Towards the Use of Decision Models (Hierarchy of Choquet Integrals) in Machine Learning and Image Processing

Christophe Labreuche [1,2]

[1] **Thales**, cortAIx-Labs, Palaiseau, France
[2]**SINCLAIR AI Lab**, Palaiseau, France
email: christophe.labreuche@thalesgroup.com

In collaboration with **Nicolas Atienza, Roman Bresson, Johanne Cohen, Eyke Hüllermeier, Michèle Sebag**

Work supported by:
FaRADAI project (ref. 101103386) funded by the European Commission under the European Defence Fund (EDF-2021-DIGIT-R)

Outline

Outline

Multi-Criteria Decision problem

Multi-Criteria Decision Aiding (MCDA)

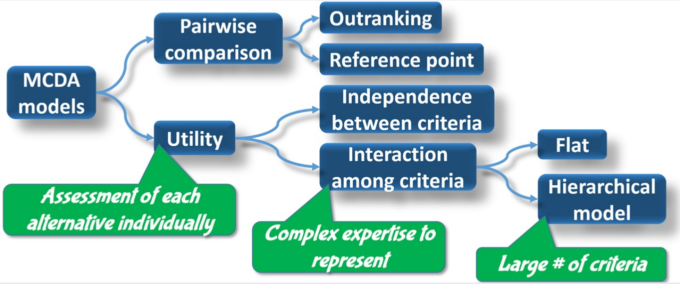
- $N = \{1, \ldots, n\}$: index set of attributes/features.
- X_i: set of values representing attribute/feature i (for $i \in N$).
- $X = X_1 \times \cdots \times X_n$: set of alternatives/instances.
  $\mathbf{x} = (x_1, \ldots, x_n) \in X$ with $x_i \in X_i$.
- Problem to solve, given a set of alternatives in X:
  - choose the *most preferred* one
  - rank the alternatives from best to worse
  - sort the alternatives into preferential categories
- $U : X \to \mathbb{R}$: utility representing preferences of decision maker over X
  - $U(\mathbf{y}) > U(\mathbf{x})$: $\mathbf{y}$ is preferred to $\mathbf{x}$

From a typical MCDA context . . .

Multi-Criteria Decision Aiding (MCDA)

Selected model: Hierarchical Choquet Integral.

Why?



Model characteristics

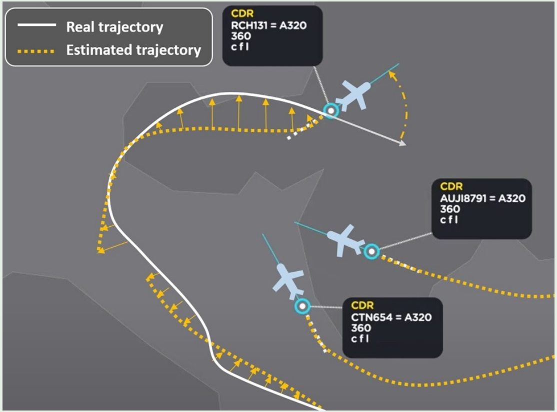
- **Model from Social Sciences** (cognitive bias)
- **Interpretable model**

Model Construction

- **Elicitation** (small & consistent data)

From a typical MCDA context . . .

Design of Tracking System for Air Traffic Management

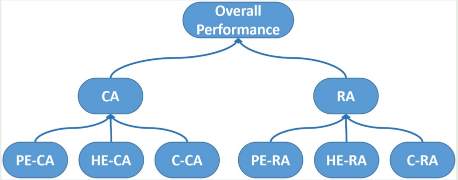


Aim: Use MCDA to select the best tracking system.
Tracking quality attributes:

- Position Error (PE)
- Heading Error (HE)
- Completeness (C)

Attributes are measured for each type of aircraft:

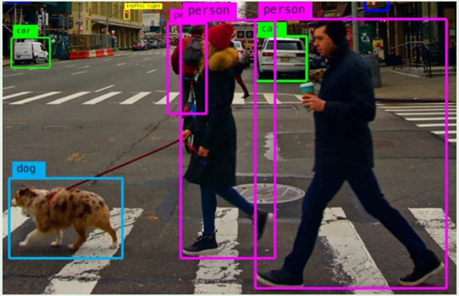
- Commercial Airplanes (CA)
- Recreational Airplanes (RA)



. . . towards the use of MCDA within Machine Learning (ML)

Object Detection in Images

Aim: Locate bounding boxes around objects of interest and classify them.



Model characteristics

- **Deep Learning**
- **Not Interpretable**

Model Construction

- **Machine Learning** (large & noisy dataset)

What we'd like to have . . .

- **Incorporate MCDA within ML to improve its interpretability**

Outline

General model

Decomposable preference model [Krantz et al'1971]

$$U(\mathbf{x}) = A(u_1(x_1), \ldots, u_n(x_n))$$

where

- $u_i : X_i \to [0,1]$: marginal utility function
- $A : [0,1]^n \to [0,1]$: aggregation function

Scale [0, 1] is typically a *satisfaction degree*.
Properties:

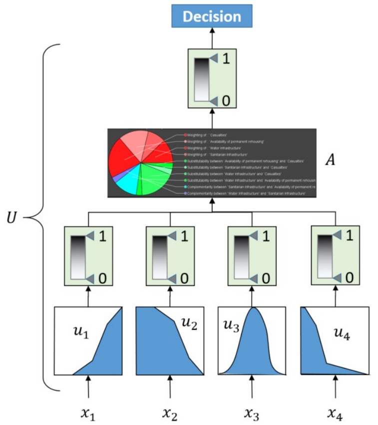
- Monotonicity

$$u_i(x_i) \geq u_i(x_i') : x_i \text{ at least as good as } x_i'$$

$$v_1 \geq v_1', \ldots, v_n \geq v_n' \;\Rightarrow\; A(\mathbf{v}) \geq A(\mathbf{v}')$$

- Idempotency:

$$A(\alpha, \ldots, \alpha) = \alpha \quad \forall \alpha \in [0,1]$$



Simplest aggregation model

Weighted sum

$$\mathrm{WS}_{\mathbf{w}}(\mathbf{v}) = \sum_{i\in N} w_i \ v_i,$$

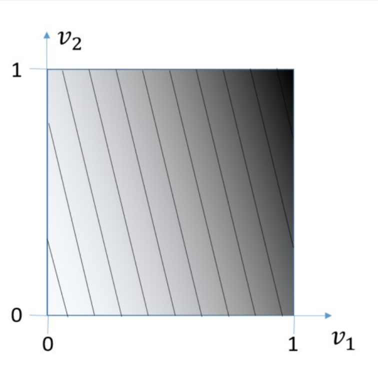
where $\mathbf{w} = (w_1, \ldots, w_n)$ are the criteria weights with

$$w_i \geq 0 \qquad \text{(monotonicity)}$$

$$\sum_{i\in N} w_i = 1 \quad \text{(idempotency)}$$

Interest of the WS:

- Very simple to understand
- Criteria weights make sense to people
  ($\Rightarrow$ *Feature Attribution* in ML)



Generalization of the Weighted Sum

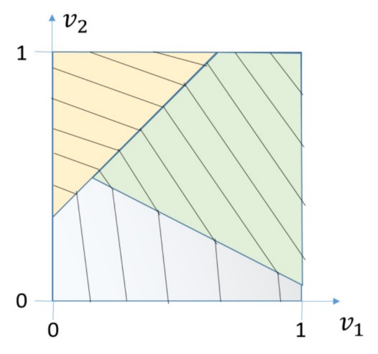
Piecewise Affine function

Model $\mathrm{PA}(\mathbf{v})$

- $\mathcal{D}$: (finite) partition of $[0, 1]^n$
- PA is a (monotone and idempotent) WS in each domain of $\mathcal{D}$
- PA is continuous

Interest of the PA:

- Universal approximator
  (⇒ see ReLU-based Neural Networks in ML)
- Might be doable to understand it
  (⇒ SP-LIME in ML [Singh et al'2016])



A Particular Piecewise Affine model

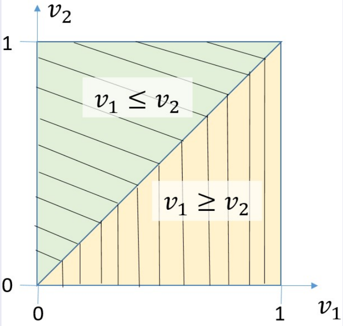
Choquet integral

Idea:

- Idempotency: it makes sense to compare v_i with v_j
- Piecewise Affine function in domains of the form $v_3 \geq v_1 \geq v_2 \geq \cdots$

$$C_m(\mathbf{v}) = \sum_{S \subseteq N} m(S) \cdot \bigwedge_{i \in S} v_i \qquad (\bigwedge \equiv \min)$$

- m: Möbius coefficients
  - Monotonicity: $\forall i \in N \; \forall S \subseteq N \setminus \{i\} \quad \sum_{T \subseteq S} m(T \cup \{i\}) \geq 0$
  - Normalization: $\sum_{S \subseteq N} m(S) = 1$
- Very versatile model:
  - Complementarity among criteria ($m(S) > 0$) $\cdots$ veto
  - Redundancy among criteria ($m(S) < 0$) $\cdots$ favor



A Particular Piecewise Affine model

Complexity of the Choquet integral

The Choquet integral contains 2^n parameters:

$$m : 2^N \to \mathbb{R}.$$

Submodels of the Choquet integral

$$C_m(\mathbf{v}) = \sum_{S \in \mathcal{S}} m(S) \cdot \min_{i \in S} v_i$$

where $\mathcal{S} \subseteq 2^N$.
Example:

- k-additive:

$$\mathcal{S} = \{ S \subseteq N \ : \ |S| \leq k \}.$$

Choquet integral

2-additive Choquet integral

$$C_w(\mathbf{v}) = \sum_{i=1}^{n} w_i \; v_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{i,j}^{\wedge} \, (v_i \wedge v_j) + \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{i,j}^{\vee} \, (v_i \vee v_j) \qquad [\wedge \equiv \min, \vee \equiv \max]$$

- **Monotonicity:** $\forall i, j \in N \quad w_i \geq 0, w_{i,j}^{\wedge} \geq 0, w_{i,j}^{\vee} \geq 0$
- **Normalization:** $\sum_{i=1}^{n} w_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{i,j}^{\wedge} + \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{i,j}^{\vee} = 1$

$$0.25 \times v_{MS} + 0.25 \times v_{MS} \wedge v_{HF} + 0.24 \times v_{MS} \wedge v_{ED} + 0.09 \times v_{ED} + 0.08 \times v_{HF} + 0.09 \times v_{ED} \wedge v_{HF}$$

- Mission Success: 28%
- Human Factors: 33%
- Economic Data: 92%

Choquet integral

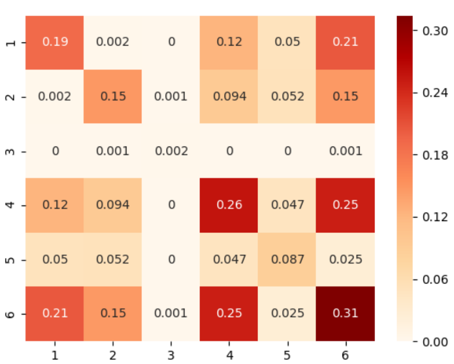
Interpretation

Importance of criteria:

$$\phi_i = w_i + \sum_{j \neq i} \frac{w^{\wedge}_{i,j} + w^{\vee}_{i,j}}{2}$$

Interaction between criteria:

$$I_{i,j} = \begin{cases} w^{\wedge}_{i,j} \text{ if } w^{\wedge}_{i,j} \neq 0 \\ -w^{\vee}_{i,j} \text{ else} \end{cases}$$

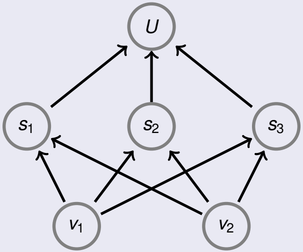


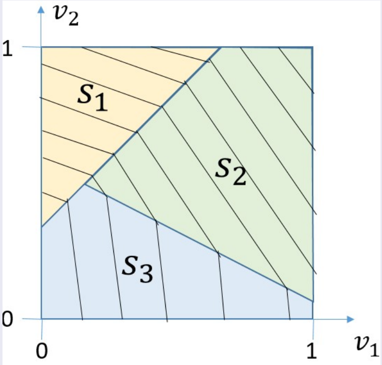
Outline

Interconnected Choquet Integrals

Theorem [Ovchinnikov'2002]

Any continuous piecewise affine function can be represented by a network of interconnected Choquet integrals.





- Layer a_i: inputs
- Layer s_j: weighted sums of the inputs (1 per affine part)
- Layer U: MinMax function that triggers the correct affine function

Interconnected Choquet Integrals

Discussion

Drawback of previous architecture

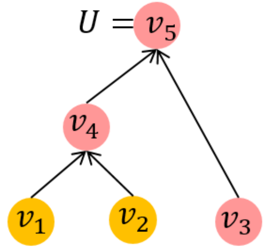
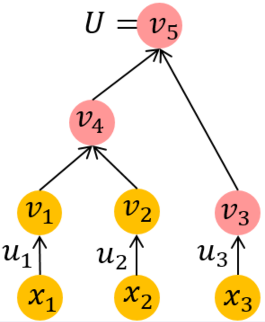
- The middle layer (s_j) might be extremely large;
- Fully connected layers are hard to understand and explain.

Modification:

- Consider a tree rather than a fully connected network: more understandable;
- The same approximation quality might be achieved with less nodes but deeper graphs.

Hierarchical models

Limitation of a flat model

HCI (Hierarchical Choquet Integral)	UHCI (Utilitaristic Hierarchical Choquet Integral)
	

Outline

Identifiability

Identifiability

Identifiability of a model class: injectivity of its parameterization.

- $\mathcal{C} = \{\mathcal{F}_\theta, \theta \in \Theta\}$ a family of functions defined on X
- Θ the parameter space
- $\mathcal{F}_\theta \in \mathcal{C}$ parameterized by θ

Then $\mathcal{C}$ is identifiable if and only if: $\forall \mathbf{x} \in X, \mathcal{F}_\theta(\mathbf{x}) = \mathcal{F}_{\theta'}(\mathbf{x}) \Rightarrow \theta = \theta'$.

Illustration

$\Theta = \mathbb{R}^2$, $X = \mathbb{R}$.
$\mathcal{C}_1 = \{\mathcal{F}_{a,b} : x \mapsto abx,\ (a,b) \in \Theta\}$ is not identifiable, as $\mathcal{F}_{3,4} = \mathcal{F}_{6,2}$
$\mathcal{C}_2 = \{\mathcal{F}_{a,b} : x \mapsto ax + b,\ (a,b) \in \Theta\}$ is identifiable

Identifiability

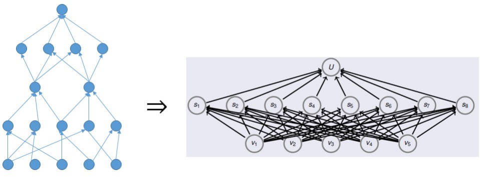
Interest of Identifiability

- It is easier to learn
- The model is interpretable

Our ambition

Identifiability of the **UHCI parameters** but also the **hierarchy**.

Not a foregone conclusion · · · wrong for graphs

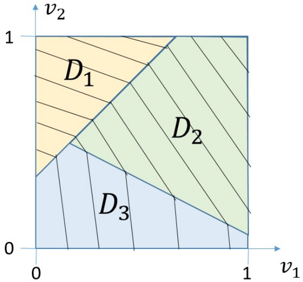


Outline

Separation frontiers of an HCI model

HCI model A: piecewise affine function

- Partition $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_p\}$ of $[0,1]^n$
- Set of affine functions $\mathcal{L} = \{L_1, \ldots, L_p\}$
- For all $j \in \{1, \ldots, p\}$ and $\mathbf{v} \in \mathcal{D}_j$, $A(\mathbf{v}) = L_j(\mathbf{v})$



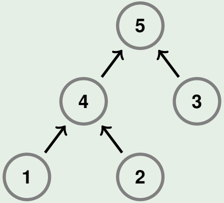
Separation frontiers of an HCI model

As A is continuous, the separation frontiers between the affine parts are hyperplanes.

Separation frontiers of an HCI model

Illustration

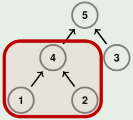
Model:



$$v_4 = \frac{v_1 + v_1 \wedge v_2}{2}$$
$$v_5 = \frac{v_3 + v_3 \wedge v_4}{2}$$

Linear parts:

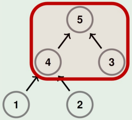
- $v_1, v_2 \mapsto v_4$ has 2 linear parts:



$$v_1 \quad \text{and} \quad \frac{v_1 + v_2}{2}$$

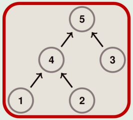
Separation frontiers ...

- ... of $v_3, v_4 \mapsto v_5$:



$$v_3 = v_4$$

- ... hence of $v_1, v_2, v_3 \mapsto v_5$:



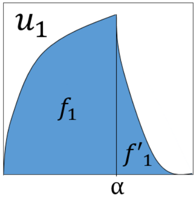
$$v_1 = v_2 \ , \ v_1 = v_3 \ \text{ and } \ \frac{v_1 + v_2}{2} = v_3$$

Separation frontiers of an UHCI model

Assumption

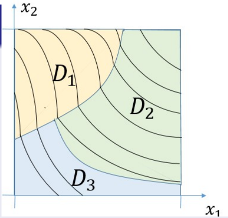
Marginal utility functions are piecewise C^1 functions

$$u_1(x_1) = \begin{cases} f_1(x_1) \text{ if } x_1 \leq \alpha \\ f_1'(x_1) \text{ else} \end{cases}$$



UHCI model U: piecewise C^1 model

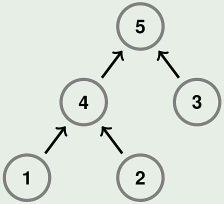
- Partition $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_p\}$ of X
- Set of C^1 functions $\mathcal{C} = \{C_1, \ldots, C_p\}$
- For all $j \in \{1, \ldots, p\}$ and $\mathbf{x} \in \mathcal{D}_j$, $U(\mathbf{x}) = C_j(\mathbf{x})$



Separation frontiers of an UHCI model

Illustration

Model:



$$v_1 = f_1(x_1) \quad v_2 = f_2(x_2) \quad v_3 = \begin{cases} f_3(x_3) \text{ if } x_3 \leq \alpha \\ f_3'(x_3) \text{ else} \end{cases}$$

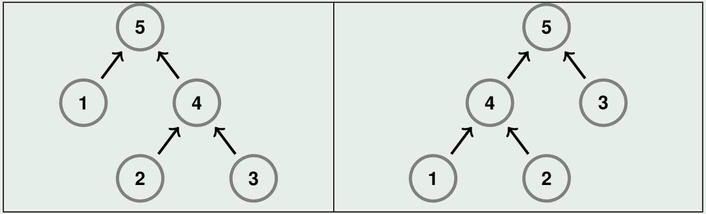
$$v_4 = \frac{v_1 + v_1 \wedge v_2}{2}$$

$$v_5 = \frac{v_3 + v_3 \wedge v_4}{2}$$

Separation of $v_1, v_2, v_3 \mapsto v_5$	Separation of $x_1, x_2, x_3 \mapsto v_5$
$v_1 = v_2$	$f_1(x_1) = f_2(x_2)$
$v_1 = v_3$	$f_1(x_1) = f_3(x_3)$, $f_1(x_1) = f_3'(x_3)$
$\frac{v_1+v_2}{2} = v_3$	$\frac{f_1(x_1)+f_2(x_2)}{2} = f_3(x_3)$, $\frac{f_1(x_1)+f_2(x_2)}{2} = f_3'(x_3)$
	$x_3 = \alpha$

Can we deduce the hierarchy from the separations?

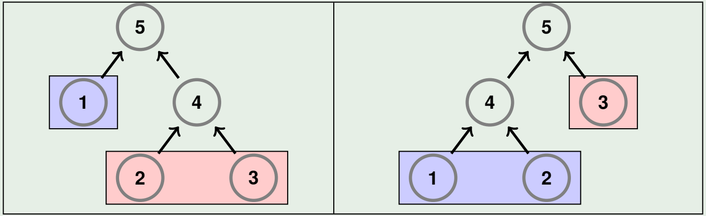
Illustration



Separation of $x_1, x_2, x_3 \mapsto v_5$	Separation of $x_1, x_2, x_3 \mapsto v_5$
$f_2(x_2) = f_3(x_3)$	$f_1(x_1) = f_2(x_2)$
$f_1(x_1) = f_2(x_2)$	$f_1(x_1) = f_3(x_3) \quad , \quad f_1(x_1) = f'_3(x_3)$
$f_1(x_1) = \frac{f_2(x_2)+f_3(x_3)}{2}$	$\frac{f_1(x_1)+f_2(x_2)}{2} = f_3(x_3)$
$f_1(x_1) = \frac{f_2(x_2)+f'_3(x_3)}{2}$	$\frac{f_1(x_1)+f_2(x_2)}{2} = f'_3(x_3)$
$x_3 = \alpha$	$x_3 = \alpha$

Can we deduce the hierarchy from the separations?

Illustration



Separation of $x_1, x_2, x_3 \mapsto v_5$	Separation of $x_1, x_2, x_3 \mapsto v_5$
$f_2(x_2) = f_3(x_3)$	$f_1(x_1) = f_2(x_2)$
$f_1(x_1) = f_2(x_2)$	$f_1(x_1) = f_3(x_3)$, $f_1(x_1) = f_3'(x_3)$
$f_1(x_1) = \frac{f_2(x_2)+f_3(x_3)}{2}$	$\frac{f_1(x_1)+f_2(x_2)}{2} = f_3(x_3)$
$f_1(x_1) = \frac{f_2(x_2)+f_3'(x_3)}{2}$	$\frac{f_1(x_1)+f_2(x_2)}{2} = f_3'(x_3)$
$x_3 = \alpha$	$x_3 = \alpha$

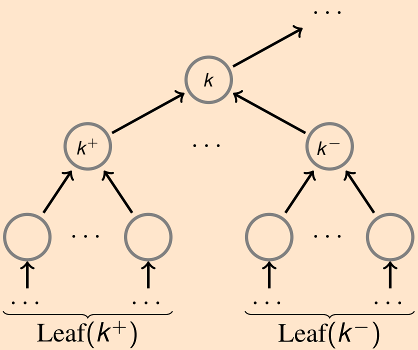
Can we deduce the hierarchy from the separations?

Theorem [*Bresson et al*, KR'2021]

The separation frontiers are of the form

- $x_i = \alpha$ for a leaf node $i \in N$;
- $\sum_{\ell \in K^+} w_\ell \; u_\ell(x_\ell) = \sum_{\ell \in K^-} w_\ell \; u_\ell(x_\ell)$ such that
  - $w_\ell > 0$ for all $\ell \in K^+ \cup K^-$
  - $\exists k \in V$ and $k^+, k^- \in \mathrm{Children}(k)$ s.t.

$$K^+ \subseteq \mathrm{Leaf}(k^+) \text{ and } K^- \subseteq \mathrm{Leaf}(k^-)$$



Outline

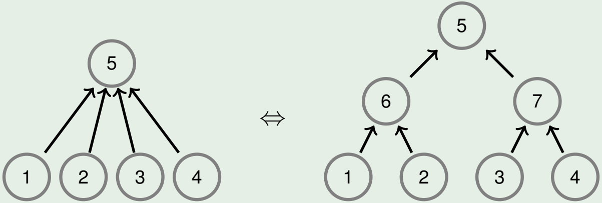
Assumptions

Fact

From the previous construction, the hierarchy cannot always be uniquely determined.

Counter-example #1

For a weighted sum, the hierarchy cannot be recovered from the expression of the model.
Example $v_5 = \frac{v_6+v_7}{2}$, $v_6 = \frac{v_1+v_2}{2}$ and $v_7 = \frac{v_3+v_4}{2}$.



Assumptions

Notation

Let $k \in V$. For a given CI, we write $\mathcal{S}_k$ the set of subsets of $\mathrm{Children}(k)$ having a non-zero Möbius coefficient.

Assumption H1

At every aggregation node $k \in V$, $\mathrm{Children}(k)$ is the only connected component of graph

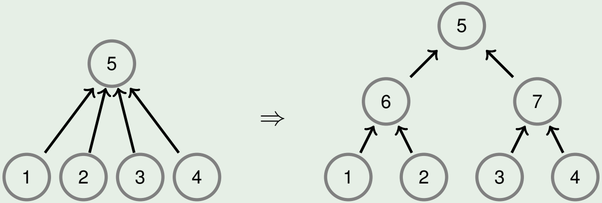
$$\langle \mathrm{Children}(k), \{(i,j)\ ,\ i \neq j \text{ s.t. } \exists S \in \mathcal{S}_k \ :\ \{i,j\} \subseteq S\} \rangle$$

Illustration of H1

Illustration

H1 forbids to have a model C_{m_k} that is (even only partly) additive.

- $v_5 = C_{m_k}(v_1, v_2, v_3, v_4) = \frac{1}{2} v_1 \wedge v_2 + \frac{1}{2} v_3 \wedge v_4$
  - violates H1: $\{1, 2\}$ and $\{3, 4\}$ are disconnected
  - $v_6 = v_1 \wedge v_2$, $v_7 = v_3 \wedge v_4$ and $v_8 = \frac{v_6 + v_7}{2}$ is equivalent

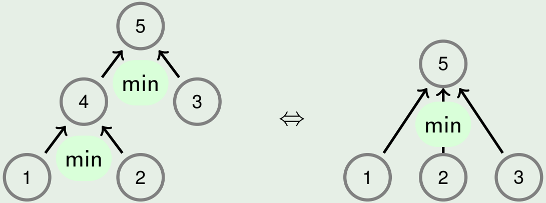


- $C_{m_k}(v_1, v_2, v_3, v_4) = \frac{1}{3} v_1 \wedge v_2 + \frac{1}{3} v_2 \wedge v_3 + \frac{1}{3} v_3 \wedge v_4$ satisfies H1

Assumptions

Counter-example #2

A pure min function.



Assumption H2

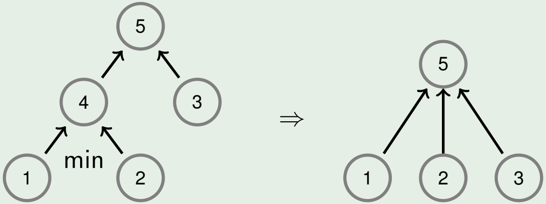
For all nodes $k \in V$:

$$|\mathcal{S}_k| \geq 2.$$

Illustration of H2

H2 (combined with H1) forbids from having a simple min between two variables.

- $v_4 = v_1 \wedge v_2$ (violating H2) and $v_5 = \frac{v_3}{2} + \frac{v_3 \wedge v_4}{2}$
- We can rewrite $v_5 = \frac{v_3}{2} + \frac{v_1 \wedge v_2 \wedge v_3}{2}$



Identifiability result

Identifiability of UHCI and its hierarchy [*Bresson et al*, KR'2021]

Let $\mathcal{F}$ and $\mathcal{F}'$ be two UHCI with potentially different hierarchies, fuzzy measures and marginal utility functions. Assume that both models fulfill H1, H2. Assume, $\forall x \in X,\ \mathcal{F}(x) = \mathcal{F}'(x)$.

Then, both models have the same hierarchy, fuzzy measures and marginal utilities.

Outline

Outline

Neuronal Representation

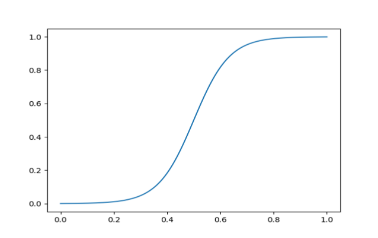
Monotonic Marginal Utility

Conditions on u_i:

- u_i is non-decreasing on X_i
- $\lim_{x_i \to -\infty} u_i(x_i) = 0$
- $\lim_{x_i \to +\infty} u_i(x_i) = 1$

Convex sum of sigmoids:

$$u_i(x_i) = \sum_{k=0}^{p} \frac{r_i^k}{1 + e^{-\left(\eta_i^k x_i - \beta_i^k\right)}} ,$$



With:

- $\sum_{k=1}^{p} r_i^k = 1$ and $\forall k,\ r_i^k \geq 0$
- $\forall k,\ \eta_i^k \geq 0$

Neuronal Representation

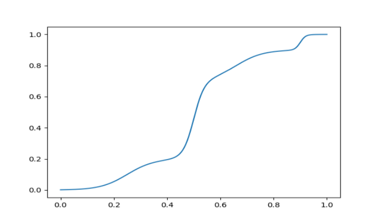
Monotonic Marginal Utility

Conditions on u_i:

- u_i is non-decreasing on X_i
- $\lim_{x_i \to -\infty} u_i(x_i) = 0$
- $\lim_{x_i \to +\infty} u_i(x_i) = 1$

Convex sum of sigmoids:

$$u_i(x_i) = \sum_{k=0}^{p} \frac{r_i^k}{1 + e^{-\left(\eta_i^k x_i - \beta_i^k\right)}},$$



With:

- $\sum_{k=1}^{p} r_i^k = 1$ and $\forall k,\ r_i^k \geq 0$
- $\forall k,\ \eta_i^k \geq 0$

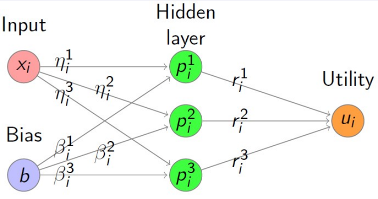
Neuronal Representation

Monotonic Marginal Utility

$$u_i(x_i) = \sum_{k=0}^{p} \frac{r_i^k}{1 + e^{-\left(\eta_i^k x_i - \beta_i^k\right)}} ,$$

where

- $\sum_{k=1}^{p} r_i^k = 1$ and $\forall k,\ r_i^k \geq 0$
- $\forall k,\ \eta_i^k \geq 0$



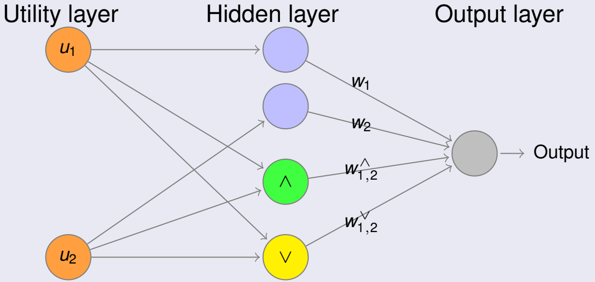
A utility module with 3 hidden nodes ($p = 3$)

Neuronal Representation

Choquet Modules

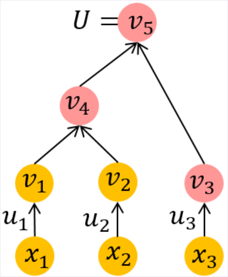
$$C_w(\mathbf{v}) = \sum_{i=1}^{n} w_i v_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(w_{i,j}^{\wedge} (v_i \wedge v_j) + w_{i,j}^{\vee} (v_i \vee v_j) \right)$$

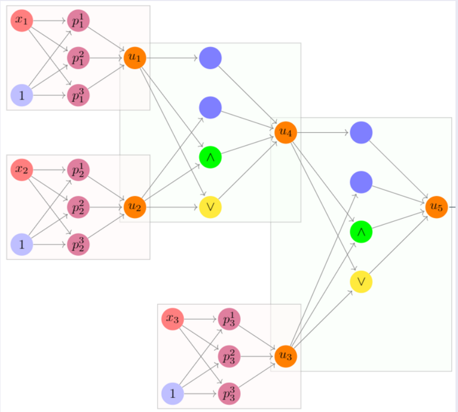
- $\forall i \in N, \forall j \in N,$
  $w_i \geq 0, w_{i,j}^{\wedge} \geq 0, w_{i,j}^{\vee} \geq 0$
- $\sum_{i=1}^{n} w_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} (w_{i,j}^{\wedge} + w_{i,j}^{\vee}) = 1$



Composition of the different parts

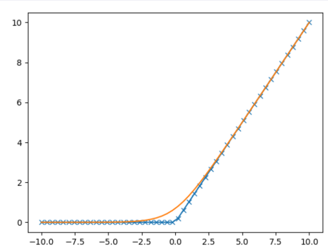
Composition of aggregation and Marginal Utility patterns [*Bresson et al*, IJCAI'2020]





Composition of the different parts

Ensuring Monotonicity and Normalization conditions

	Monotonicity	Normalization
Utility function	*clipping*: $r_i^k \leftarrow max(r_i^k, 0)$	$r_i^k \leftarrow \frac{r_i^k}{\sum_j r_i^j}$
Aggregation	$\begin{array}{cccc} \mathbb{R} & \rightarrow & \mathbb{R}^+ \\ z_i & \mapsto & w_i = \mathrm{softmax}(z_i) \end{array}$ 	$w_i \leftarrow \frac{w_i}{\sum_j w_j}$ $z_i \leftarrow \mathrm{softmax}^{-1}(w_i)$

Outline

Experimental Results - Performance

Dataset	MLP	Logistic Reg.	CUR	NCI	NCI+U	NHCI	NHCI+U
CPU	**0.015 ± 0.021**	0.091±0.051	0.024 ± 0.025	0.045±0.039	0.023±0.024	0.030±0.027	0.023±0.026
CEV	**0.004 ± 0.004**	0.110±0.023	0.084±0.067	0.059±0.012	0.051±0.023	0.035±0.009	0.019±0.017
LEV	**0.135 ± 0.021**	0.161± 0.022	0.143±0.0213	**0.136 ± 0.022**	**0.135 ± 0.019**	N/A	N/A
MPG	0.113 ± 0.036	0.090 ± 0.030	0.112 ± 0.099	0.086 ± 0.027	**0.079 ± 0.027**	0.085 ± 0.029	0.082 ± 0.027
DB	0.143 ± 0.069	0.164± 0.071	0.235 ± 0.017	0.139±0.067	**0.132± 0.068**	0.141 ± 0.068	**0.132 ± 0.066**
MG	0.179 ± 0.028	0.196 ± 0.027	**0.166± 0.022**	0.195 ± 0.027	**0.166 ± 0.026**	0.201 ± 0.030	0.181 ± 0.028
Journal	**0.180 ±0.063**	0.250±0.070	0.218±0.086	0.207±0.065	0.197±0.060	0.219±0.065	0.216±0.062
Boston	0.124 ± 0.030	0.145±0.033	0.1360± 0.085	0.127±0.031	0.129±0.032	**0.121±0.032**	0.129±0.031
Titanic	**0.182 ± 0.025**	0.202 ± 0.027	0.185 ± 0.041	0.192±0.0264	0.193 ± 0.027	0.203±0.027	0.194±0.027

Table 1 Neur-HCI, Classification setting: Classification error (average and variance over 1,000 runs).

Dataset	MLP	Linear Reg.	NCI	NCI+U	NHCI	NHCI+U
CPU	**0.0005 ± 0.0016**	0.0022±0.0019	0.0023±0.0032	0.0009±0.0013	0.0026±0.0023	0.0009±0.0011
CEV	**0.0094 ± 0.003**	0.0434±0.0442	0.0437±0.0037	0.0264±0.0027	0.0197±0.0017	0.0176±0.0017
LEV	0.0312 ± 0.0254	**0.0252±0.0029**	**0.0252±0.0031**	**0.0252±0.0029**	N/A	N/A
MPG	**0.0047 ± 0.0008**	0.0089±0.0019	0.0084±0.0018	0.0056±0.0013	0.0091±0.0018	0.0057±0.0012
Journal	0.0410 ± 0.010	0.0524±0.0128	0.0631±0.0127	**0.0385±0.0112**	0.0629 ± 0.0127	0.0391 ± 0.0117
Boston	0.0079 ± 0.0030	0.0174±0.0038	0.0157 ±0.0037	**0.0072±0.0023**	0.0151 ± 0.0033	0.0077 ± 0.0023

Table 2 Neur-HCI, Regression setting: Mean square error (average and variance over 1,000 runs)

Dataset	MLP	Linear Reg.	NCI	NCI+U	NHCI	NHCI+U
CPU	**0.0005 ± 0.002**	**0.0006 ± 0.003**	0.0007 ± 0.003	**0.0006 ± 0.003**	0.0009 ± 0.003	0.0010 ± 0.004
CEV	0.0174 ± 0.012	0.0642±0.011	0.0243±0.005	0.0099±0.002	0.0165±0.004	**0.0088±0.003**
LEV	**0.0178 ± 0.025**	**0.0179±0.023**	**0.0178 ±0.024**	**0.0177±0.023**	N/A	N/A
MPG	**0.0613 ± 0.012**	0.0642±0.011	**0.0610±0.011**	**0.0612±0.011**	0.0633±0.012	0.0621±0.011
DB	0.1355 ± 0.0796	0.1257±0.079	0.1216±0.081	**0.0942±0.069**	0.1231 ± 0.092	**0.0962 ± 0.081**
MG	0.2601 ± 0.046	0.2661±0.047	0.2668±0.045	**0.2381±0.037**	0.2701±0.052	0.2446 ±0.036
Journal	0.1801 ± 0.064	0.1802±0.065	0.1761±0.063	0.1838±0.066	**0.1711±0.063**	0.1889±0.065
Boston	**0.0659 ± 0.016**	0.0790±0.014	0.0790±0.015	**0.0669±0.012**	0.0752 ± 0.014	0.0681 ± 0.014
Titanic	**0.1521 ± 0.027**	0.1651 ± 0.029	0.1632 ±0.028	**0.1533 ±0.028**	0.166 ± 0.028	**0.1542 ± 0.029**
Arguments 1	0.0157 ± 0.015	0.0195±0.016	0.0145±0.012	**0.0141±0.012**	**0.0141±0.012**	**0.0140±0.012**
Arguments 2	0.0588 ± 0.028	0.0653±0.031	0.0644±0.028	0.0581±0.027	**0.0572±0.027**	**0.0572±0.028**
Arguments 3	**0.0740 ± 0.039**	0.0941±0.042	0.0783±0.040	0.0784±0.040	**0.0761±0.039**	**0.0771±0.041**

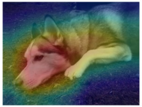
Table 3 Neur-HCI, Ranking setting: percentage of mis-ordered pairs (average and variance

Outline

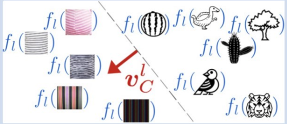
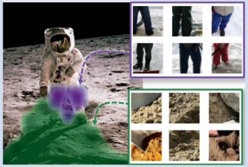
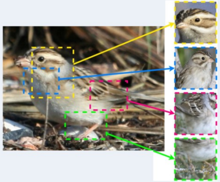
Outline

Main approaches of XAI for Image Processing

Feature Attribution

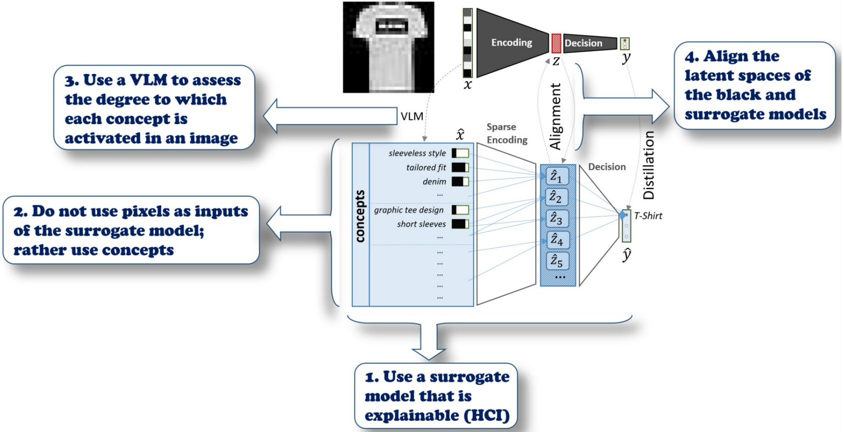
Test Image	
Explanation for class « Siberian Husky »	
Explanation for class « Transverse Flute »	

* *Checkermallo, 2016*

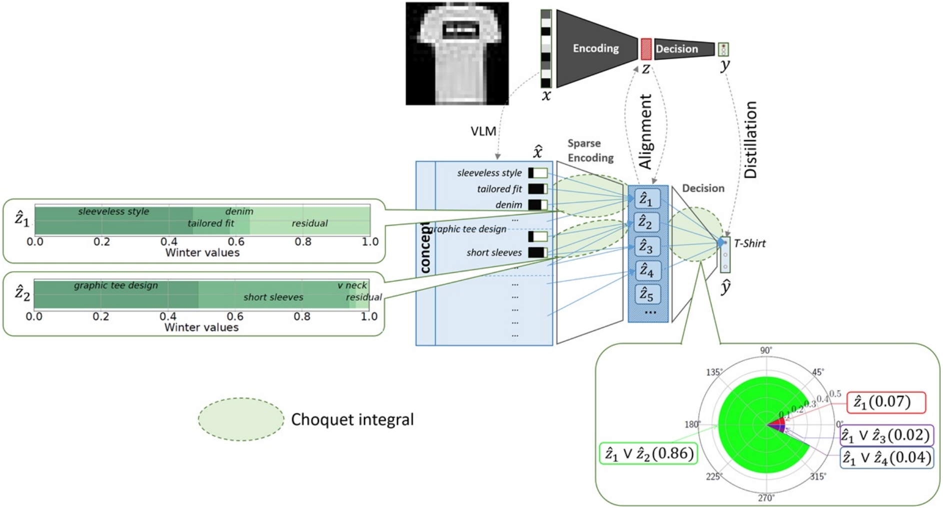
	Explicit Concepts	Implicit Concepts
Explain the Model	 * *Kim et al*. Quantitative testing with concept activation vectors (TCAV). 2018	 * *Fell et al*. CRAFT: Concept Recursive Activation FacTorization for Explainability. 2023
Explanator	?	 * *Chen et al*. This Looks Like that. 2019

Outline

CB2: Cut The Back-Box



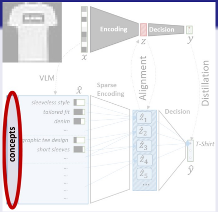
CB2: Cut The Back-Box



CB2: Cut The Back-Box

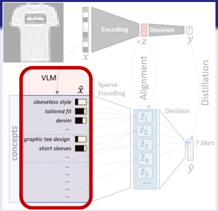
Choice of a set $\mathcal{C}$ of concepts

- Provided by domain expert
- Domain Ontology, *Concept-Net* ontology
- Most frequent words in dictionary



Conceptual Representation

- VLM (Visual Language Model) with pivotal representation
  - ϕ_v : images $\to \mathbb{R}^m$
  - ϕ_t : text $\to \mathbb{R}^m$
- Degree of relevance of concept c in image $\mathbf{x}$: $\hat{x}_c = \dfrac{\langle \phi_v(\mathbf{x}), \phi_t(c) \rangle}{\|\phi_t(c)\|}$

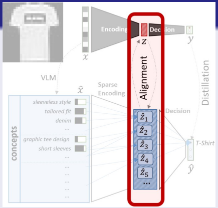


CB2: Cut The Back-Box

Alignment

- $f : z \mapsto \hat{z}$ and $g : \hat{z} \mapsto z$
- Alignment loss:

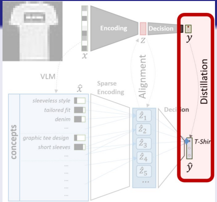
$$\mathcal{L}_{\text{Align}}(H) = \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \left[\|\hat{\mathbf{z}}(\hat{\mathbf{x}}) - f(\mathbf{z}(\mathbf{x}))\|^2 + \|\mathbf{z}(\mathbf{x}) - g(\hat{\mathbf{z}}(\hat{\mathbf{x}}))\|^2 \right]$$



Distillation

- Distillation loss

$$\mathcal{L}_{\text{Dist}}(H) = \sum_{j=1}^{L} \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \left[H_j(\hat{\mathbf{x}}) \log(y_j(x)) + (1 - H_j(\hat{\mathbf{x}})) \log(1 - y_j(x)) \right]$$



Outline

Epilogue

Take-away messages

UHCI model is a good model

- can be learnt from data
  - very versatile Neural Network architecture
- is interpretable
  - hierarchy is uniquely determined
  - explained through pie charts, importance/interaction coefficients
- can be used for image processing
  - as a surrogate model of DL
  - taking as inputs relevant concepts

Epilogue

Some Extensions

- Other models from Decision Theory
  - Generalized Additive Independence
  - MR-Sort
  - ...
- Learn the hierarchy
- Other types of explanations
  - Counterfactuals / Anchors
  - Causality: actual causes

References

- C. Labreuche, S. Fossier. *Explaining Multi-Criteria Decision Aiding Models with an Extended Shapley Value*, **IJCAI'2018**
- C. Labreuche, S. Destercke. *How to handle missing values in Multi-Criteria Decision Aiding?*, **IJCAI'2019**
- R. Bresson, J. Cohen, E. Hullermeier, C. Labreuche, M. Sebag. *Neural Representation and Learning of Hierarchical 2-additive Choquet Integrals*, **IJCAI'2020**
- R. Bresson, J. Cohen, E. Hullermeier, C. Labreuche, M. Sebag. *On the Identifiability of Hierarchical Decision Models*, **KR'2021**
- C. Labreuche.*Explanation with the Winter Value: Efficient Computation for Hierarchical Choquet Integrals*, **IJAR'2022**
- C. Labreuche, R. Bresson. Necessary and Sufficient Explanations of Multi-Criteria Decision Aiding Models, with and without interacting Criteria. conference **xAI'2023**
- N. Atienza, R. Bresson, C. Rousselot, P. Caillou, J. Cohen, C. Labreuche, M. Sebag. Cutting the Black Box: Conceptual Interpretation of a Deep Neural Net with Multi-Modal Embeddings and Multi-Criteria Decision Aid. **IJCAI'2024**