Learning Monotone Nonlinear Models using the Choquet Integral

Ali Fallah Tehrani, Weiwei Cheng, Krzysztof Dembczynski, Eyke Hüllermeier

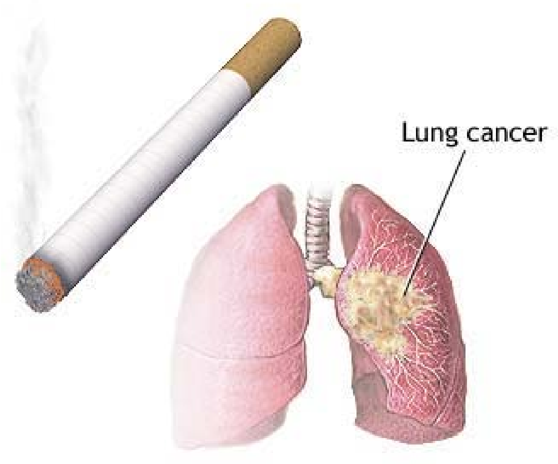
Knowledge Engineering & Bioinformatics Lab
Department of Mathematics and Computer Science
Marburg University, Germany

Philipps Universität Marburg

ECMLPKDD, Athens Greece, September 2011

Monotonicity

Incorporating background knowledge, such as ***monotonicity***, into the learning process is an important aspect in machine learning research.



For example, the higher the tobacco consumption, the more likely a patient suffers a lung cancer.

Monotonicity

For a linear model $Y = \sum_{i=1}^{m} \alpha_i X_i + \epsilon$:

- Monotonicity is easy to ensure (signs of coefficients);
- Easy to interpret. The direction and strength of influence of each predictor are reflected by the corresponding coefficient;
- ***But*** lack of flexibility.

For a nonlinear model, e.g., $Y = \sum_{i=1}^{m} \alpha_i X_i + \sum_{1 \le i < j \le m} \alpha_{ij} X_i X_j + \epsilon$:

- More flexible;
- ***But*** difficult to find simple global constraints to ensure monotonicity, as $\partial Y / \partial X_i = \alpha_i + \sum_{j \neq i} \alpha_{ij} X_j$, which depends on all other attributes;
- Harder to interpret.

Outline of the Talk

Contribution:

- We propose the use of the Choquet integral as a flexible and expressive aggregation operator, which is monotone and provides important insights into the data.
- As an example, we generalize logistic regression using the Choquet integral, leading to choquistic regression.

Outline:

(1) Introduction to non-additive measures and Choquet integral

(2) Choquistic regression as a generalization of logistic regression

(3) First experimental results

Additive & Non-Additive Measures

Let $C = \{c_1, \ldots, c_m\}$ be a finite set and $\mu(\cdot)$ a measure $2^C \to [0,1]$. For each $A \subseteq C$, we interpret $\mu(A)$ as the *weight* of the set A.

C = {speaking Chinese, coding in Java, coding in C}

For an additive measure:

$\mu(A \cup B) = \mu(A) + \mu(B), \ \forall A, B \subseteq C$ such that $A \cap B = \emptyset$.

μ ({speaking Chinese}) = 0.2
μ ({coding in Java}) = 0.4
μ ({coding in C}) = 0.4

μ({speaking Chinese, coding in Java}) = 0.6
μ({speaking Chinese, coding in C}) = 0.6
$\mu(C)$ = 1

A (non-additive) measure is normalized and monotone:

$\mu(\emptyset) = 0, \ \mu(C) = 1$, and $\mu(A) \leq \mu(B) \quad \forall\, A \subseteq B \subseteq C$.

μ ({speaking Chinese}) = 0
μ ({coding in Java}) = 0
μ ({coding in C}) = 0

μ({speaking Chinese, coding in Java}) = 1
μ({speaking Chinese, coding in C}) = 0.7
$\mu(C)$ = 1

Importance of Criteria & Interaction

For an additive measure:

- There is no possibility to model interaction between criteria.
- $\mu(\{c_i\})$ is a natural quantification of the importance of c_i.

For a non-additive measure:

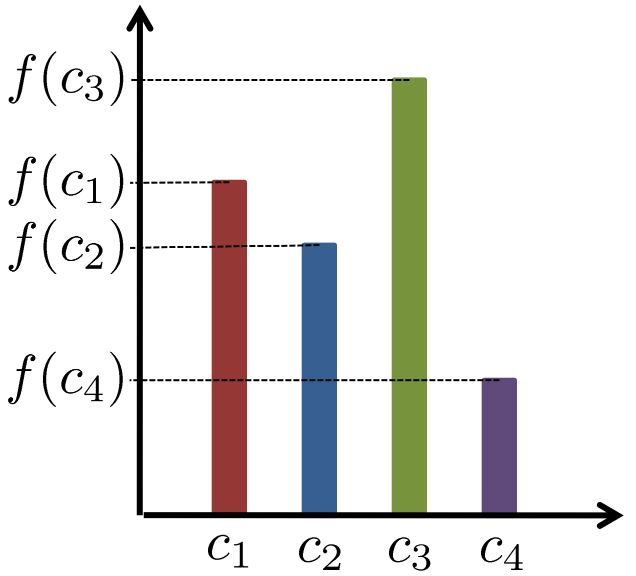
- Importance of criteria can be measured by the Shapley index:

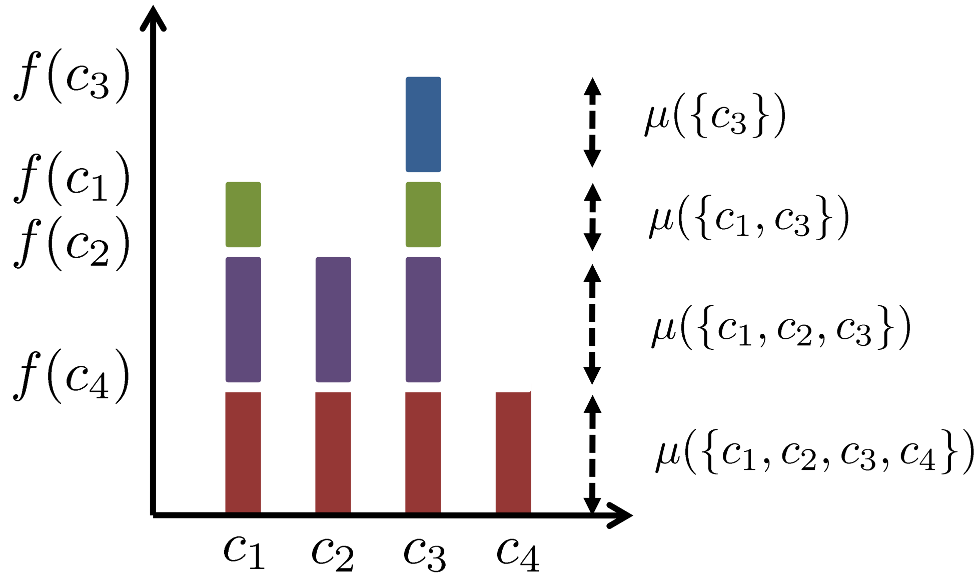
$$\varphi(c_i) = \sum_{A \subseteq C \setminus \{c_i\}} \frac{1}{m \binom{m-1}{|A|}} \Big(\mu(A \cup \{c_i\}) - \mu(A) \Big) .$$

- Interactions between criteria can be measured by the interaction index:

$$I_{i,j} = \sum_{A \subseteq C \setminus \{c_i, c_j\}} \frac{\mu(A \cup \{c_i, c_j\}) - \mu(A \cup \{c_i\}) - \mu(A \cup \{c_j\}) + \mu(A)}{(m-1) \binom{m-2}{|A|}} .$$

Discrete Choquet Integral: A Brief Intro

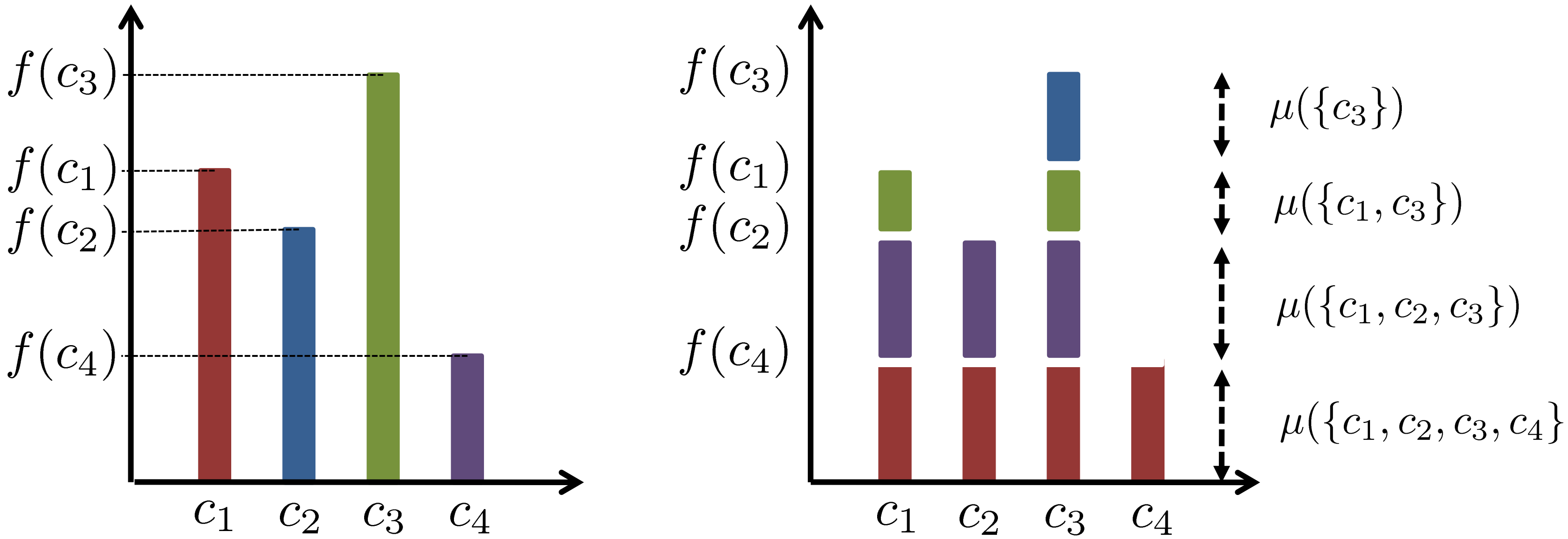




$$\mathcal{C}_\mu(f) = \sum_{i=1}^{4} w_i \cdot f(c_i) = \sum_{i=1}^{4} \mu(\{c_i\}) \cdot f(c_i)$$

$$\mathcal{C}_\mu(f) = \sum_{i=1}^{4} \mu\Big(A_{(i)}\Big) \cdot \Big(f(c_{(i)}) - f(c_{(i-1)})\Big)$$

Discrete Choquet Integral: A Brief Intro



The **discrete Choquet integral** of $f : C \to \mathbb{R}_+$ with respect to μ is defined as follows:

$$\mathcal{C}_\mu(f) = \sum_{i=1}^{m} \left(f(c_{(i)}) - f(c_{(i-1)}) \right) \cdot \mu(A_{(i)}) \ ,$$

where $(\cdot)$ is a permutation of $\{1, \ldots, m\}$ such that $0 \leq f(c_{(1)}) \leq f(c_{(2)}) \leq \ldots \leq f(c_{(m)})$, and $A_{(i)} = \{c_{(i)}, \ldots, c_{(m)}\}$.

In our case, $f(c_i) = x_i$ is the value of the i-th variable.

From Logistic to Choquistic Regression

Logistic

$$\mathbf{P}(y=1\,|\,\boldsymbol{x}) = \Big(1+\exp\big(\boxed{-w_0-\boldsymbol{w}^\top\boldsymbol{x}}\big)\Big)^{-1}$$

Choquistic

$$\mathbf{P}(y=1\,|\,\boldsymbol{x}) = \Big(1+\exp\big(\boxed{-\gamma\,(\mathcal{C}_\mu(\boldsymbol{x})-\beta)}\big)\Big)^{-1}$$

Choquet integral of (normalized) attribute values

- It can be shown that, by choosing the parameters in a proper way, logistic regression is indeed **a special case of choquistic regression**.

Choquistic Regression: Interpretation

Interpretation of choquistic regression as a **two-stage process**:

(1) a (latent) utility degree $u = \mathcal{C}_\mu(\boldsymbol{x}) \in [0,1]$ is determined by the Choquet integral

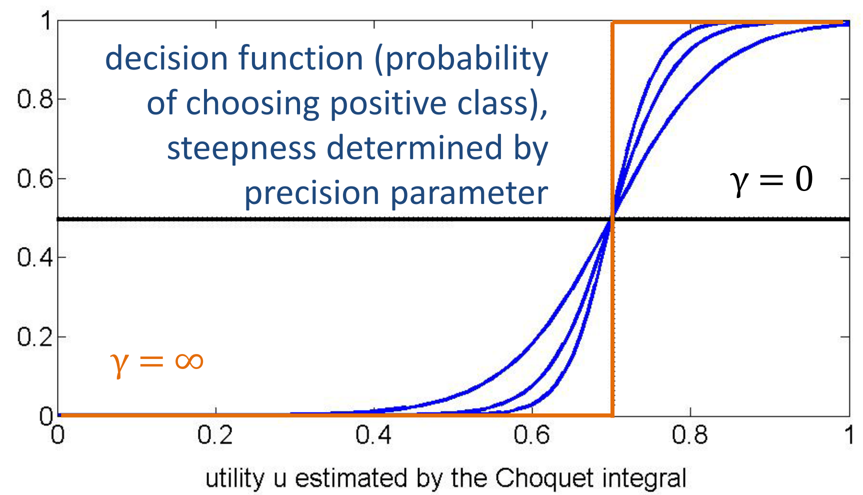
(2) a discrete choice is made by thresholding u at β

Thresholding:

$$\mathbf{P}(y=1) = \frac{1}{1+\exp\left(-\gamma\left(\mathcal{C}_\mu(\boldsymbol{x})-\beta\right)\right)}$$

precision of the model (γ)

utility threshold (β)



Choquistic Regression: Interpretation

- The non-additive measure μ specifies the **importance** of subsets of predictor variables, i.e., their influence on the probability of the positive class.

- Due to the non-additivity of the measure, it becomes possible to model **interaction effects**, thereby expressing complementarity and redundancy of variables.

  For example, what is the **joint effect** of {*smoking,age*} on the probability of cancer, as opposed to the sum of their individual influences?

- Formally, measures like **Shapley index** and **interaction index** can be used, respectively, to quantify the importance of individual and the interaction between different variables.

- **Monotonicity** is obviously ensured by the Choquet integral.

Choquistic Regression: Parameter Estimation

- We need to identify the following model parameters:
  - the non-additive measure μ
  - The utility threshold β
  - The precision parameter γ
- The non-additive measure, in its most general form, has a number of parameters which is exponential in the number of attributes.
  → *critical from a computational complexity point of view*
- We follow a **maximum likelihood** (ML) approach; the Choquet integral is expressed in terms of its **Möbius transform**:

$$\mathcal{C}_\mu(f) = \sum_{T \subseteq C} \boldsymbol{m}(T) \times \min_{c_i \in T} f(c_i) \ .$$

Choquistic Regression: Parameter Estimation

- ML estimation leads to a **constrained optimization problem**:

$$\min_{\boldsymbol{m},\gamma,\beta} \; \gamma \sum_{i=1}^{n} (1 - y^{(i)}) \left(\mathcal{C}_{\boldsymbol{m}}(\boldsymbol{x}^{(i)}) - \beta\right) + \sum_{i=1}^{n} \log\left(1 + \exp(-\gamma\left(\mathcal{C}_{\boldsymbol{m}}(\boldsymbol{x}^{(i)}) - \beta\right))\right)$$

subject to:

$$0 \leq \beta \leq 1$$

$$0 < \gamma$$

conditions on utility threshold and precision

$$\sum_{T \subseteq C} \boldsymbol{m}(T) = 1$$

$$\sum_{B \subseteq A \setminus \{c_i\}} \boldsymbol{m}(B \cup \{c_i\}) \geq 0 \quad \forall A \subseteq C, \forall c_i \in C$$

normalization and monotonicity of the non-additive measure

→ solution with sequential quadratic programming

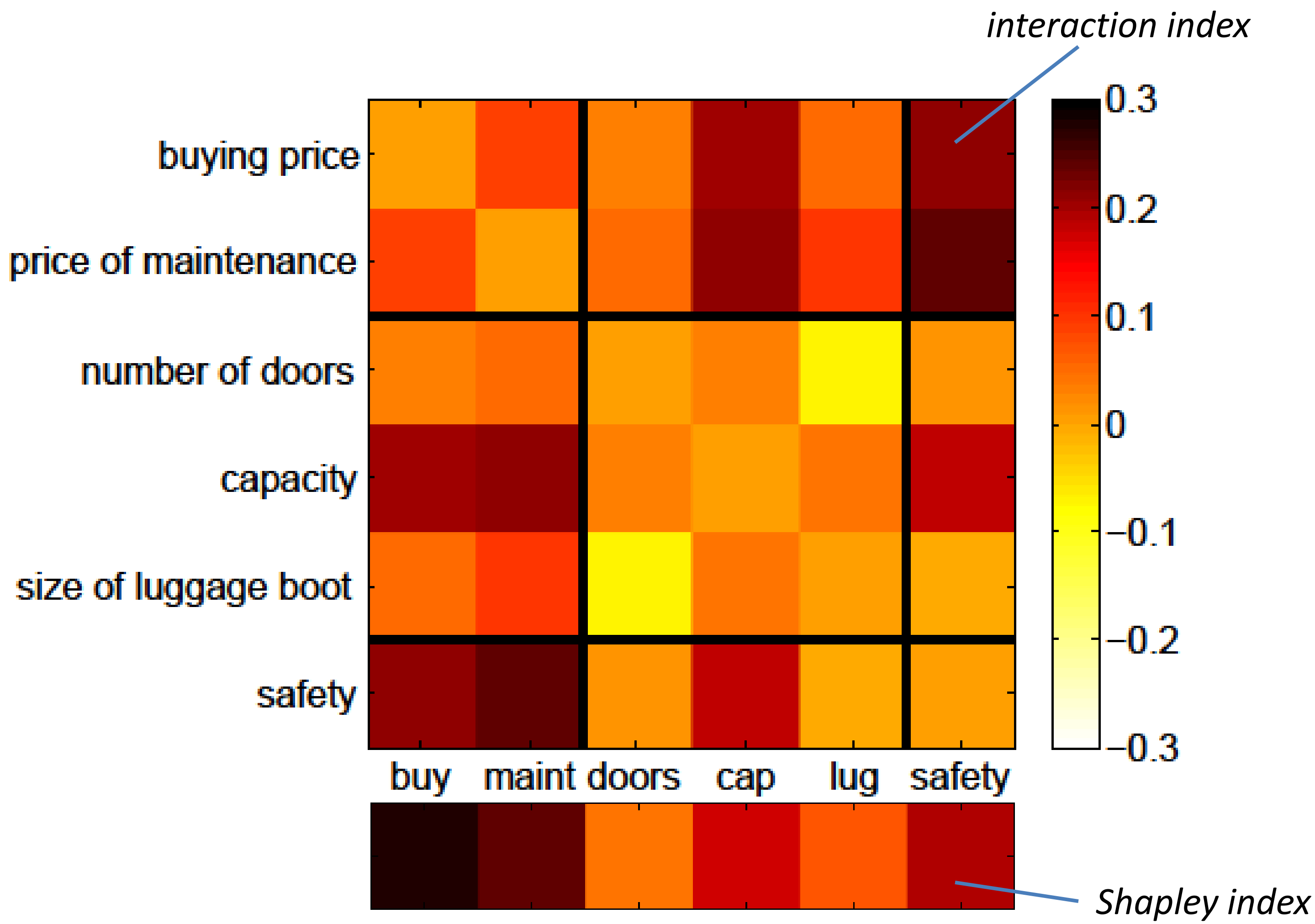
Experimental Evaluation

KE BI

	dataset	CR	LR	KLR-ply	KLR-rbf	MORE
	DBS	.2226±.0380 (4)	.1803±.0336 (1)	.2067±.0447 (3)	.1922±.0501 (2)	.2541±.0142 (5)
	CPU	.0457±.0338 (2)	.0430±.0318 (1)	.0586±.0203 (3)	.0674±.0276 (4)	.1033±.0681 (5)
	BCC	.2939±.0100 (4)	.2761±.0265 (1)	.3102±.0386 (5)	.2859±.0329 (3)	.2781±.0219 (2)
	MPG	.0688±.0098 (2)	.0664±.0162 (1)	.0729±.0116 (4)	.0705±.0122 (3)	.0800±.0198 (5)
20%	ESL	.0764±.0291 (3)	.0747±.0243 (1)	.0752±.0117 (2)	.0794±.0134 (4)	.1035±.0332 (5)
	MMG	.1816±.0140 (3)	.1752±.0106 (2)	.1970±.0095 (4)	.2011±.0123 (5)	.1670±.0120 (1)
	ERA	.2997±.0123 (2)	.2922±.0096 (1)	.3011±.0132 (3)	.3259±.0172 (5)	.3040±.0192 (4)
	LEV	.1527±.0138 (1)	.1644±.0106 (4)	.1570±.0116 (2)	.1577±.0124 (3)	.1878±.0242 (5)
	CEV	.0441±.0128 (1)	.1689±.0066 (5)	.0571±.0078 (3)	.0522±.0085 (2)	.0690±.0408 (4)
	avg. rank	**2.4**	**1.9**	**3.3**	**3.4**	**4**
	DBS	.1560±.0405 (3)	.1443±.0371 (2)	.1845±.0347 (5)	.1628±.0269 (4)	.1358±.0432 (1)
	CPU	.0156±.0135 (1)	.0400±.0106 (3)	.0377±.0153 (2)	.0442±.0223 (5)	.0417±.0198 (4)
	BCC	.2871±.0358 (4)	.2647±.0267 (2)	.2706±.0295 (3)	.2879±.0269 (5)	.2616±.0320 (1)
	MPG	.0641±.0175 (1)	.0684±.0206 (2)	.1462±.0218 (5)	.1361±.0197 (4)	.0700±.0162 (3)
50%	ESL	.0660±.0135 (1)	.0697±.0144 (3)	.0704±.0128 (5)	.0699±.0148 (4)	.0690±.0171 (2)
	MMG	.1736±.0157 (3)	.1710±.0161 (2)	.1859±.0141 (4)	.1900±.0169 (5)	.1604±.0139 (1)
	ERA	.3008±.0135 (3)	.3054±.0140 (4)	.2907±.0136 (1)	.3084±.0152 (5)	.2928±.0168 (2)
	LEV	.1357±.0122 (1)	.1641±.0131 (4)	.1500±.0098 (3)	.1482±.0112 (2)	.1658±.0202 (5)
	CEV	.0346±.0076 (1)	.1667±.0093 (5)	.0357±.0113 (2)	.0393±.0090 (3)	.0443±.0080 (4)
	avg. rank	**2**	**3**	**3.3**	**4.1**	**2.6**
	DBS	.1363±.0380 (2)	.1409±.0336 (4)	.1422±.0498 (5)	.1386±.0521 (3)	.0974±.0560 (1)
	CPU	.0089±.0126 (1)	.0366±.0068 (4)	.0329±.0295 (2)	.0384±.0326 (5)	.0342±.0232 (3)
	BCC	.2631±.0424 (2)	.2669±.0483 (3)	.2784±.0277 (4)	.2937±.0297 (5)	.2526±.0472 (1)
	MPG	.0526±.0263 (1)	.0538±.0282 (2)	.0669±.0251 (4)	.0814±.0309 (5)	.0656±.0248 (3)
80%	ESL	.0517±.0235 (1)	.0602±.0264 (2)	.0654±.0228 (3)	.0718±.0188 (5)	.0657±.0251 (4)
	MMG	.1584±.0255 (2)	.1683±.0231 (3)	.1798±.0293 (4)	.1853±.0232 (5)	.1521±.0249 (1)
	ERA	.2855±.0257 (1)	.2932±.0261 (4)	.2885±.0302 (2)	.2951±.0286 (5)	.2894±.0278 (3)
	LEV	.1312±.0186 (1)	.1662±.0171 (5)	.1518±.0104 (3)	.1390±.0129 (2)	.1562±.0252 (4)
	CEV	.0221±.0091 (1)	.1643±.0184 (5)	.0376±.0091 (3)	.0262±.0067 (2)	.0408±.0090 (4)
	avg. rank	**1.3**	**3.6**	**3.3**	**4.1**	**2.7**

monotone classifier   nonlinear classifier

Importance & Interactions (Car Evaluation)



Conclusions & Outlook

- We advocate the use of the discrete Choquet integral as an aggregation operator in machine learning, especially in learning monotone models.
- As a concrete application, we have proposed choquistic regression, a generalization of logistic regression.
- First experimental results confirm advantages of the Choquet integral.
- **Ongoing work:** Restriction to k-additive measures, for a properly chosen k
  - full flexibility is normally not needed and may even lead to overfitting the data
  - advantages from a computational point of view
  - key question: how to find a suitable k in an efficient way?