t \in [-1, 1],\ t \in \tau_L,\ \gamma_t = \gamma = 512,\ t \in \tau_L.$

- Equal misclassification weights,

$$W_{kk'} = 0.5,\ k, k' = 1, \ldots, K,\ k \neq k'.$$

Computational experience

- Logistic CDF:

$$F(\cdot;\mu,\gamma)=\frac{1}{1+\exp\left(-\left(\cdot-\mu\right)\gamma\right)},\ \mu\in\mathbb{R},\ \gamma>0.$$

$\mu_t\in[-1,1],\ t\in\tau_L,\ \gamma_t=\gamma=512,\ t\in\tau_L.$

- Equal misclassification weights,

$$W_{kk'}=0.5,\ k,k'=1,\dots,K,\ k\neq k'.$$

- 10 hold-out runs: training subset (75%) and test subset (25%).

Computational experience

- Logistic CDF:

$$F(\cdot;\mu,\gamma) = \frac{1}{1+\exp\left(-\left(\cdot-\mu\right)\gamma\right)},\ \mu \in \mathbb{R},\ \gamma > 0.$$

$\mu_t \in [-1,1],\ t \in \tau_L,\ \gamma_t = \gamma = 512,\ t \in \tau_L.$

- Equal misclassification weights,

$$W_{kk'} = 0.5,\ k,k' = 1,\ldots,K,\ k \neq k'.$$

- 10 hold-out runs: training subset (75%) and test subset (25%).
- Performance measure: average accuracy over the 10 test subsets.

Computational experience

ORCT compared with:

- **CART** (Breiman et al. 1984).
- **OCT-H** (Bertsimas and Dunn 2017).

Computational experience

$D = 1$

Data set	ORCT average time (in secs)	Out-of-sample accuracy		
		ORCT	CART	OCT-H
Connectionist-bench-sonar	22	**76.3**	70.0	70.4
Wisconsin	24	**96.4**	92.0	93.1
Credit-approval	22	83.7	85.7	**87.9**
Pima-indians-diabetes	21	**75.8**	74.2	71.6
Statlog-project-German-credit	28	**72.8**	72.1	71.6
Ozone-level-detection-one	94	96.7	95.6	**96.8**
Spambase	72	**89.8**	89.2	83.6

Computational experience

$D = 1$

Data set	ORCT average time (in secs)	Out-of-sample accuracy		
		ORCT	CART	OCT-H
Connectionist-bench-sonar	22	**76.3**	70.0	70.4
Wisconsin	24	**96.4**	92.0	93.1
Credit-approval	22	83.7	85.7	**87.9**
Pima-indians-diabetes	21	**75.8**	74.2	71.6
Statlog-project-German-credit	28	**72.8**	72.1	71.6
Ozone-level-detection-one	94	96.7	95.6	**96.8**
Spambase	72	**89.8**	89.2	83.6

$D = 2$

Data set	ORCT average time (in secs)	Out-of-sample accuracy		
		ORCT	CART	OCT-H
Iris	17	**95.9**	92.7	95.1
Wine	23	**96.6**	88.6	91.1
Seeds	20	**94.2**	90.2	90.6
Thyroid-disease-ann-thyroid	145	92.2	**99.1**	92.5
Car-evaluation	71	**90.8**	88.1	87.5

Contents

1 Classic Classification Trees
- CARTs

2 Optimal Classification Trees
- ORCTs

3 Current and future research
- Sparsity on ORCTs at depth 1
- Sparsity on ORCTs at any depth

Sparsity on ORCTs at depth 1

$$
\begin{aligned}
\text{min} \quad & \sum_{k=1}^{2} \sum_{i \in I_k} \sum_{t \in \tau_L} P_{it} \sum_{k' \neq k} C_{k't} W_{kk'} \\
\text{s.t.} \quad & C_{12} + C_{22} = 1, \\
& C_{13} + C_{23} = 1, \\
& C_{12} + C_{13} \geq 1, \\
& C_{22} + C_{23} \geq 1, \\
& a_{j1} \in [-1, 1], \; j = 1, \ldots, p, \\
& C_{12}, \; C_{13}, \; C_{22}, \; C_{23} \in [0, 1], \\
& \mu_1 \in [-1, 1].
\end{aligned}
$$

Sparsity on ORCTs at depth 1

A lasso penalization to ORCT

$$
\begin{aligned}
\min \quad & \sum_{k=1}^{2} \sum_{i \in I_k} \sum_{t \in \tau_L} P_{it} \sum_{k' \neq k} C_{k't} W_{kk'} + \lambda \|\boldsymbol{a}_1\|_1 \\
\text{s.t.} \quad & C_{12} + C_{22} = 1, \\
& C_{13} + C_{23} = 1, \\
& C_{12} + C_{13} \geq 1, \\
& C_{22} + C_{23} \geq 1, \\
& a_{j1} \in [-1, 1], \; j = 1, \ldots, p, \\
& C_{12}, \; C_{13}, \; C_{22}, \; C_{23} \in [0, 1], \\
& \mu_1 \in [-1, 1].
\end{aligned}
$$

Sparsity on ORCTs at depth 1

A lasso penalization to ORCT

$$
\begin{aligned}
\min \quad & \sum_{k=1}^{2} \sum_{i \in I_k} \sum_{t \in \tau_L} P_{it} \sum_{k' \neq k} C_{k' t} W_{kk'} + \lambda \sum_{j=1}^{p} |a_{j1}| \\
\text{s.t.} \quad & C_{12} + C_{22} = 1, \\
& C_{13} + C_{23} = 1, \\
& C_{12} + C_{13} \geq 1, \\
& C_{22} + C_{23} \geq 1, \\
& a_{j1} \in [-1, 1], \; j = 1, \ldots, p, \\
& C_{12}, \; C_{13}, \; C_{22}, \; C_{23} \in [0, 1], \\
& \mu_1 \in [-1, 1].
\end{aligned}
$$

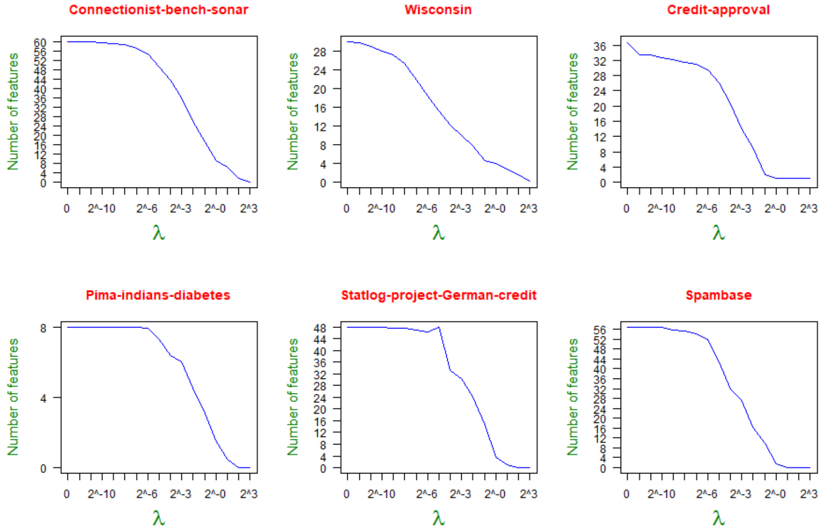
Sparsity on ORCTs at depth 1

A lasso penalization to ORCT

$$a_{j1} = a_{j1}^{+} - a_{j1}^{-}$$

$$\begin{aligned}
\min \quad & \sum_{k=1}^{2} \sum_{i \in I_k} \sum_{t \in \tau_L} P_{it} \sum_{k' \neq k} C_{k't} W_{kk'} + \lambda \sum_{j=1}^{p} \left(a_{j1}^{+} + a_{j1}^{-} \right) \\
\text{s.t.} \quad & C_{12} + C_{22} = 1, \\
& C_{13} + C_{23} = 1, \\
& C_{12} + C_{13} \geq 1, \\
& C_{22} + C_{23} \geq 1, \\
& a_{j1}^{+},\ a_{j1}^{-} \in [0,1],\ j = 1, \ldots, p, \\
& C_{12},\ C_{13},\ C_{22},\ C_{23} \in [0,1]\,, \\
& \mu_1 \in [-1,1]\,.
\end{aligned}$$

Sparsity on ORCTs at depth 1



Sparsity on ORCTs at depth 1

Theorem

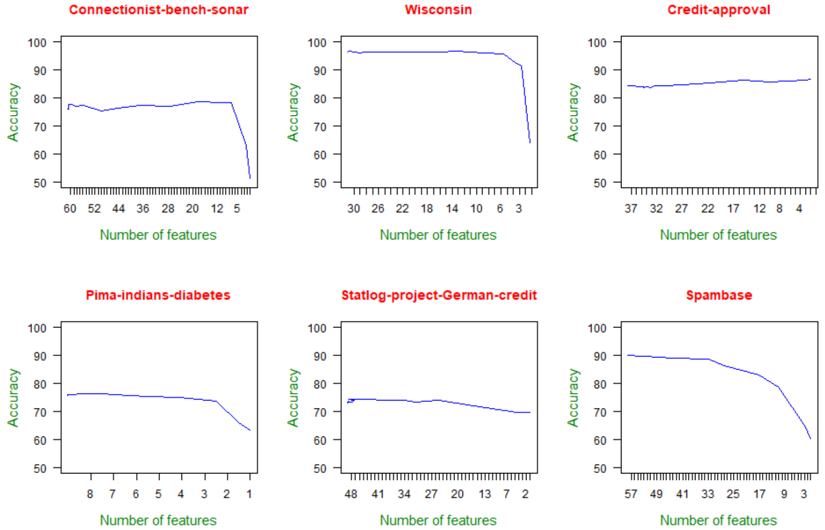
Let $F \in \mathcal{C}^1$ a CDF with f as its corresponding PDF. **There exists a minimum λ from which $a_1 = 0$ is an optimal solution to the lasso penalization of ORCT at depth 1**:

$$\lambda = \max\left\{\lambda_{\mu_1=-1}, \lambda_{\mu_1=1}\right\},$$

where

$$\lambda_{\mu_1} = \frac{1}{p} f\left(-\frac{\mu_1}{p}\right) \max_{j=1,\dots,p} \left| -W_{21} \sum_{i \in I_2} x_{ij} + W_{12} \sum_{i \in I_1} x_{ij} \right|.$$

Sparsity on ORCTs at depth 1



Sparsity on ORCTs at any depth

CURRENT RESEARCH

Sparsity on ORCTs at any depth

CURRENT RESEARCH

A **Sparse oblique cuts**. A generalization of the previous model.

$$
\begin{aligned}
\min \quad & \sum_{k=1}^{K} \sum_{i \in I_k} \sum_{t \in \tau_L} P_{it} \sum_{k' \neq k} C_{k't} W_{kk'} + \lambda \sum_{t \in \tau_L} \|\boldsymbol{a}_t\|_1 \\
\text{s.t.} \quad & \sum_{k=1}^{K} C_{kt} = 1, \ t \in \tau_L, \\
& \sum_{t \in \tau_L} C_{kt} \geq 1, \ k = 1, \dots, K, \\
& a_{jt} \in [-1, 1], \ j = 1, \dots, p, \ t \in \tau_B, \\
& C_{kt} \in [0, 1], \ k = 1, \dots, K, \ t \in \tau_L, \\
& \mu_t \in [-1, 1], \ t \in \tau_B.
\end{aligned}
$$

Sparsity on ORCTs at any depth

CURRENT RESEARCH

A **Sparse oblique cuts**. A generalization of the previous model.

$$\min \quad \sum_{k=1}^{K} \sum_{i \in I_k} \sum_{t \in \tau_L} P_{it} \sum_{k' \neq k} C_{k't} W_{kk'} + \lambda \sum_{t \in \tau_L} \|\boldsymbol{a}_t\|_1$$

$$\text{s.t.} \quad \sum_{k=1}^{K} C_{kt} = 1, \ t \in \tau_L,$$

$$\sum_{t \in \tau_L} C_{kt} \geq 1, \ k = 1, \dots, K,$$

$$a_{jt} \in [-1, 1], \ j = 1, \dots, p, \ t \in \tau_B,$$

$$C_{kt} \in [0, 1], \ k = 1, \dots, K, \ t \in \tau_L,$$

$$\mu_t \in [-1, 1], \ t \in \tau_B.$$

B **Sparse ORCT**.

Bibliography

Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen.
Classification and regression trees.
CRC press, 1984.

Dimitris Bertsimas and Jack Dunn.
Optimal classification trees.
Machine Learning, 106(7):1039–1082, 2017.

Oktay Günlük, Jayant Kalagnanam, Matt Menickelly, and Katya Scheinberg.
Optimal Decision Trees for Categorical Data via Integer Programming.
arXiv preprint arXiv:1612.03225v2 2018.

Sicco Verwer and Yingqian Zhang.
Learning decision trees with flexible constraints and objectives using integer optimization.
In *International Conference on AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems*, pages 94–103. Springer, 2017.

Sicco Verwer, Yingqian Zhang, and Qing Chuan Ye.
Auction optimization using regression trees and linear models as integer programs.
Artificial Intelligence, 244:368–395, 2017.

Thank you for your attention!

mmolero@us.es