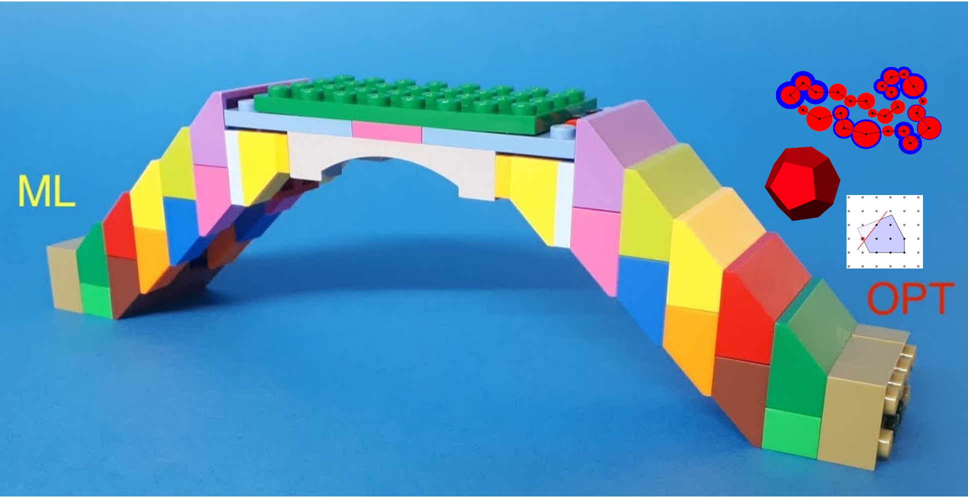
A polyhedral study of Multivariate Decision Trees

Carla Michini Zachary Zhou
University of Wisconsin-Madison

MIP 2024
University of Kentucky, June 4, 2024


ML
OPT

Decision trees for classification

To Get Results Advertise Your Wants in the Daily Globe

The Boston Daily Globe.

Advertise Your April 19th Sales in The Daily Globe

EVENING EDITION—7:30 O'CLOCK

ALL DROWNED BUT 868

About 1232 Lost Lives in the Titantic's Plunge, Greatest Sea Disaster for Years.

EXCITING EVENTS BEFORE TITANIC'S FINAL PLUNGE

Virginian and Parisian Found None Alive.

Women and Children Safe But Few Notable Men.

Carpathia Has Survivors—On Way to New York.

Only Partial List of Them Received Owing to Interruption.

BAY STATE PEOPLE SAVED

CHARLES M. HAYS SAVED.

President of Grand Trunk Railroad One of Few Men Rescued.

THROUGH FIELD OF ICE.

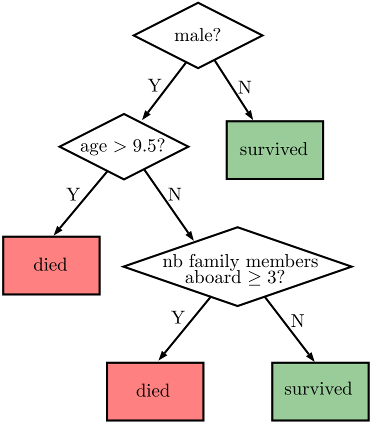
Wireless Says Carpathia is Going Slowly Toward New York.

TITANIC'S PASSENGER LISTS.

SOME WHO MAY BE SAVED

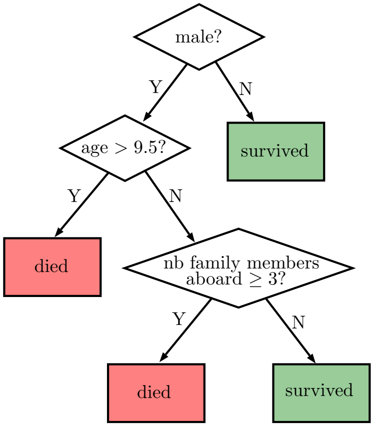
NEARLY ALL MEN LOST.

All the Boats Launched Have Been Accounted For by Carpathia.

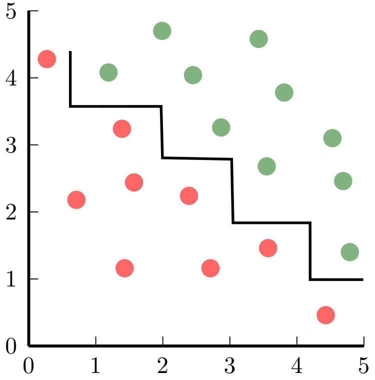


Decision trees for classification

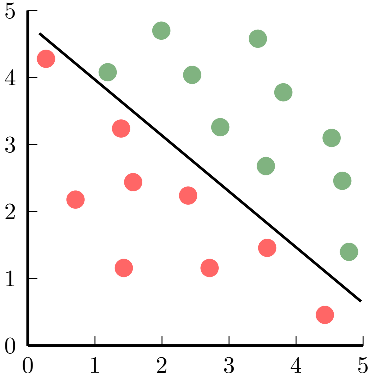
- Binary Tree
- Highly Interpretable
- Branching nodes $\mathcal{B}$ and leaf nodes $\mathcal{L}$
- At each branching node a branching rule
- At each leaf node a class label
- Tree depth D



Univariate vs multivariate branching rules



Univariate branching rules



Multivariate branching rules

Previous approaches to compute optimal decision trees

- Heuristics: CART [Breiman et. al, 1984], ID3 [Quinlan, 1986]

Previous approaches to compute optimal decision trees

- Heuristics: CART [Breiman et. al, 1984], ID3 [Quinlan, 1986]
- Dynamic Programming: Nijssen & Fromon (2007), Aglin et al. (2020), Hu et al. (2020), Lin et al. (2020). Demirović et al. (2021).

Previous approaches to compute optimal decision trees

- Heuristics: CART [Breiman et. al, 1984], ID3 [Quinlan, 1986]
- Dynamic Programming: Nijssen & Fromon (2007), Aglin et al. (2020), Hu et al. (2020), Lin et al. (2020). Demirović et al. (2021).
- SAT and Constraint Programming: Narodytska et al. (2018), Avellaneda (2020), Janota and Morgado (2020), Verhaeghe et al. (2020), Schidler & Szeider (2021).

Previous approaches to compute optimal decision trees

- Heuristics: CART [Breiman et. al, 1984], ID3 [Quinlan, 1986]
- Dynamic Programming: Nijssen & Fromon (2007), Aglin et al. (2020), Hu et al. (2020), Lin et al. (2020). Demirović et al. (2021).
- SAT and Constraint Programming: Narodytska et al. (2018), Avellaneda (2020), Janota and Morgado (2020), Verhaeghe et al. (2020), Schidler & Szeider (2021).
- MIP: Bertsimas & Dunn (2017), Aghaei et al. (2019, 2021), Günlük et al. (2019), Verwer & Zhan (2019), Dash et al. (2020), Zhu et al. (2020), Lawless & Günlük 2021, Boutilier et al.(2022, 2023).

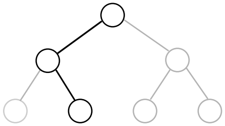
Previous approaches to compute optimal decision trees

- MIP: Bertsimas & Dunn (2017), Aghaei et al. (2019, 2021), Günlük et al. (2019), Verwer & Zhan (2019), Dash et al. (2020), Zhu et al. (2020), Lawless & Günlük 2021, Boutilier et al.(2022, 2023).

Previous approaches to compute optimal decision trees

- MIP: Bertsimas & Dunn (2017), Aghaei et al. (2019, 2021), Günlük et al. (2019), Verwer & Zhan (2019), Dash et al. (2020), Zhu et al. (2020), Lawless & Günlük 2021, Boutilier et al.(2022, 2023).

Two main ingredients:

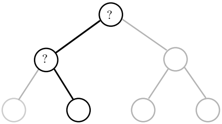


1. How to route
2. How to split

Previous approaches to compute optimal decision trees

- MIP: Bertsimas & Dunn (2017), Aghaei et al. (2019, 2021), Günlük et al. (2019), Verwer & Zhan (2019), Dash et al. (2020), Zhu et al. (2020), Lawless & Günlük 2021, Boutilier et al.(2022, 2023).

Two main ingredients:

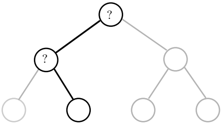


1. How to route
2. How to split

Previous approaches to compute optimal decision trees

- MIP: Bertsimas & Dunn (2017), Aghaei et al. (2019, 2021), Günlük et al. (2019), Verwer & Zhan (2019), Dash et al. (2020), Zhu et al. (2020), Lawless & Günlük 2021, Boutilier et al.(2022, 2023).

Two main ingredients:



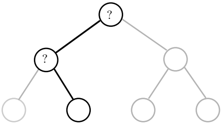
1. How to route
2. How to split

Big-M

Previous approaches to compute optimal decision trees

- MIP: Bertsimas & Dunn (2017), Aghaei et al. (2019, 2021), Günlük et al. (2019), Verwer & Zhan (2019), Dash et al. (2020), Zhu et al. (2020), Lawless & Günlük 2021, Boutilier et al.(2022, 2023).

Two main ingredients:

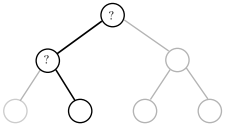


1. How to route
2. How to split

Previous approaches to compute optimal decision trees

- MIP: Bertsimas & Dunn (2017), Aghaei et al. (2019, 2021), Günlük et al. (2019), Verwer & Zhan (2019), Dash et al. (2020), Zhu et al. (2020), Lawless & Günlük 2021, Boutilier et al.(2022, 2023).

Two main ingredients:



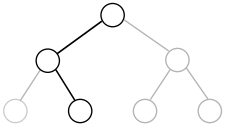
1. How to route
2. How to split

GOAL: stronger formulation, polyhedral study

Previous approaches to compute optimal decision trees

- MIP: Bertsimas & Dunn (2017), Aghaei et al. (2019, 2021), Günlük et al. (2019), Verwer & Zhan (2019), Dash et al. (2020), Zhu et al. (2020), Lawless & Günlük 2021, Boutilier et al.(2022, 2023).

Two main ingredients:



1. How to route
2. How to split

Big-M

GOAL: stronger formulation, polyhedral study

Problem setting

INPUT:

- N (distinct) datapoints, K classes
- Training set $(x^i, y^i) \in [0,1]^p \times [K]$, $i \in [N]$
- Max tree depth D

OUTPUT:

- Multivariate decision tree maximizing training accuracy

DECISIONS:

- $\forall t \in \mathcal{B}$: branching rule?
- $\forall t \in \mathcal{L}$: class label?
- $\forall i \in [N]$: route of x^i?
- $\forall i \in [N]$: is x^i correctly classified?

Problem setting

INPUT:

- N (distinct) datapoints, K classes
- Training set $(x^i, y^i) \in [0,1]^p \times [K]$, $i \in [N]$
- Max tree depth D

OUTPUT:

- Multivariate decision tree maximizing training accuracy

DECISIONS:

- $\forall t \in \mathcal{B}$: branching rule?$(a_t, b_t) \in \mathbb{R}^{p+1}$
- $\forall t \in \mathcal{L}$: class label?
- $\forall i \in [N]$: route of x^i?
- $\forall i \in [N]$: is x^i correctly classified?

Problem setting

INPUT:

- N (distinct) datapoints, K classes
- Training set $(x^i, y^i) \in [0,1]^p \times [K]$, $i \in [N]$
- Max tree depth D

OUTPUT:

- Multivariate decision tree maximizing training accuracy

DECISIONS:

- $\forall t \in \mathcal{B}$: branching rule? $(a_t, b_t) \in \mathbb{R}^{p+1}$
- $\forall t \in \mathcal{L}$: class label? $c_{kt} \in \{0,1\} \quad k \in [K], t \in \mathcal{L}$
- $\forall i \in [N]$: route of x^i?
- $\forall i \in [N]$: is x^i correctly classified?

Problem setting

INPUT:

- N (distinct) datapoints, K classes
- Training set $(x^i, y^i) \in [0,1]^p \times [K]$, $i \in [N]$
- Max tree depth D

OUTPUT:

- Multivariate decision tree maximizing training accuracy

DECISIONS:

- $\forall t \in \mathcal{B}$: branching rule?$(a_t, b_t) \in \mathbb{R}^{p+1}$
- $\forall t \in \mathcal{L}$: class label? $c_{kt} \in \{0,1\} \quad k \in [K], t \in \mathcal{L}$
- $\forall i \in [N]$: route of x^i?$w_{it} \in \{0,1\} \quad i \in [N], t \in \mathcal{B} \cup \mathcal{L}$
- $\forall i \in [N]$: is x^i correctly classified?

Problem setting

INPUT:

- N (distinct) datapoints, K classes
- Training set $(x^i, y^i) \in [0,1]^p \times [K]$, $i \in [N]$
- Max tree depth D

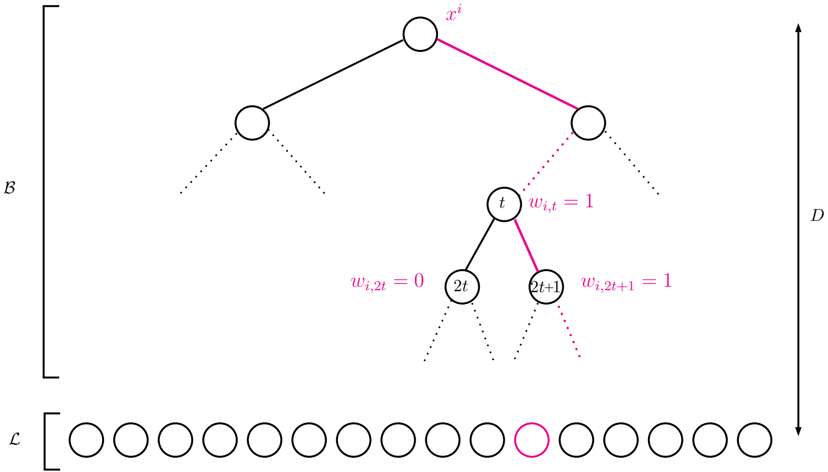
OUTPUT:

- Multivariate decision tree maximizing training accuracy

DECISIONS:

- $\forall t \in \mathcal{B}$: branching rule? $(a_t, b_t) \in \mathbb{R}^{p+1}$
- $\forall t \in \mathcal{L}$: class label? $c_{kt} \in \{0,1\} \quad k \in [K], t \in \mathcal{L}$
- $\forall i \in [N]$: route of x^i? $w_{it} \in \{0,1\} \quad i \in [N], t \in \mathcal{B} \cup \mathcal{L}$
- $\forall i \in [N]$: is x^i correctly classified? $z_{it} \in \{0,1\} \quad i \in [N], t \in \mathcal{L}$

Binary routings

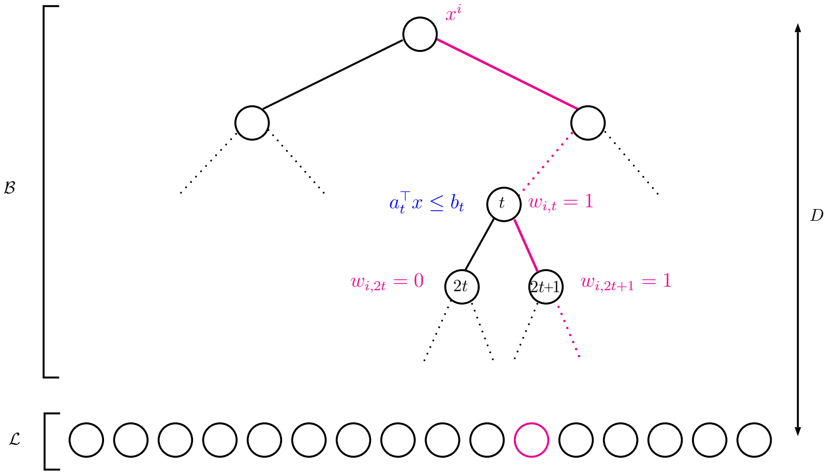


Routing constraints:

$$\sum_{t \in \mathcal{L}} w_{it} = 1 \qquad \forall i \in [N],$$

$$w_{it} = w_{i,2t} + w_{i,2t+1} \qquad \forall i \in [N],\ t \in \mathcal{B},$$

Realizable routings



$$\{(a_t, b_t) \in \mathbb{R}^{p+1} : a_t^\top x^i \le b_t - 1 \quad \forall\, i \in [N] : w_{i,2t} = 1,$$
$$a_t^\top x^i \ge b_t + 1 \quad \forall\, i \in [N] : w_{i,2t+1} = 1\} \neq \emptyset.$$

[∃ mv branching rules realizing the routings]

Realizable routings

For a tree of depth D, let R_D be the set of realizable routings, and define $W_D = \text{conv}(R_D)$.

GOAL: Polyhedral characterization of W_D?

QUESTIONS:

1. Facets of W_1?
2. From W_1 to W_D?
3. From W_D to a polyhedral description of multivariate decision trees.

Problem formulation

Let P_D be a polyhedron such that $R_D = P_D \cap \{0,1\}^{[N]\times(\mathcal{B}\cup\mathcal{L})}$.

$$
\begin{aligned}
\underset{w,c,z}{\text{maximize}} \quad & \sum_{i\in[N]} \sum_{t\in\mathcal{L}} z_{it} \\
\text{subject to} \quad & w \in P_D \cap \{0,1\}^{[N]\times(\mathcal{B}\cup\mathcal{L})} \\
& \sum_{k\in[K]} c_{tk} = 1 && \forall t \in \mathcal{L}, \\
& z_{it} \le \min\{w_{it}, c_{t,y^i}\} && \forall i \in [N],\ t \in \mathcal{L}, \\
& c_{tk} \in \{0,1\} && \forall t \in \mathcal{L},\ k \in [K], \\
& z_{it} \in \{0,1\} && \forall i \in [N],\ t \in \mathcal{L}.
\end{aligned}
$$

Let S be the feasible set of the above problem.

Problem formulation

Let P_D be a polyhedron such that $R_D = P_D \cap \{0,1\}^{[N]\times(\mathcal{B}\cup\mathcal{L})}$.

$$
\begin{array}{lll}
\underset{w,c,z}{\text{maximize}} & \displaystyle\sum_{i\in[N]}\sum_{t\in\mathcal{L}} z_{it} & \\
\text{subject to} & w \in P_D \cap \{0,1\}^{[N]\times(\mathcal{B}\cup\mathcal{L})} & \\
& \displaystyle\sum_{k\in[K]} c_{tk} = 1 & \forall t \in \mathcal{L}, \\
& z_{it} \le \min\{w_{it}, c_{t,y^i}\} & \forall i \in [N],\ t \in \mathcal{L}, \\
& \cancel{c_{tk} \in \{0,1\}}\ c_{tk} \ge 0 & \forall t \in \mathcal{L},\ k \in [K], \\
& \cancel{z_{it} \in \{0,1\}}\ z_{it} \ge 0 & \forall i \in [N],\ t \in \mathcal{L}.
\end{array}
$$

Let $\cancel{S}$ S' be the feasible set of the above problem.

Problem formulation

Let P_D be a polyhedron such that $R_D = P_D \cap \{0,1\}^{[N]\times(\mathcal{B}\cup\mathcal{L})}$.

$$
\begin{aligned}
\underset{w,c,z}{\text{maximize}} \quad & \sum_{i\in[N]} \sum_{t\in\mathcal{L}} z_{it} \\
\text{subject to} \quad & w \in P_D \cap \{0,1\}^{[N]\times(\mathcal{B}\cup\mathcal{L})} \\
& \sum_{k\in[K]} c_{tk} = 1 && \forall t \in \mathcal{L}, \\
& z_{it} \le \min\{w_{it}, c_{t,y^i}\} && \forall i \in [N],\ t \in \mathcal{L}, \\
& \cancel{c_{tk} \in \{0,1\}}\ c_{tk} \ge 0 && \forall t \in \mathcal{L},\ k \in [K], \\
& \cancel{z_{it} \in \{0,1\}}\ z_{it} \ge 0 && \forall i \in [N],\ t \in \mathcal{L}.
\end{aligned}
$$

Let $\cancel{S}$ S' be the feasible set of the above problem.

Lemma. $\text{conv}(S) = \text{conv}(S')$.

Possible choices for P_D: baseline formulation

Define P_D as the projection on the w variables of the (w, a, b) s.t.:

$$
\begin{aligned}
&\|a_t\|_1 \leq 1, b_t \leq 1 && t \in \mathcal{B} \\
&a_t^\top x^i \leq b_t + M_i(1 - w_{i,2t}) && i \in [N], t \in \mathcal{B} \\
&a_t^\top x^i - \varepsilon \geq b_t - (M_i + \varepsilon)(1 - w_{i,2t+1}) && i \in [N], t \in \mathcal{B}
\end{aligned}
$$

[w satisfies the routing constraints]

where ε is a small positive constant and the big-M values are

$$M_i = \max_{j \in [p]} \{x_j^i\} + 1 \quad \forall i \in [N].$$

Similar to Bertsimas and Dunn (2017), but tighter LP relaxation (Boutilier, M. & Zhou, 2023)

Possible choices for P_D: shattering inequalities

Let $\mathcal{I}$ be the set of pairs $(I_L, I_R) \in [N]^2$ such that:

1. $I_L \cap I_R = \emptyset$ are disjoint
2. $\{x^i\}_{i \in I_L}$ and $\{x^i\}_{i \in I_R}$ are NOT linearly separable
3. $\forall j \in I_L \cup I_R$, $\{x^i\}_{i \in I_L} \setminus \{j\}$ and $\{x^i\}_{i \in I_R} \setminus \{j\}$ are linearly separable

Possible choices for P_D: shattering inequalities

Let $\mathcal{I}$ be the set of pairs $(I_L, I_R) \in [N]^2$ such that:

1. $I_L \cap I_R = \emptyset$ are disjoint
2. $\{x^i\}_{i \in I_L}$ and $\{x^i\}_{i \in I_R}$ are NOT linearly separable
3. $\forall j \in I_L \cup I_R$, $\{x^i\}_{i \in I_L} \setminus \{j\}$ and $\{x^i\}_{i \in I_R} \setminus \{j\}$ are linearly separable

Possible choices for P_D: shattering inequalities

Let $\mathcal{I}$ be the set of pairs $(I_L, I_R) \in [N]^2$ such that:

1. $I_L \cap I_R = \emptyset$ are disjoint
2. $\{x^i\}_{i \in I_L}$ and $\{x^i\}_{i \in I_R}$ are NOT linearly separable
3. $\forall j \in I_L \cup I_R$, $\{x^i\}_{i \in I_L} \setminus \{j\}$ and $\{x^i\}_{i \in I_R} \setminus \{j\}$ are linearly separable

Shattering inequalities [Boutilier, M. & Zhou, 2022] at node $t \in \mathcal{B}$:

$$\sum_{i \in I_L} w_{i,2t} + \sum_{i \in I_R} w_{i,2t+1} \leq |I_L| + |I_R| - 1, \quad \forall (I_L, I_R) \in \mathcal{I},\ t \in \mathcal{B}.$$

Possible choices for P_D: shattering inequalities

Let $\mathcal{I}$ be the set of pairs $(I_L, I_R) \in [N]^2$ such that:

1. $I_L \cap I_R = \emptyset$ are disjoint
2. $\{x^i\}_{i \in I_L}$ and $\{x^i\}_{i \in I_R}$ are NOT linearly separable
3. $\forall j \in I_L \cup I_R$, $\{x^i\}_{i \in I_L} \setminus \{j\}$ and $\{x^i\}_{i \in I_R} \setminus \{j\}$ are linearly separable

Define P_D as the w vectors s.t.: [w satisfies the routing constraints]
[w satisfies all shattering inequalities]

The binary vectors in P_D are the realizable routings.

Shattering inequalities [Boutilier, M. & Zhou, 2022] at node $t \in \mathcal{B}$:

$$\sum_{i \in I_L} w_{i,2t} + \sum_{i \in I_R} w_{i,2t+1} \leq |I_L| + |I_R| - 1, \quad \forall (I_L, I_R) \in \mathcal{I},\ t \in \mathcal{B}.$$

Our contributions

QUESTIONS:

1. Facets of W_1?

2. From W_1 to W_D?

3. From W_D to a polyhedral description of $\text{conv}(S)$?

Our contributions

QUESTIONS:

1. Facets of W_1?

   Main result 1. If the dataset is in general position, then the shattering inequalities are facets of W_1.

2. From W_1 to W_D?

3. From W_D to a polyhedral description of $\text{conv}(S)$?

Our contributions

QUESTIONS:

1. Facets of W_1?

   Main result 1. If the dataset is in general position, then the shattering inequalities are facets of W_1.

2. From W_1 to W_D?

   Main result 2. A facet of W_1 involving at least two variables is also a facet of W_D, for $D \geq 2$.

3. From W_D to a polyhedral description of $\mathrm{conv}(S)$?

Our contributions

QUESTIONS:

1. Facets of W_1?

   Main result 1. If the dataset is in general position, then the shattering inequalities are facets of W_1.

2. From W_1 to W_D?

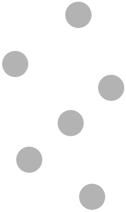
   Main result 2. A facet of W_1 involving at least two variables is also a facet of W_D, for $D \geq 2$.

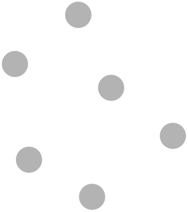
3. From W_D to a polyhedral description of $\text{conv}(S)$?

   Main result 3. A facet of W_D is a facet of $\text{conv}(S)$ iff it is not contained in $\{w : w_{it} = 0\}$, $i \in [N]$, $t \in \mathcal{L}$.

The general position assumption

- A finite set of points in $\mathbb{R}^p$ are in general position if no n points lie in an $(n-2)$-dimensional affine subspace for $n = 2, \ldots, p+1$.





- The dataset is in general position if $x^1, \ldots, x^N$ are in general position.
- If the dataset is in general position each shattering inequality has $p+2$ nonzero coefficients (related to VC dimension of linear classifiers [Vapnik, 1998]).

The general position assumption

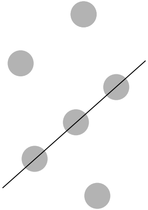
- A finite set of points in $\mathbb{R}^p$ are in general position if no n points lie in an $(n-2)$-dimensional affine subspace for $n = 2, \ldots, p+1$.

NOT in general position

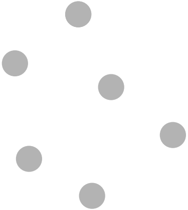
- The dataset is in general position if $x^1, \ldots, x^N$ are in general position.
- If the dataset is in general position each shattering inequality has $p+2$ nonzero coefficients (related to VC dimension of linear classifiers [Vapnik, 1998]).

The general position assumption

- A finite set of points in $\mathbb{R}^p$ are in general position if no n points lie in an $(n-2)$-dimensional affine subspace for $n = 2, \ldots, p+1$.



NOT in general position



in general position

- The dataset is in general position if $x^1, \ldots, x^N$ are in general position.
- If the dataset is in general position each shattering inequality has $p+2$ nonzero coefficients (related to VC dimension of linear classifiers [Vapnik, 1998]).

The general position assumption

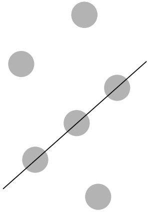
- A finite set of points in $\mathbb{R}^p$ are in general position if no n points lie in an $(n-2)$-dimensional affine subspace for $n = 2, \ldots, p+1$.

NOT in general position — in general position

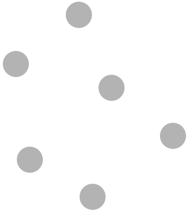
- The dataset is in general position if $x^1, \ldots, x^N$ are in general position.
- If the dataset is in general position each shattering inequality has $p+2$ nonzero coefficients (related to VC dimension of linear classifiers [Vapnik, 1998]).

The general position assumption

- A finite set of points in $\mathbb{R}^p$ are in general position if no n points lie in an $(n-2)$-dimensional affine subspace for $n = 2, \ldots, p+1$.



NOT in general position

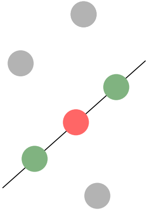


in general position

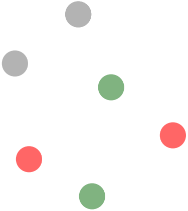
- The dataset is in general position if $x^1, \ldots, x^N$ are in general position.
- If the dataset is in general position each shattering inequality has $p+2$ nonzero coefficients (related to VC dimension of linear classifiers [Vapnik, 1998]).

The general position assumption

- A finite set of points in $\mathbb{R}^p$ are in general position if no n points lie in an $(n-2)$-dimensional affine subspace for $n = 2, \ldots, p+1$.



NOT in general position



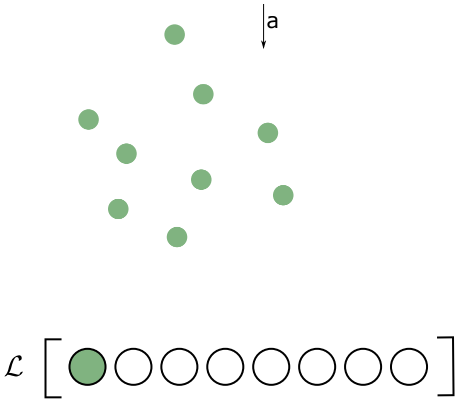
in general position

- The dataset is in general position if $x^1, \ldots, x^N$ are in general position.
- If the dataset is in general position each shattering inequality has $p+2$ nonzero coefficients (related to VC dimension of linear classifiers [Vapnik, 1998]).

Dimension of W_D

Theorem. $\dim(W_D) = N(|\mathcal{L}| - 1)$.

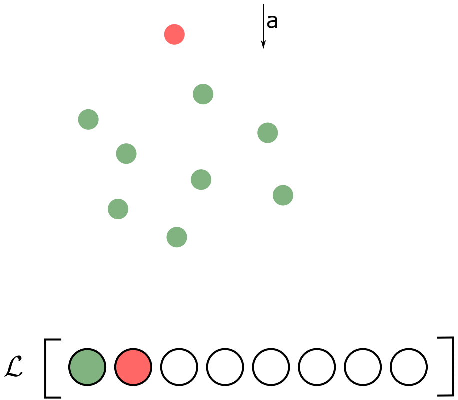
Proof sketch. $\exists a \in \mathbb{R}^p$ s.t. $a^\top x^i \neq a^\top x^j$ for all distinct $i, j \in [N]$. Suppose wlog that $a^\top x^i < a^\top x^{i+1} \Rightarrow \forall i \in [N-1]$ we can linearly separate the first i datapoints from the rest.



Dimension of W_D

Theorem. $\dim(W_D) = N(|\mathcal{L}| - 1)$.

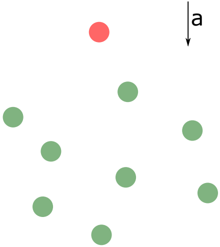
Proof sketch. $\exists a \in \mathbb{R}^p$ s.t. $a^\top x^i \neq a^\top x^j$ for all distinct $i, j \in [N]$. Suppose wlog that $a^\top x^i < a^\top x^{i+1} \Rightarrow \forall i \in [N-1]$ we can linearly separate the first i datapoints from the rest.

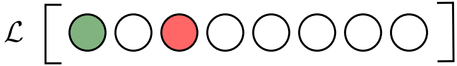


Dimension of W_D

Theorem. $\dim(W_D) = N(|\mathcal{L}| - 1)$.

Proof sketch. $\exists a \in \mathbb{R}^p$ s.t. $a^\top x^i \neq a^\top x^j$ for all distinct $i, j \in [N]$. Suppose wlog that $a^\top x^i < a^\top x^{i+1} \Rightarrow \forall i \in [N-1]$ we can linearly separate the first i datapoints from the rest.

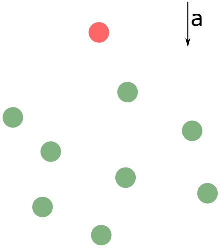


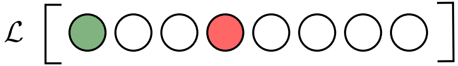


Dimension of W_D

Theorem. $\dim(W_D) = N(|\mathcal{L}| - 1)$.

Proof sketch. $\exists a \in \mathbb{R}^p$ s.t. $a^\top x^i \neq a^\top x^j$ for all distinct $i, j \in [N]$. Suppose wlog that $a^\top x^i < a^\top x^{i+1} \Rightarrow \forall i \in [N-1]$ we can linearly separate the first i datapoints from the rest.

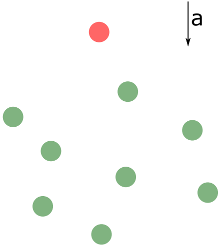


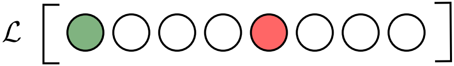


Dimension of W_D

Theorem. $\dim(W_D) = N(|\mathcal{L}| - 1)$.

Proof sketch. $\exists a \in \mathbb{R}^p$ s.t. $a^\top x^i \neq a^\top x^j$ for all distinct $i, j \in [N]$. Suppose wlog that $a^\top x^i < a^\top x^{i+1} \Rightarrow \forall i \in [N-1]$ we can linearly separate the first i datapoints from the rest.

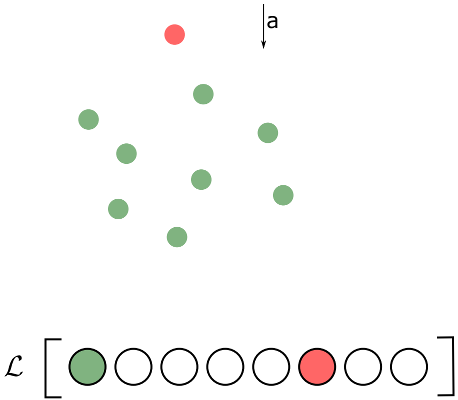




Dimension of W_D

Theorem. $\dim(W_D) = N(|\mathcal{L}| - 1)$.

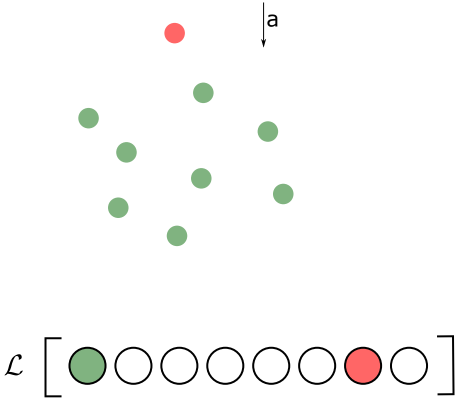
Proof sketch. $\exists a \in \mathbb{R}^p$ s.t. $a^\top x^i \neq a^\top x^j$ for all distinct $i, j \in [N]$. Suppose wlog that $a^\top x^i < a^\top x^{i+1} \Rightarrow \forall i \in [N-1]$ we can linearly separate the first i datapoints from the rest.



Dimension of W_D

Theorem. $\dim(W_D) = N(|\mathcal{L}| - 1)$.

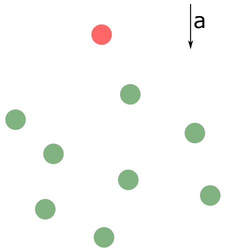
Proof sketch. $\exists a \in \mathbb{R}^p$ s.t. $a^\top x^i \neq a^\top x^j$ for all distinct $i, j \in [N]$. Suppose wlog that $a^\top x^i < a^\top x^{i+1} \Rightarrow \forall i \in [N-1]$ we can linearly separate the first i datapoints from the rest.

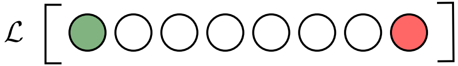


Dimension of W_D

Theorem. $\dim(W_D) = N(|\mathcal{L}| - 1)$.

Proof sketch. $\exists a \in \mathbb{R}^p$ s.t. $a^\top x^i \neq a^\top x^j$ for all distinct $i, j \in [N]$. Suppose wlog that $a^\top x^i < a^\top x^{i+1} \Rightarrow \forall i \in [N-1]$ we can linearly separate the first i datapoints from the rest.

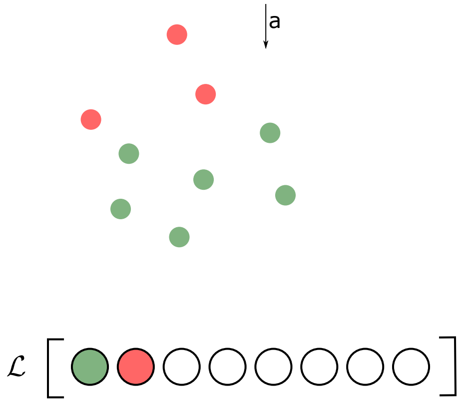




Dimension of W_D

Theorem. $\dim(W_D) = N(|\mathcal{L}| - 1)$.

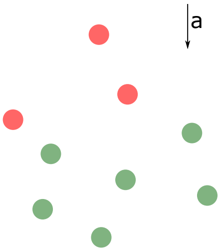
Proof sketch. $\exists a \in \mathbb{R}^p$ s.t. $a^\top x^i \neq a^\top x^j$ for all distinct $i, j \in [N]$. Suppose wlog that $a^\top x^i < a^\top x^{i+1} \Rightarrow \forall i \in [N-1]$ we can linearly separate the first i datapoints from the rest.

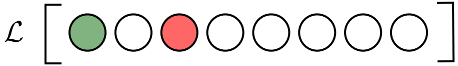


Dimension of W_D

Theorem. $\dim(W_D) = N(|\mathcal{L}| - 1)$.

Proof sketch. $\exists a \in \mathbb{R}^p$ s.t. $a^\top x^i \neq a^\top x^j$ for all distinct $i, j \in [N]$. Suppose wlog that $a^\top x^i < a^\top x^{i+1} \Rightarrow \forall i \in [N-1]$ we can linearly separate the first i datapoints from the rest.

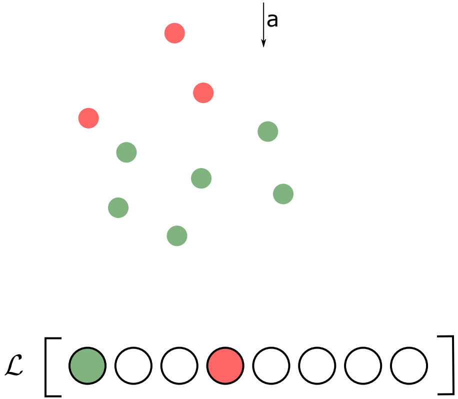




Dimension of W_D

Theorem. $\dim(W_D) = N(|\mathcal{L}| - 1)$.

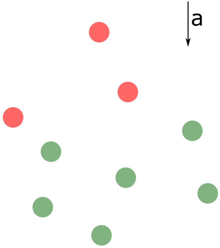
Proof sketch. $\exists a \in \mathbb{R}^p$ s.t. $a^\top x^i \neq a^\top x^j$ for all distinct $i, j \in [N]$. Suppose wlog that $a^\top x^i < a^\top x^{i+1} \Rightarrow \forall i \in [N-1]$ we can linearly separate the first i datapoints from the rest.

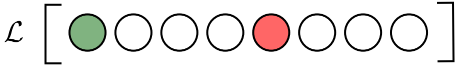


Dimension of W_D

Theorem. $\dim(W_D) = N(|\mathcal{L}| - 1)$.

Proof sketch. $\exists a \in \mathbb{R}^p$ s.t. $a^\top x^i \neq a^\top x^j$ for all distinct $i, j \in [N]$. Suppose wlog that $a^\top x^i < a^\top x^{i+1} \Rightarrow \forall i \in [N-1]$ we can linearly separate the first i datapoints from the rest.

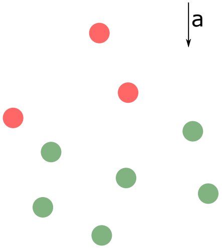


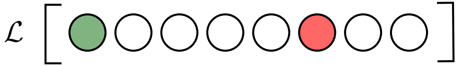


Dimension of W_D

Theorem. $\dim(W_D) = N(|\mathcal{L}| - 1)$.

Proof sketch. $\exists a \in \mathbb{R}^p$ s.t. $a^\top x^i \neq a^\top x^j$ for all distinct $i, j \in [N]$. Suppose wlog that $a^\top x^i < a^\top x^{i+1} \Rightarrow \forall i \in [N-1]$ we can linearly separate the first i datapoints from the rest.

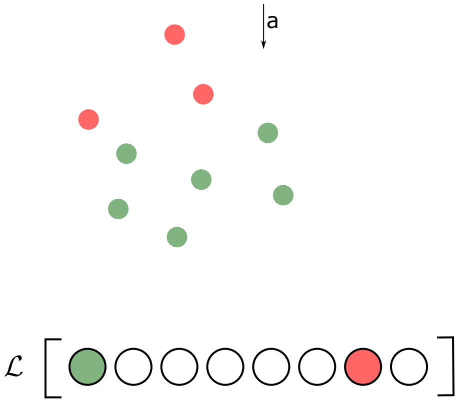




Dimension of W_D

Theorem. $\dim(W_D) = N(|\mathcal{L}| - 1)$.

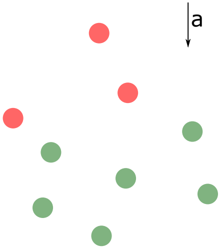
Proof sketch. $\exists a \in \mathbb{R}^p$ s.t. $a^\top x^i \neq a^\top x^j$ for all distinct $i, j \in [N]$. Suppose wlog that $a^\top x^i < a^\top x^{i+1} \Rightarrow \forall i \in [N-1]$ we can linearly separate the first i datapoints from the rest.

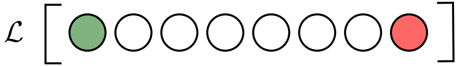


Dimension of W_D

Theorem. $\dim(W_D) = N(|\mathcal{L}| - 1)$.

Proof sketch. $\exists a \in \mathbb{R}^p$ s.t. $a^\top x^i \neq a^\top x^j$ for all distinct $i, j \in [N]$. Suppose wlog that $a^\top x^i < a^\top x^{i+1} \Rightarrow \forall i \in [N-1]$ we can linearly separate the first i datapoints from the rest.

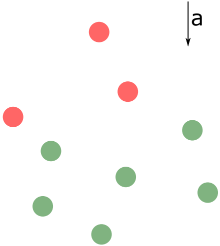




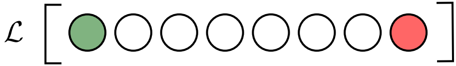
Dimension of W_D

Theorem. $\dim(W_D) = N(|\mathcal{L}| - 1)$.

Proof sketch. $\exists a \in \mathbb{R}^p$ s.t. $a^\top x^i \neq a^\top x^j$ for all distinct $i, j \in [N]$. Suppose wlog that $a^\top x^i < a^\top x^{i+1} \Rightarrow \forall i \in [N-1]$ we can linearly separate the first i datapoints from the rest.



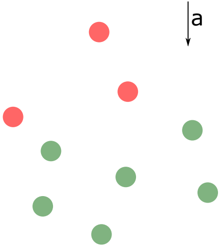
$\Rightarrow N(|\mathcal{L}| - 1)$ linearly independent points



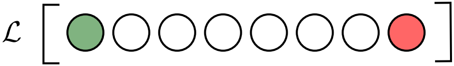
Dimension of W_D

Theorem. $\dim(W_D) = N(|\mathcal{L}| - 1)$.

Proof sketch. $\exists a \in \mathbb{R}^p$ s.t. $a^\top x^i \neq a^\top x^j$ for all distinct $i, j \in [N]$. Suppose wlog that $a^\top x^i < a^\top x^{i+1} \Rightarrow \forall i \in [N-1]$ we can linearly separate the first i datapoints from the rest.



$\Rightarrow N(|\mathcal{L}| - 1) + 1$ affinely independent points

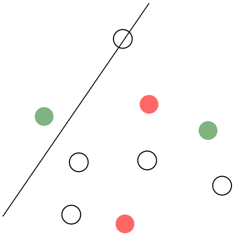


Very good partitions

$(I_L, I_R) \in \mathcal{I}$ is a good partition if $\forall i \notin I_L \cup I_R$, there exists a hyperplane $a^\top x = b$ that traverses x^i and correctly separates all but one datapoint in (I_L, I_R).

A good partition is called very good if, $a^\top x^k \neq b$ for all $k \in [N] \setminus (I_L \cup I_R \cup \{i\})$

Lemma. Good partition $\Rightarrow$ very good partition

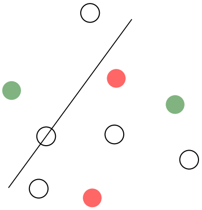


Very good partitions

$(I_L, I_R) \in \mathcal{I}$ is a good partition if $\forall i \notin I_L \cup I_R$, there exists a hyperplane $a^\top x = b$ that traverses x^i and correctly separates all but one datapoint in (I_L, I_R).

A good partition is called very good if, $a^\top x^k \neq b$ for all $k \in [N] \setminus (I_L \cup I_R \cup \{i\})$

Lemma. Good partition $\Rightarrow$ very good partition

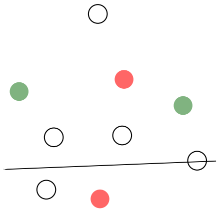


Very good partitions

$(I_L, I_R) \in \mathcal{I}$ is a good partition if $\forall i \notin I_L \cup I_R$, there exists a hyperplane $a^\top x = b$ that traverses x^i and correctly separates all but one datapoint in (I_L, I_R).

A good partition is called very good if, $a^\top x^k \neq b$ for all $k \in [N] \setminus (I_L \cup I_R \cup \{i\})$

Lemma. Good partition $\Rightarrow$ very good partition

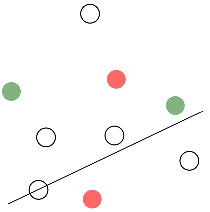


Very good partitions

$(I_L, I_R) \in \mathcal{I}$ is a good partition if $\forall i \notin I_L \cup I_R$, there exists a hyperplane $a^\top x = b$ that traverses x^i and correctly separates all but one datapoint in (I_L, I_R).

A good partition is called very good if, $a^\top x^k \neq b$ for all $k \in [N] \setminus (I_L \cup I_R \cup \{i\})$

Lemma. Good partition $\Rightarrow$ very good partition

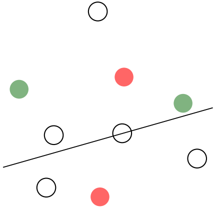


Very good partitions

$(I_L, I_R) \in \mathcal{I}$ is a good partition if $\forall i \notin I_L \cup I_R$, there exists a hyperplane $a^\top x = b$ that traverses x^i and correctly separates all but one datapoint in (I_L, I_R).

A good partition is called very good if, $a^\top x^k \neq b$ for all $k \in [N] \setminus (I_L \cup I_R \cup \{i\})$

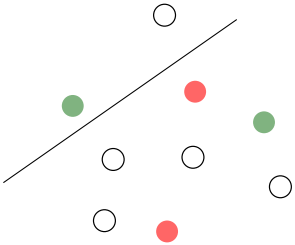
Lemma. Good partition $\Rightarrow$ very good partition

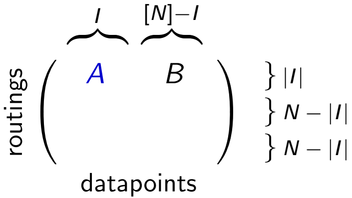


Very good partitions

Theorem. If $I = (I_L, I_R) \in \mathcal{I}$ is a very good partition, then the shattering inequality associated with (I_L, I_R) and $t = 1$ is facet-defining for W_1.

Proof sketch. $\forall i \in I$, define a routing that partitions all the points in $I_L \cup I_R$ correctly, except for x^i $\Rightarrow$ A has full affine rank

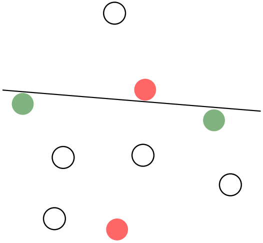


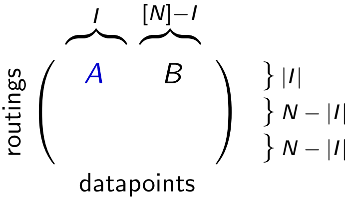


Very good partitions

Theorem. If $I = (I_L, I_R) \in \mathcal{I}$ is a very good partition, then the shattering inequality associated with (I_L, I_R) and $t = 1$ is facet-defining for W_1.

Proof sketch. $\forall i \in I$, define a routing that partitions all the points in $I_L \cup I_R$ correctly, except for $x^i \Rightarrow A$ has full affine rank

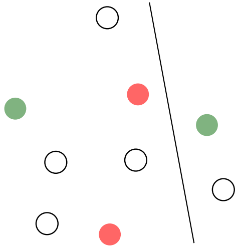


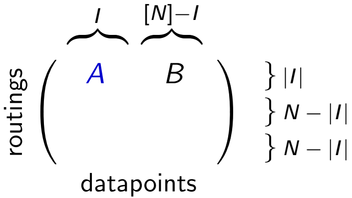


Very good partitions

Theorem. If $I = (I_L, I_R) \in \mathcal{I}$ is a very good partition, then the shattering inequality associated with (I_L, I_R) and $t = 1$ is facet-defining for W_1.

Proof sketch. $\forall i \in I$, define a routing that partitions all the points in $I_L \cup I_R$ correctly, except for $x^i \Rightarrow A$ has full affine rank

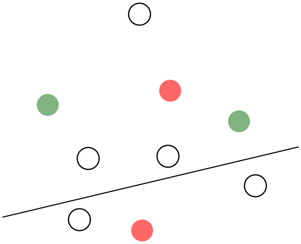


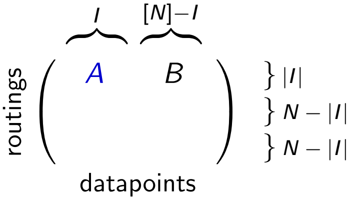


Very good partitions

Theorem. If $I = (I_L, I_R) \in \mathcal{I}$ is a very good partition, then the shattering inequality associated with (I_L, I_R) and $t = 1$ is facet-defining for W_1.

Proof sketch. $\forall i \in I$, define a routing that partitions all the points in $I_L \cup I_R$ correctly, except for $x^i \Rightarrow A$ has full affine rank

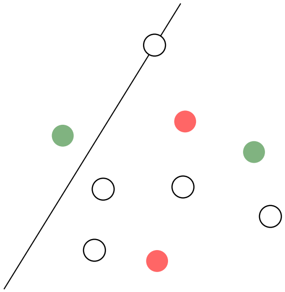




Very good partitions

Theorem. If $I = (I_L, I_R) \in \mathcal{I}$ is a very good partition, then the shattering inequality associated with (I_L, I_R) and $t = 1$ is facet-defining for W_1.

Proof sketch. $\forall j \in [N] \setminus I$, $\exists$ hyperplane through x^j and $i \in I$ s.t. all the points in $I_L \cup I_R$ but x^i are correctly separated.



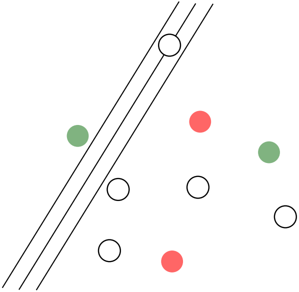
$$\text{routings}\ \begin{pmatrix} \overbrace{A}^{I} & \overbrace{B}^{[N]-I} \\ C & D \\ C & D' \end{pmatrix} \begin{matrix} \}\ |I| \\ \}\ N-|I| \\ \}\ N-|I| \end{matrix}$$

datapoints

Very good partitions

Theorem. If $I = (I_L, I_R) \in \mathcal{I}$ is a very good partition, then the shattering inequality associated with (I_L, I_R) and $t = 1$ is facet-defining for W_1.

Proof sketch. By shifting this hyperplane, define two routings that route all observations identically, except for x^j



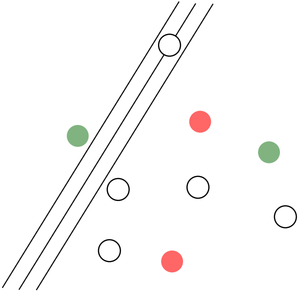
$$\text{routings}\begin{pmatrix} \overbrace{A}^{I} & \overbrace{B}^{[N]-I} \\ C & D \\ C & D' \end{pmatrix}\begin{matrix} \}\,|I| \\ \}\,N-|I| \\ \}\,N-|I| \end{matrix}$$

datapoints

Very good partitions

Theorem. If $I = (I_L, I_R) \in \mathcal{I}$ is a very good partition, then the shattering inequality associated with (I_L, I_R) and $t = 1$ is facet-defining for W_1.

Proof sketch. By shifting this hyperplane, define two routings that route all observations identically, except for x^j
$\Rightarrow \begin{pmatrix} C & D \end{pmatrix} - \begin{pmatrix} C & D' \end{pmatrix} = \begin{pmatrix} 0 & I \end{pmatrix}$



$$\text{routings}\;\begin{pmatrix} \overbrace{A}^{I} & \overbrace{B}^{[N]-I} \\ C & D \\ C & D' \end{pmatrix} \begin{matrix} \}\ |I| \\ \}\ N-|I| \\ \}\ N-|I| \end{matrix}$$

datapoints

Facets of W_1

Theorem. If the dataset is in general position, then every $(I_L, I_R) \in \mathcal{I}$ is a good partition.

Facets of W_1

Theorem. If the dataset is in general position, then every $(I_L, I_R) \in \mathcal{I}$ is a very good partition.

Facets of W_1

Theorem. If the dataset is in general position, then every $(I_L, I_R) \in \mathcal{I}$ is a very good partition.

$$\Downarrow$$

Main result 1. If the dataset is in general position, then the shattering inequalities are facets of W_1.

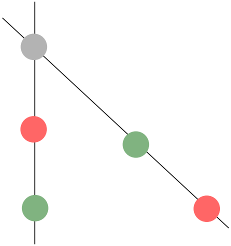
Facets of W_1

Theorem. If the dataset is in general position, then every $(I_L, I_R) \in \mathcal{I}$ is a very good partition.

$$\Downarrow$$

Main result 1. If the dataset is in general position, then the shattering inequalities are facets of W_1.

If the dataset is not in general position, shattering inequalities might not be facets of W_1.



not a very good partition

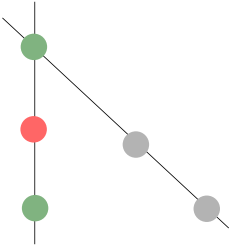
Facets of W_1

Theorem. If the dataset is in general position, then every $(I_L, I_R) \in \mathcal{I}$ is a very good partition.

$$\Downarrow$$

Main result 1. If the dataset is in general position, then the shattering inequalities are facets of W_1.

If the dataset is not in general position, shattering inequalities might not be facets of W_1.



not a very good partition

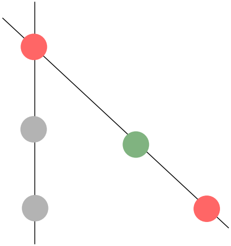
Facets of W_1

Theorem. If the dataset is in general position, then every $(I_L, I_R) \in \mathcal{I}$ is a very good partition.

$$\Downarrow$$

Main result 1. If the dataset is in general position, then the shattering inequalities are facets of W_1.

If the dataset is not in general position, shattering inequalities might not be facets of W_1.



not a very good partition

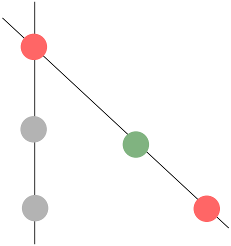
Facets of W_1

Theorem. If the dataset is in general position, then every $(I_L, I_R) \in \mathcal{I}$ is a very good partition.

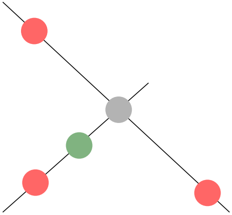
$$\Downarrow$$

Main result 1. If the dataset is in general position, then the shattering inequalities are facets of W_1.

If the dataset is not in general position, shattering inequalities might not be facets of W_1.



not a very good partition



not a very good partition

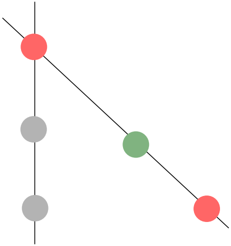
Facets of W_1

Theorem. If the dataset is in general position, then every $(I_L, I_R) \in \mathcal{I}$ is a very good partition.

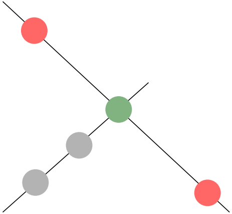
$$\Downarrow$$

Main result 1. If the dataset is in general position, then the shattering inequalities are facets of W_1.

If the dataset is not in general position, shattering inequalities might not be facets of W_1.



not a very good partition



not a very good partition

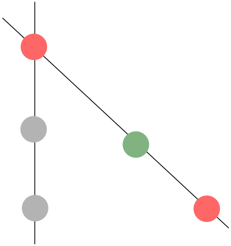
Facets of W_1

Theorem. If the dataset is in general position, then every $(I_L, I_R) \in \mathcal{I}$ is a very good partition.

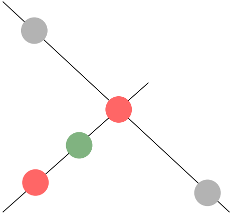
$$\Downarrow$$

Main result 1. If the dataset is in general position, then the shattering inequalities are facets of W_1.

If the dataset is not in general position, shattering inequalities might not be facets of W_1.



not a very good partition



not a very good partition

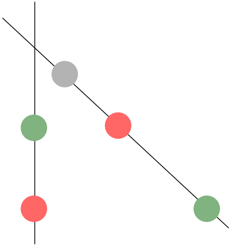
Facets of W_1

Theorem. If the dataset is in general position, then every $(I_L, I_R) \in \mathcal{I}$ is a very good partition.

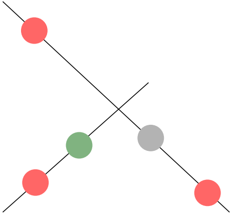
$$\Downarrow$$

Main result 1. If the dataset is in general position, then the shattering inequalities are facets of W_1.

But even when the dataset is not in general position, shattering inequalities could be facets of W_1.



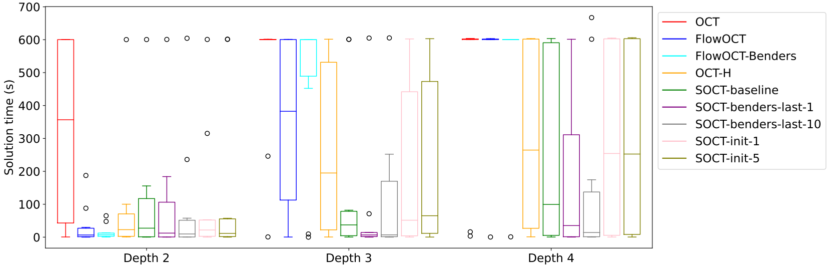
very good partition



very good partition

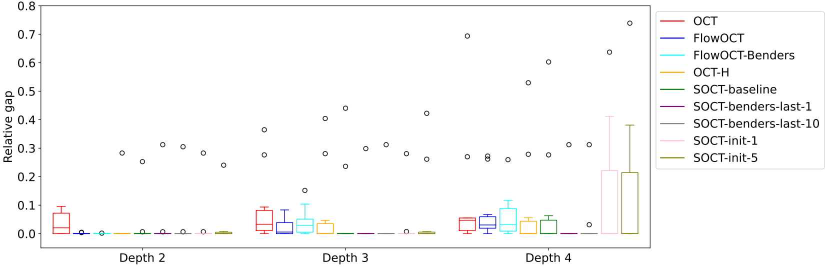
Computational experiments

- The numerical experiments by Boutilier, M. and Zhou (2022, 2023) have shown that the MIP formulations using shattering inequalities outperform other MIP formulations in terms of solution time and relative gap.



Computational experiments

- The numerical experiments by Boutilier, M. and Zhou (2022, 2023) have shown that the MIP formulations using shattering inequalities outperform other MIP formulations in terms of solution time and relative gap.



Computational experiments

- The numerical experiments by Boutilier, M. and Zhou (2022, 2023) have shown that the MIP formulations using shattering inequalities outperform other MIP formulations in terms of solution time and relative gap.
- To validate our theoretical findings, we perform numerical experiments that specifically exploit shattering inequalities defined at the root node —the only ones that are guaranteed to be facets of W_D if the dataset is in general position.

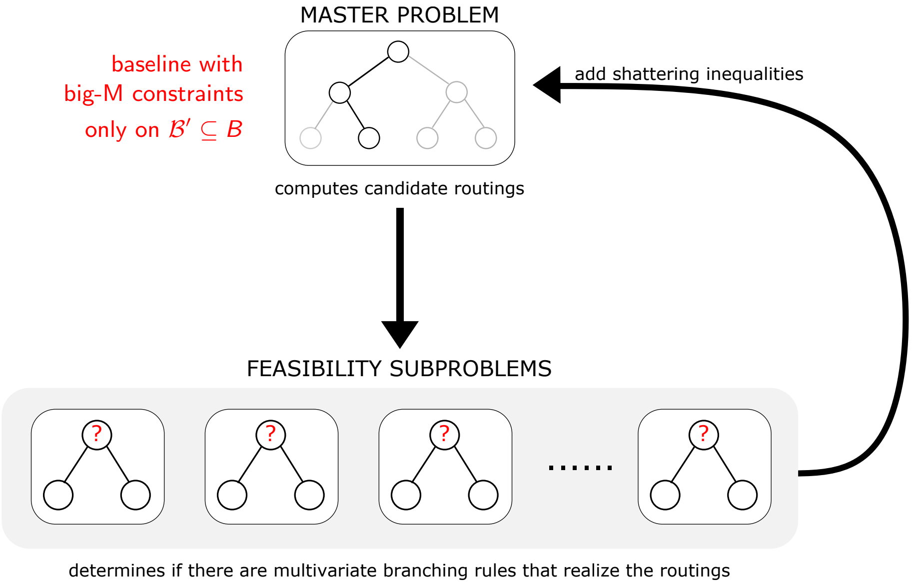
Computational experiments

- The numerical experiments by Boutilier, M. and Zhou (2022, 2023) have shown that the MIP formulations using shattering inequalities outperform other MIP formulations in terms of solution time and relative gap.
- To validate our theoretical findings, we perform numerical experiments that specifically exploit shattering inequalities defined at the root node —the only ones that are guaranteed to be facets of W_D if the dataset is in general position.
- We use both numerical and categorical datasets to test whether having a datapoints in general position impacts computational performance.

Computational experiments

- The numerical experiments by Boutilier, M. and Zhou (2022, 2023) have shown that the MIP formulations using shattering inequalities outperform other MIP formulations in terms of solution time and relative gap.
- To validate our theoretical findings, we perform numerical experiments that specifically exploit shattering inequalities defined at the root node —the only ones that are guaranteed to be facets of W_D if the dataset is in general position.
- We use both numerical and categorical datasets to test whether having a datapoints in general position impacts computational performance.
- We compare against the baseline model.

Decomposition



Separation

$\forall t \in \mathcal{B} \setminus \mathcal{B}'$, given candidate routing w^*, check feasibily of $(\star)$:

$$\begin{aligned} & a_t^\top x^i \leq b_t - 1 && \forall\, i \in [N] : w^*_{i,2t} = 1 \\ & a_t^\top x^i \geq b_t + 1 && \forall\, i \in [N] : w^*_{i,2t+1} = 1 \\ & (a_t, b_t) \in \mathbb{R}^{p+1} && \end{aligned}$$

If the system is infeasible, each Irreducible Infeasible Subsystem (IIS) provides:

- a subset I of datapoints that cannot be shattered
- a partition of I that cannot be separated

$$\Rightarrow \sum_{i \in I : w^*_{i,2t} = 1} w_{i,2t} + \sum_{i \in I : w^*_{i,2t+1} = 1} w_{i,2t+1} \leq |I| - 1$$

Separation

$\forall t \in \mathcal{B} \setminus \mathcal{B}'$, given candidate routing w^*, check feasibily of $(\star)$:

$$
\begin{aligned}
& a_t^\top x^i \leq b_t - 1 && \forall\, i \in [N] : w^*_{i,2t} = 1 \\
& a_t^\top x^i \geq b_t + 1 && \forall\, i \in [N] : w^*_{i,2t+1} = 1 \\
& (a_t, b_t) \in \mathbb{R}^{p+1}
\end{aligned}
$$

If the system is infeasible, each Irreducible Infeasible Subsystem (IIS) provides:

- a subset I of datapoints that cannot be shattered
- a partition of I that cannot be separated

$\Rightarrow$ For a binary w^* yielding an infeasible system $(\star)$, each vertex of the dual of $(\star)$ corresponds to an IIS [Gleeson and Ryan, 1990]

Separation

We define Separation(w, `nodes`, `n_cuts`):

- w is the candidate routing to separate (possibly fractional)
- `nodes` is the subset of $\mathcal{B}$ for which we generate shattering inequalities
- `n_cuts` is the maximum number of cuts to generate for each $t \in$ `nodes`.

Note: If w is binary, then the separation is exact.

Separation

We define Separation(w, `nodes`, `n_cuts`):

- w is the candidate routing to separate (possibly fractional)
- `nodes` is the subset of $\mathcal{B}$ for which we generate shattering inequalities
- `n_cuts` is the maximum number of cuts to generate for each $t \in$ `nodes`.

Two models:

1. `Root-`*x* calls Separation(w,1,x), $x \in \{1, 5, 10, 15, 20\}$ over the LP relaxation of the master problem, adding cuts up front as initial cuts.
2. `Root-`*x*`-Ben-`*y* uses hybrid decomposition approach with $\mathcal{B}' = \emptyset$. Calls Separation(w,1,x) to add initial cuts to the master problem. Additional cuts are iteratively added to the master problem by calling Separation(w,$\mathcal{B}$,y); $x, y \in \{1, 5, 10\}$.

Experimental setup

- 15 datasets from the UCI Machine Learning Repository
- Python 3.8.10, Gurobi 10.0, 3.00 GHz 6-core Intel Corei5-8500 processor and 16 GB RAM
- 10 minute time limit
- Code available at `https://github.com/zachzhou777/S-OCT`

First set of experiments

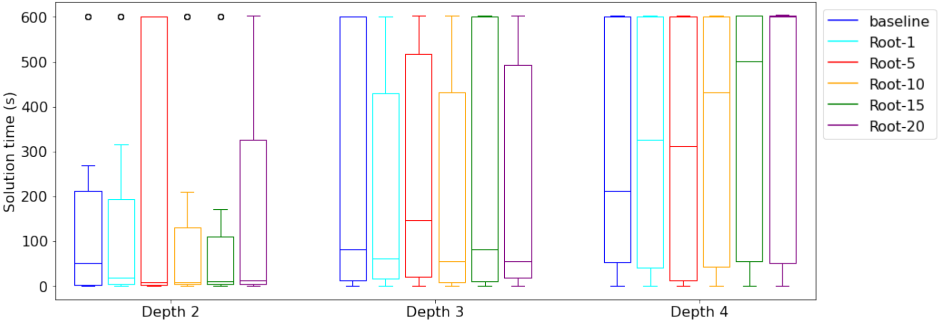
GOAL: does adding shattering inequalities at the root node improve computational performance?

We compared:

- baseline model: only big-M constraints
- Root-x, $x \in \{1, 5, 10, 15, 20\}$

First set of experiments

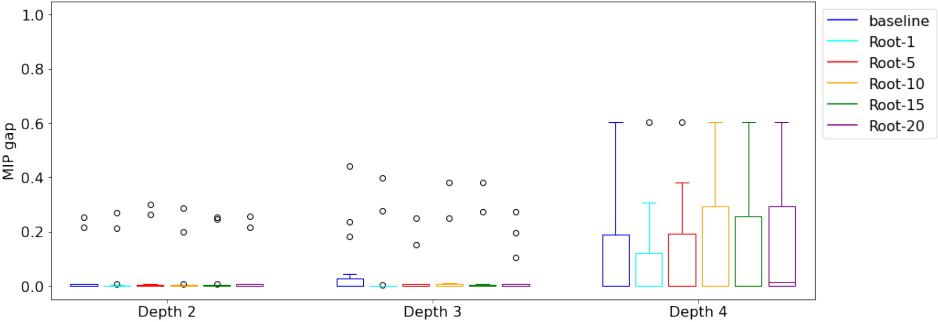
GOAL: does adding shattering inequalities at the root node improve computational performance?



Solution time at depths 2, 3 and 4

First set of experiments

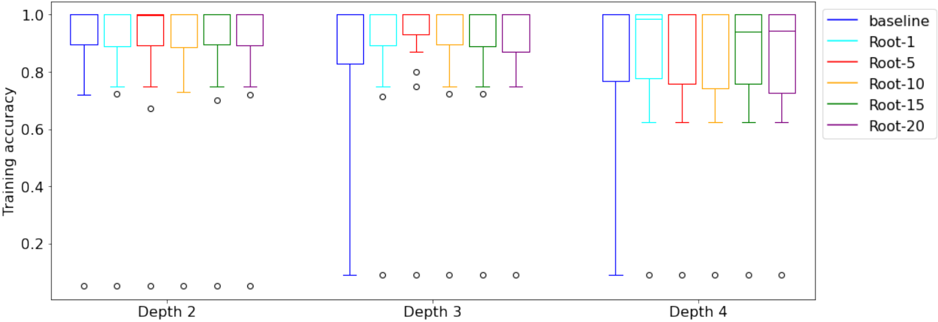
GOAL: does adding shattering inequalities at the root node improve computational performance?



Relative gap at depths 2, 3 and 4

First set of experiments

GOAL: does adding shattering inequalities at the root node improve computational performance?



Training accuracy at depths 2, 3 and 4

Second set of experiments

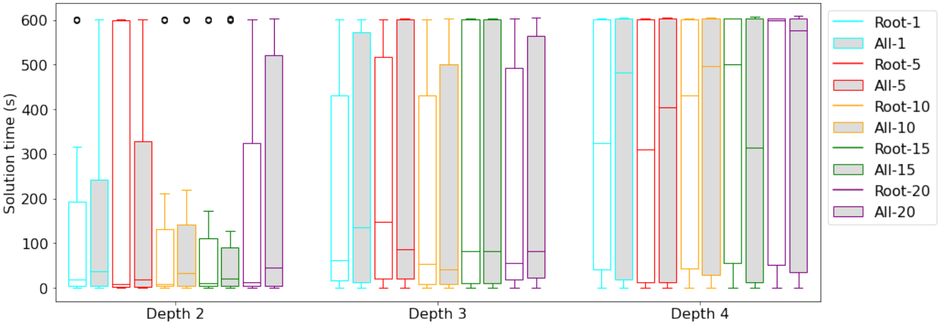
GOAL: does adding shattering inequalities at all the nodes improve computational performance?

We compared:

- Root-x, $x \in \{1, 5, 10, 15, 20\}$
- All-x, $x \in \{1, 5, 10, 15, 20\}$

Second set of experiments

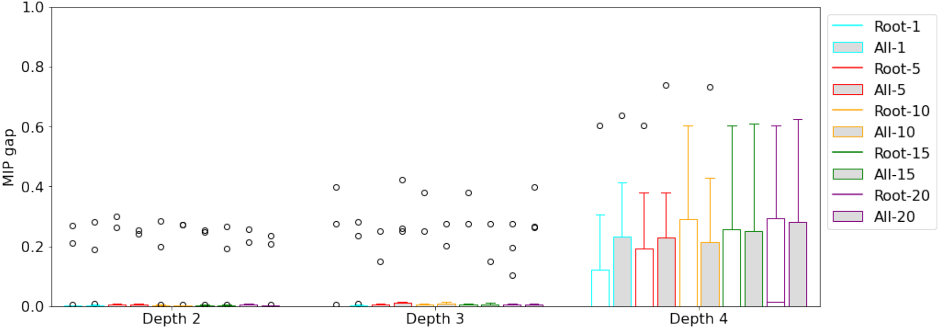
GOAL: does adding shattering inequalities at all the nodes improve computational performance?



Solution time at depths 2, 3 and 4

Second set of experiments

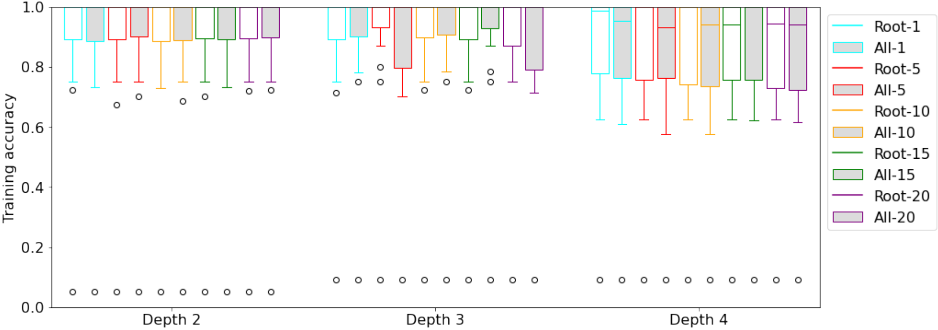
GOAL: does adding shattering inequalities at all the nodes improve computational performance?



Relative gap at depths 2, 3 and 4

Second set of experiments

GOAL: does adding shattering inequalities at all the nodes improve computational performance?



Training accuracy at depths 2, 3 and 4

Second set of experiments

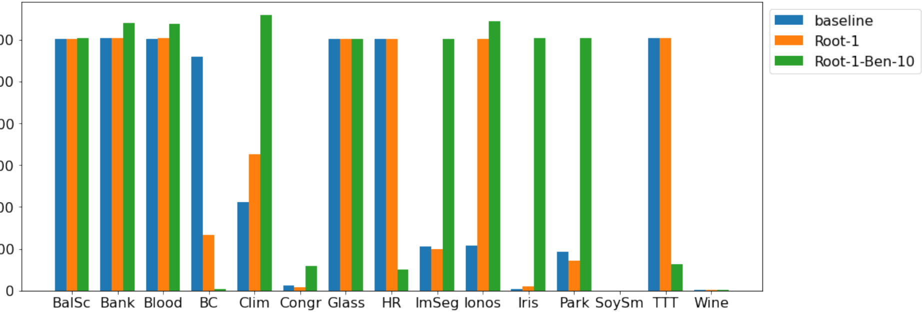
GOAL: does adding shattering inequalities at all the nodes improve computational performance?

We compared:

- baseline model
- Root-x, $x \in \{1, 5, 10\}$
- Root-x-Ben-y, $x, y \in \{1, 5, 10\}$

Second set of experiments

GOAL: does adding shattering inequalities at all the nodes improve computational performance?



Solution time at depth 4

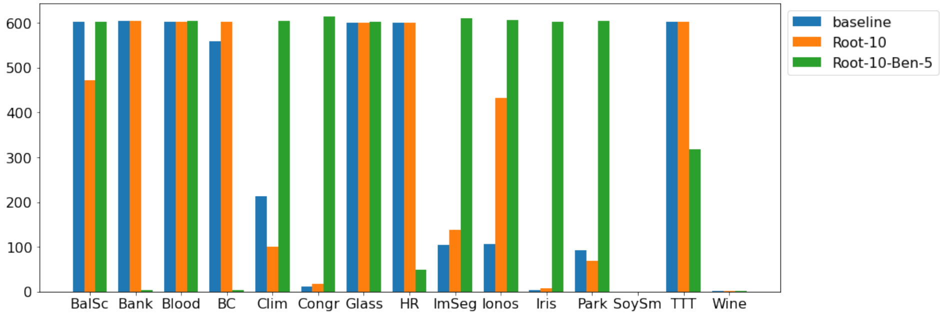
Second set of experiments

GOAL: does adding shattering inequalities at all the nodes improve computational performance?

Solution time at depth 4

Second set of experiments

GOAL: does adding shattering inequalities at all the nodes improve computational performance?



Solution time at depth 4

Conclusion

- Shattering inequalities are sparse and capture the combinatorial structure of the problem.
- We have established conditions s.t. the shattering inequalities are facets (dataset in general position, very good partitions).
- Computational experiments show that shattering inequalities at the root node are useful to reduce MIP gap.
- Future directions: more (combinatorial) cuts, robust multivariate decision trees.

References

Justin Boutillier, Carla Michini and Zachary Zhou. Shattering Inequalities for Learning Optimal Decision Trees. Proceedings of CPAIOR 2022.

C.Michini and Z.Zhou. Optimal multivariate decision trees, Constraints 28, 549–577 (2023).

C.Michini and Z.Zhou. A polyhedral study of multivariate decision trees, submitted, 2024.