NN-based Hierarchical Choquet Integrals: explainability by design

Michèle Sebag

TAU, CNRS – INRIA – LISN, U. Paris-Saclay

Paris, June 8th, 2022

Joint work: Roman Bresson, Johanne Cohen, Eyke Hüllermeier, Christophe Labreuche

Multi-criteria decision aids

Decision Problems

Sobrie, 2015

- Selection: choosing the best *alternative* in a given set;
- Ranking: ordering the *alternatives* by preference;
- Sorting: assigning to any *alternative* a preference label (e.g. *good*, *bad*...)

Scoring Model

$$\mathcal{F} : X \to \mathbb{R}$$

$$\mathcal{F}(\mathbf{x}) \geq \mathcal{F}(\mathbf{y}) \iff \mathbf{x} \text{ is at least as good as } \mathbf{y}$$

$$\mathcal{F}(\mathbf{x}) > \mathcal{F}(\mathbf{y}) \iff \mathbf{x} \text{ is better than } \mathbf{y}$$

Requirement: Trustability

Jiang et al., 18; Varshney 16

- intelligibility
- unicity of the interpretation = identifiability of the model

Building models

MCDA constraint-based approches

- formally-valid models
- time-expensive (requires an interaction with an expert)
- can lead to global inconsistencies
- vulnerable to errors

Machine Learning

- learn from data
- statistically valid
- (in general) not interpretable

Summary of the talk

- Present MCDA, Hierarchical Choquet Integrals
- Learn HCIs from data
- PhD: identifiability property

Formal background

Notations

$$\text{Variables} x_1, \ldots x_n; \qquad x_i \text{ defined on domain} X_i$$

Marginal Utilities

$$u_i : X_i \mapsto \mathbb{R}$$

A marginal utility denotes a preference relation on the values of a attribute i:

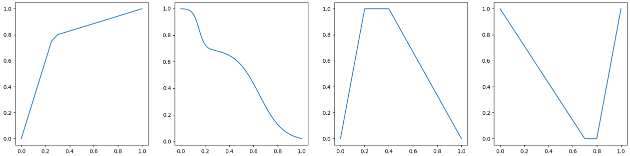
$$u_i(a_i) \geq u_i(b_i) \iff a_i \succeq_i b_i$$

Criterion == attribute i and marginal utility u_i

Marginal utilities

Constraints

- $\max_{x_i \in X_i}(u_i(x_i)) = 1$ (total satisfaction)
- $\min_{x_i \in X_i}(u_i(x_i)) = 0$ (total dissatisfaction)
- u_i is continuous, piecewise-C^1
- u_i is either monotonic or peak-shaped or valley-shaped



Four types of monotonicities

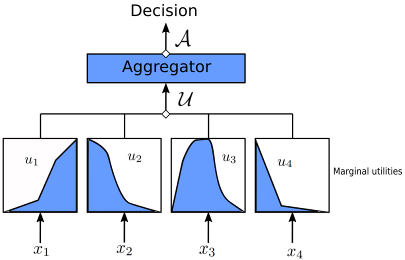
Decomposable Models

Krantz et al. 71

Definition $\mathcal{F}$ is decomposable iff

$$\mathcal{F}(\mathbf{x}) = \mathcal{A}(u_1(x_1), ..., u_n(x_n)) \quad (1)$$

with aggregator $\mathcal{A}$ non-decreasing.



Decomposable model with dimension 4

Example

Houses

House	Surface m^2	Garden m^2	Garage (yes/no)	Road km	Transp. km	Downtown km	Price €
h_1	50	100	No	0.1	0.	0.	400,000
h_2	110	150	Yes	0.5	3.	4.	500,000
h_3	150	0	No	1.	0.5	0.5	450,000
h_4	150	30	No	0.1	5.	3.	300,000
h_5	500	1000	Yes	5.	5.	10.	1,500,000

Notations

- 7 features
- Domains $X_1 = X_2 = X_4 = X_5 = X_6 = X_7 = \mathbb{R}_+$
- Domain $X_3 = \{\text{Yes, No}\}$

Aggregators

Grabish & Perny 03

Weighted Sum of utilities

$$\mathcal{A}(u_1, \ldots u_n) = \sum_{i=i}^{n} w_i a_i \text{ with, } \forall i \in N,\ w_i \geq 0 \text{ and } \sum_{i=i}^{n} w_i = 1.$$

Fuzzy measure μ on set N

$\mu : 2^N \rightarrow [0, 1]$

- $\mu(N) = 1$
- $\mu(\emptyset) = 0$
- $\forall B \subseteq A \subseteq N,\ \mu(B) \leq \mu(A)$

Representation power Let A, B be two non-empty, disjoint **coalitions** of criteria (subsets of N).

- $\mu(A \cup B) \leq \mu(A) + \mu(B)$: redundancy (downtown / public transportation)
- $\mu(A \cup B) = \mu(A) + \mu(B)$: independence (downtown / garden)
- $\mu(A \cup B) \geq \mu(A) + \mu(B)$: synergy (garage / road)

Choquet Integral

Definition — Choquet 54

Given μ, the **Choquet integral** (CI)$(a_1, \ldots, a_n)$ is defined as:

$$C_\mu(a) = \sum_{i=1}^{N} \mu(\{\tau(i), \tau(i+1), ..., \tau(n)\})(a_{\tau(i)} - a_{\tau(i-1)}) \qquad (2)$$

with τ a permutation in N s.t. $\forall i \in N$, $a_{\tau(i)} \leq a_{\tau(i+1)}$ and $a_{\tau(0)} = 0$.

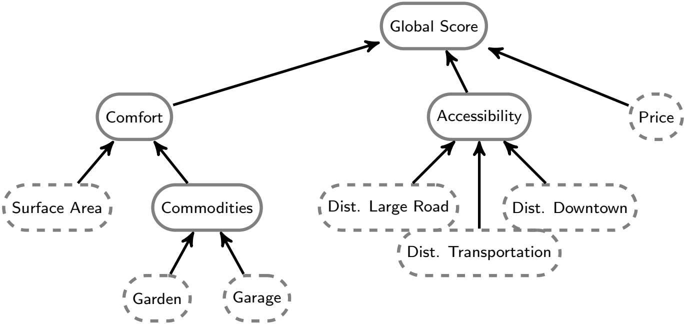
Properties — Grabisch & Labreuche 08

- continuous
- non-decreasing w.r.t. arguments
- piecewise linearity
- interpretable
- 1-Lipschitz

Hierarchical Aggregation

Hierarchical CI: a hierarchical model with CI aggregators.

Example



- $S_{Global} = \mathcal{A}_{Global}(S_{Comfort}, S_{Accessibility}, S_{Price})$
- $S_{Comfort} = \mathcal{A}_{Comfort}(S_{Area}, S_{Commodities})$
- ...

Models

Grabisch 16

Utilitaristic Hierarchical Choquet Integral

$$\mathcal{F} : X = (X_1, \ldots, X_n) \rightarrow [0,1]$$
$$x \mapsto \mathcal{A}(u_1(x_1), ..., u_n(x_n))$$

with $\mathcal{F}$ an HCI and, $\forall i \in N$, u_i a marginal utility.

SoA on Learning MAUT UHCIs

UHCI learning

- U: marginal utilities
- H: hierarchical models
- CI: Choquet integral

U	H	CI	References
		✓	Grabisch et al. 08; Alavi et al. 09 Fallah Tehrani et al. 12; Hüllermeier & Fallah Tehrani 12; Benabbou et al. 17; Havens & Anderson 18; Bourdache et al. 19
✓			Bous & Pirlot 13
✓		✓	Fallah Tehrani et al. 14
✓	✓		Huang et al. 08; Senge & Hüllermeier 11
✓	✓	✓	Bresson et al. 20

Overview

Neur-HCI

- Neural Network framework
- Learns the parameters of a UHCI
  - weights of the CIs
  - marginal utilities
- all UHCI constraints verified by design

Settings

- Regression
- Classification
- Pairwise preferences

Structure

Tree-Hierarchy given by expert

Learning

- Marginal Utilities
  - Monotonic
  - Peak or valleyed-shape
  - Selector
- Aggregators
  - 2-additive
  - $0-1$-based 3-additive
  - general

Representation theorems

- For each type of module, in large sample limit.

Monotonic Marginal Utility

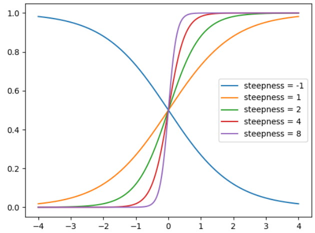
Requirement

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

Search space: a convex sum of sigmoids:

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

Parameters learned : r_i^k, η_i^k, β_i^k.
p is a hyperparameter.



With:

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

Monotonic Marginal Utility

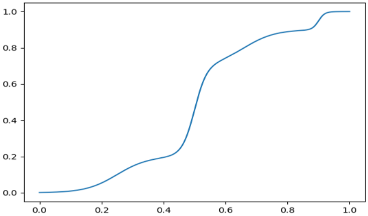
Requirement

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

Search space: a convex sum of sigmoids:

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

Parameters learned : r_i^k, η_i^k, β_i^k.
p is a hyperparameter.



With:

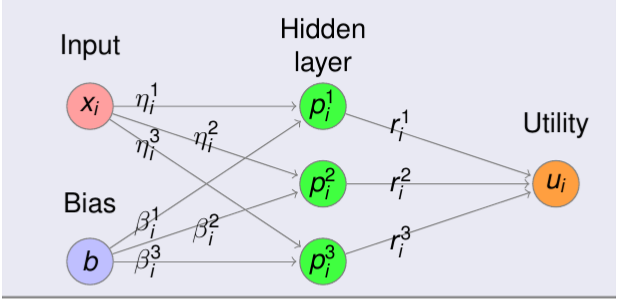
- $\sum_{k=1}^{p} r_i^k = 1$ and $\forall k,\ r_i^k \geq 0$
- $\forall k,\ \eta_i^k \geq 0$

Monotonic Marginal Utility Module

Structure

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

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



Other Marginal Utility Modules

Single-peaked/single-valleyed utilities

- Additional parameter: location of maximum/minimum
- built from monotonic modules

Selectors

- selects automatically the best type of monotonicity
- based on a switching mechanism

2-additive Choquet integral

Characterization

$$C_w(a) = \sum_{i=1}^{n} w_i a_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{i,j}^{\wedge}(a_i \wedge a_j) + \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{i,j}^{\vee}(a_i \vee a_j) \qquad (3)$$

where $\wedge$ (resp $\vee$) denote the min (resp max) operators, with:

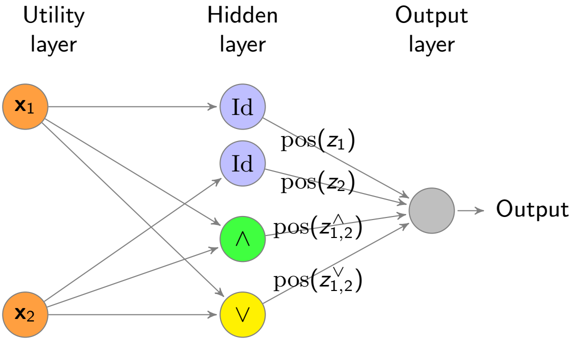
- **Monotonicity:** $\forall i \in N, \forall j \in N, w_i \geq 0, w_{i,j}^{\wedge} \geq 0, w_{i,j}^{\vee} \geq 0$
- **Normalization:** $\sum_{i=1}^{n} w_i + \sum_{1 \leq i < j \leq n} w_{i,j}^{\wedge} + \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{i,j}^{\vee} = 1$

Choquet Modules

2-additive CI

$$C_w(a) = \sum_{i=1}^{n} \mathrm{pos}(z_i)a_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(\mathrm{pos}(z_{i,j}^{\wedge})(a_i \wedge a_j) + \mathrm{pos}(z_{i,j}^{\vee})(a_i \vee a_j)\right)$$

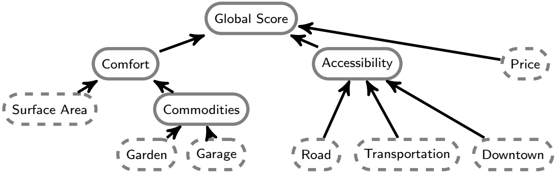
$z_i, z_{i,j}^{\vee}, z_{i,j}^{\wedge} \in \mathbb{R}$; pos non-negative.

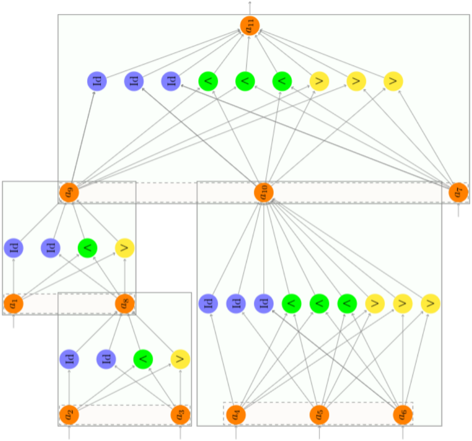


Other Aggregator Models

- A subset of the 3-additive CIs
- General CIs

Assembled Network - HCI





Assembled Network - UHCI

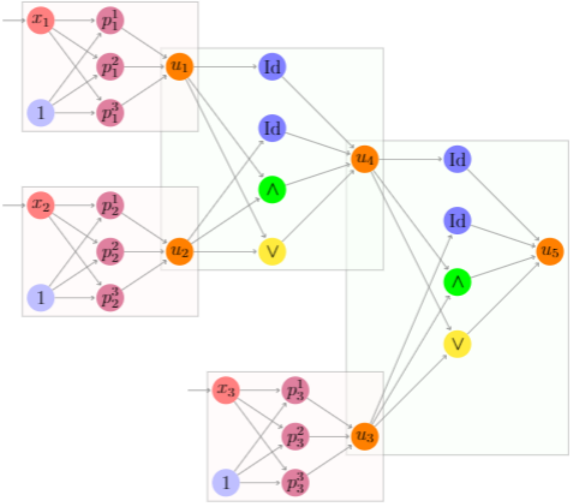


Figure 8.5: Example of the combination of the two modules on a tree with three monotonic marginal utility functions and two aggregation nodes.

Experimental validation on benchmarks

Settings

- Regression (mean squared error)
- Classification (binary) (Misclassification rate)
- Pairwise Learning (Ranking inversion rate)

Experimental setting

- 1000 random splits of each dataset train/test
- Baselines:
  - same-size MLP
  - linear model
  - Choquistic utilitaristic regression (binary classif.)
- Benchmarks: 9 datasets (CPU, CEV, LEV, MPG, DB, MG, Journal, Boston)

Performance on Real Data

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	**0.135 ± 0.068**
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	0.191 ± 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).

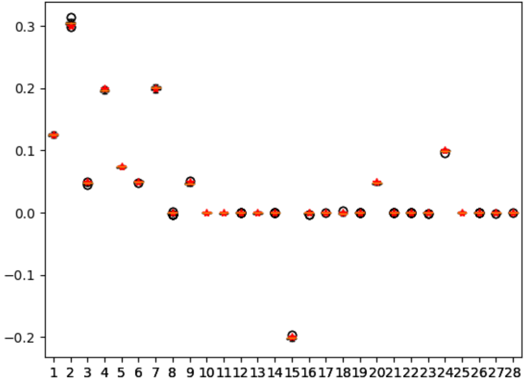
Dataset	MLP	Linear Reg.	NCI	NCI+U	NHCI	NHCI+U
CPU	**0.0005 ± 0.0016**	0.0022±0.0019	0.0023±0.0032	0.0009±0.0013	0.0026±0.0023	0.0009±0.0011
CEV	**0.0094 ± 0.003**	0.0434±0.0442	0.0437±0.0037	0.0264±0.0027	0.0197±0.0017	0.0176±0.0017
LEV	0.0312 ± 0.0254	**0.0252±0.0029**	**0.0252±0.0031**	**0.0252±0.0029**	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 over 1,000 runs)

On artificial data (large sample limit



Distribution of the weights of 50 CIs
trained on 50 different datasets of $10,000$ examples
generated by a ground-truth model (red stars).

Artificial Data, followed

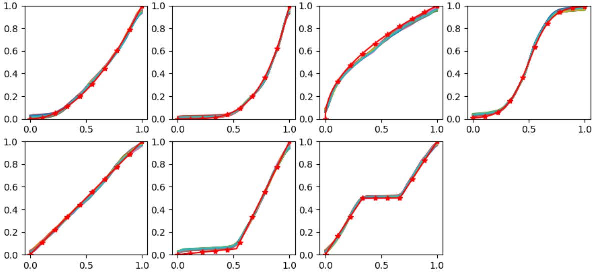


Figure: Marginal utilities of 50 UHCI models trained on 50 different datasets of 10, 000 examples generated by a ground-truth model (red stars).

Sensitivity to Noise

Noise	Tree	Flat
0.	$7.1 \times 10^{-7} \pm 4.4 \times 10^{-6}$	$2.5 \times 10^{-6} \pm 7.2 \times 10^{-7}$
0.05	$1.1 \times 10^{-4} \pm 4.6 \times 10^{-5}$	$1.7 \times 10^{-4} \pm 4.8 \times 10^{-5}$
0.1	$4.4 \times 10^{-4} \pm 1.6 \times 10^{-4}$	$6.3 \times 10^{-4} \pm 1.8 \times 10^{-4}$
0.2	$1.6 \times 10^{-3} \pm 7.6 \times 10^{-4}$	$1.7 \times 10^{-3} \pm 6.1 \times 10^{-4}$

Table: Mean Squared Error in regression setting, 50 models training on 50 different datasets of 300 noisy examples. Tested on noiseless examples, the error is lower than the noise.

Real Data - MPG Dataset

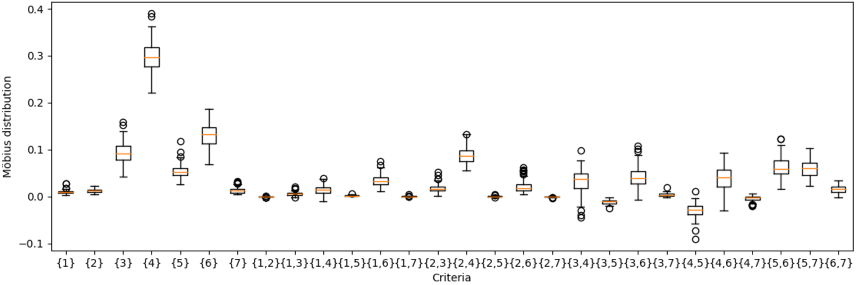


Figure: Distribution of weights, 50 models trained on different splits of the MPG dataset.

Real Data - MPG Dataset

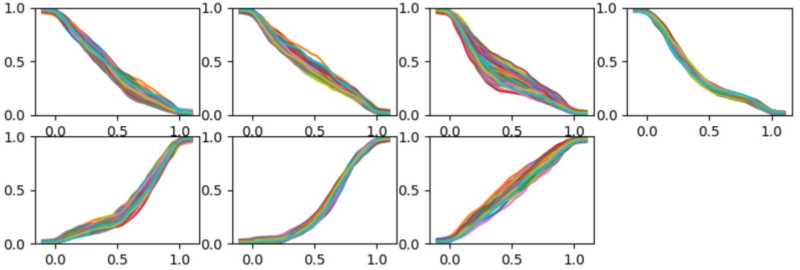


Figure: Marginal utilities, 50 models trained on different splits of the MPG dataset.

Conclusion

- This was made possible as HCI constraints can be included in NN architecture *by design*

Next

- Revise a NN with Hierarchical Choquet head.