Learning Monotone Nonlinear Models Using the Choquet Integral

Ali Fallah Tehrani[1], Weiwei Cheng[1], Krzysztof Dembczyński[1,2], and Eyke Hüllermeier[1]

[1] Mathematics and Computer Science, University of Marburg, Germany
[2] Institute of Computing Science, Poznań University of Technology, Poland
{fallah,cheng,dembczynski,eyke}@mathematik.uni-marburg.de

Abstract. The learning of predictive models that guarantee monotonicity in the input variables has received increasing attention in machine learning in recent years. While the incorporation of monotonicity constraints is rather simple for certain types of models, it may become a more intricate problem for others. By trend, the difficulty of ensuring monotonicity increases with the flexibility or, say, nonlinearity of a model. In this paper, we advocate the so-called Choquet integral as a tool for learning monotone nonlinear models. While being widely used as a flexible aggregation operator in different fields, such as multiple criteria decision making, the Choquet integral is much less known in machine learning so far. Apart from combining monotonicity and flexibility in a mathematically sound and elegant manner, the Choquet integral has additional features making it attractive from a machine learning point of view. Notably, it offers measures for quantifying the importance of individual predictor variables and the interaction between groups of variables. As a concrete application of the Choquet integral, we propose a generalization of logistic regression. The basic idea of our approach, referred to as choquistic regression, is to replace the linear function of predictor variables, which is commonly used in logistic regression to model the log odds of the positive class, by the Choquet integral.

1 Introduction

A proper specification of the type of dependency between a set of predictor (input) variables $X_1, \ldots, X_m$ and the target (output) variable Y is an important prerequisite for successful model induction. The specification of a corresponding hypothesis space imposes an inductive bias that, amongst others, allows for the incorporation of background knowledge in the learning process. An important type of background knowledge is *monotonicity*: Everything else being equal, the increase (decrease) of a certain input variable X_i can only produce an increase in the output variable Y (e.g., a real number in regression, a class in ordered classification, or the probability of the positive class in binary classification). Adherence to this kind of background knowledge may not only be beneficial for model induction, but is often even considered as a hard constraint. For example,

D. Gunopulos et al. (Eds.): ECML PKDD 2011, Part III, LNAI 6913, pp. 414–429, 2011.

no medical doctor will accept a model in which the probability of cancer is *not* monotone increasing in tobacco consumption.

The simplest type of dependency is a linear relationship:

$$Y = \sum_{i=1}^{m} \alpha_i X_i + \epsilon \,, \tag{1}$$

where $\alpha_1, \ldots, \alpha_m$ are real coefficients and ϵ is an error term. Monotonicity can be guaranteed quite easily for (1), since monotonicity in X_i is equivalent to the constraint $\alpha_i \geq 0$. Another important advantage of (1) is its comprehensibility. In particular, the direction and strength of influence of each predictor X_i are directly reflected by the corresponding coefficient α_i.

Perhaps the sole disadvantage of a linear model is its inflexibility and, coming along with this, the supposed absence of any *interaction* between the variables: The effect of an increase of X_i is always the same, namely $\partial Y/\partial X_i = \alpha_i$, regardless of the values of all other attributes. In many real applications, this assumption is not tenable. Instead, more complex, nonlinear models are needed to properly capture the dependencies between the inputs X_i and the output Y.

An increased flexibility, however, typically comes at the price of a loss in terms of the two previous criteria: comprehensibility is hampered, and monotonicity is more difficult to assure. In fact, as soon as an interaction between attributes is allowed, the influence of an increase in X_i may depend on all other variables, too. As a simple example, consider the extension of (1) by the addition of *interaction terms*, a model which is often used in statistics:

$$Y = \sum_{i=1}^{m} \alpha_i X_i + \sum_{1 \leq i < j \leq m} \alpha_{ij} X_i X_j + \epsilon \,. \tag{2}$$

For this model, $\partial Y/\partial X_i$ is given by $\alpha_i + \sum_{j \neq i} \alpha_{ij} X_j$ and depends on the values of *all* other attributes, which means that, depending on the context as specified by these values, the monotonicity condition may change from one case to another. Consequently, it is difficult to find simple *global* constraints on the coefficients that assure monotonicity. For example, assuming that all attributes are non-negative, it is clear that $\alpha_i \geq 0$ and $\alpha_{ij} \geq 0$ for all $1 \leq i \leq j \leq m$ will imply monotonicity. While being sufficient, however, these constraints are non-necessary conditions, and may therefore impose restrictions on the model space that are more far-ranging than desired; besides, negative interactions cannot be modeled in this way. Quite similar problems occur for commonly used nonlinear methods in machine learning, such as neural networks and kernel machines.

In this paper, we advocate the use of the (discrete) Choquet integral as a tool that is interesting in this regard. As will be argued in more detail later on, the Choquet integral combines the aforementioned properties in a quite convenient and mathematically elegant way: By its very nature as an integral, it is a monotone operator, while at the same time allowing for interactions between attributes. Moreover, the existence of natural measures for quantifying the *importance* of individual and the *interaction* between groups of features, it provides important insights into the model, thereby supporting interpretability.

The rest of this paper is organized as follows. In the next section, we give a brief overview of related work. In Section 3, we recall the basic definition of the Choquet integral and some related notions. In Section 4, we propose a generalization of logistic regression in which the Choquet integral is used to model the log odds of the positive class. Experimental results are presented in Section 5, prior to concluding the paper with a few remarks in Section 6.

2 Related Work

As already mentioned, the problem of monotone classification has received increasing attention in the machine learning community in recent years,[1] despite having been introduced in the literature much earlier [1]. Meanwhile, several machine learning algorithms have been modified so as to guarantee monotonicity in attributes, including nearest neighbor classification [2], neural networks [3], decision tree learning [4,5], rule induction [6], as well as methods based on isotonic regression [7] and piecewise linear models [8].

Instead of modifying learning algorithms so as to guarantee monotone models, another idea is to modify the training data. To this end, data pre-processing methods such as re-labeling techniques have been developed. Such methods seek to repair inconsistencies in the training data, so that (standard) classifiers learned on that data will automatically be monotone [9, 10].

Although the Choquet integral has been widely applied as an aggregation operator in multiple criteria decision making [11, 12, 13], it has been used much less in the field of machine learning so far. There are, however, a few notable exceptions. First, the problem of extracting a Choquet integral (or, more precisely, the non-additive measure on which it is defined) in a data-driven way has been addressed in the literature. Essentially, this is a parameter identification problem, which is commonly formalized as a constraint optimization problem, for example using the sum of squared errors as an objective function [14, 15]. To this end, [16] proposed an approach based on the use of quadratic forms, while an alternative heuristic, gradient-based method called HLMS (Heuristic Least Mean Squares) was introduced in [17]. In [18, 19], the Choquet integral is used in the context of ordinal classification. Besides, the Choquet integral has been used as an aggregation operator in the context of ensemble learning, i.e., for combining the predictions of different classifiers [20].

3 The Discrete Choquet Integral

In this section, we given a brief introduction to the (discrete) Choquet integral, which, to the best of our knowledge, is not widely known in the field of machine learning so far. Since the Choquet integral can be seen as a generalization of the standard (Lebesque) integral to the case of non-additive measures, we start with a reminder of this type of measure.

[1] For example, a workshop on "Learning Monotone Models from Data" was organized at ECML/PKDD 2009 in Bled, Slovenia.

3.1 Non-additive Measures

Let $C = \{c_1, \ldots, c_m\}$ be a finite set and $\mu : 2^C \to [0, 1]$ a measure. For each $A \subseteq C$, we interpret $\mu(A)$ as the *weight* or, say, the *importance* of the set of elements A. As an illustration, one may think of C as a set of criteria (binary features) relevant for a job, like "speaking French" and "programming Java", and of $\mu(A)$ as the evaluation of a candidate satisfying criteria A (and not satisfying $C \backslash A$). The term "criterion" is indeed often used in the decision making literature, where it suggests a monotone "the higher the better" influence.

A standard assumption on a measure $\mu(\cdot)$, which is, for example, at the core of probability theory, is additivity: $\mu(A \cup B) = \mu(A) + \mu(B)$ for all $A, B \subseteq C$ such that $A \cap B = \emptyset$. Unfortunately, additive measures cannot model any kind of interaction between elements: Extending a set of elements A by a set of elements B always increases the weight $\mu(A)$ by the weight $\mu(B)$, regardless of A and B.

Suppose, for example, that the elements of two sets A and B are *complementary* in a certain sense. For instance, $A = \{\texttt{French}, \texttt{Spanish}\}$ and $B = \{\texttt{Java}\}$ could be seen as complementary, since both language skills and programming skills are important for the job. Formally, this can be expressed in terms of a positive interaction: $\mu(A \cup B) > \mu(A) + \mu(B)$. In the extreme case, when language skills and programming skills are indeed essential, $\mu(A \cup B)$ can be high although $\mu(A) = \mu(B) = 0$ (suggesting that a candidate lacking either language or programming skills is completely unacceptable). Likewise, elements can interact in a negative way: If two sets A and B are partly *redundant* or *competitive*, then $\mu(A \cup B) < \mu(A) + \mu(B)$. For example, $A = \{\texttt{C}, \texttt{C\#}\}$ and $B = \{\texttt{Java}\}$ might be seen as redundant, since one programming language does in principle suffice.

The above considerations motivate the use of non-additive measures, also called capacities or fuzzy measures, which are simply normalized and monotone [21]:

$$\mu(\emptyset) = 0, \mu(C) = 1 \quad \text{and} \quad \mu(A) \leq \mu(B) \text{ for all } A \subseteq B \subseteq C \; . \tag{3}$$

A useful representation of non-additive measures, that we shall explore later on for learning Choquet integrals, is in terms of the *Möbius transform*:

$$\mu(B) = \sum_{A \subseteq B} \boldsymbol{m}(A) \tag{4}$$

for all $B \subseteq C$, where the Möbius transform $\boldsymbol{m}_\mu$ of the measure μ is defined as follows:

$$\boldsymbol{m}_\mu(A) = \sum_{B \subseteq A} (-1)^{|A|-|B|} \mu(B) \; . \tag{5}$$

The value $\boldsymbol{m}_\mu(A)$ can be interpreted as the weight that is *exclusively* allocated to A, instead of being indirectly connected with A through the interaction with other subsets.

A measure μ is said to be k-order additive, or simply k-additive, if k is the smallest integer such that $\boldsymbol{m}(A) = 0$ for all $A \subseteq C$ with $|A| > k$. This property is interesting for several reasons. First, as can be seen from (4), it means that a measure μ can formally be specified by significantly fewer than 2^m values, which

are needed in the general case. Second, k-additivity is also interesting from a semantic point of view: As will become clear in the following, this property simply means that there are no interaction effects between subsets $A, B \subseteq C$ whose cardinality exceeds k.

3.2 Importance of Criteria and Interaction

An additive (i.e., k-additive with $k = 1$) measure μ can be written as follows:

$$\mu(A) = \sum_{c_i \in A} w_i \ ,$$

with $w_i = \mu(\{c_i\})$ the weight of c_i. Due to (3), these weights are non-negative and such that $\sum_{i=1}^{m} w_i = 1$. In this case, there is obviously no interaction between the criteria c_i, i.e., the influence of a c_i on the value of μ is independent of the presence or absence of any other c_j. Besides, the weight w_i is a natural quantification of the *importance* of c_i.

Measuring the importance of a criterion c_i becomes obviously more involved when μ is non-additive. Besides, one may then also be interested in a measure of *interaction* between the criteria, either pairwise or even of a higher order. In the literature, measures of that kind have been proposed, both for the importance of single as well as the interaction between several criteria.

Given a fuzzy measure μ on C, the *Shaply value* (or importance index) of c_i is defined as a kind of average increase in importance due to adding c_i to another subset $A \subset C$:

$$\varphi(c_i) = \sum_{A \subseteq C \setminus \{c_i\}} \frac{1}{m \binom{m-1}{|A|}} \Big(\mu(A \cup \{c_i\}) - \mu(A)\Big) \ . \tag{6}$$

The Shaply value of μ is the vector $\varphi(\mu) = (\varphi(c_1), \ldots, \varphi(c_m))$. One can show that $0 \leq \varphi(c_i) \leq 1$ and $\sum_{i=1}^{m} \varphi(c_i) = 1$. Thus, $\varphi(c_i)$ is a measure of the *relative* importance of c_i. Obviously, $\varphi(c_i) = \mu(\{c_i\})$ if μ is additive.

The *interaction index* between criteria c_i and c_j, as proposed by Murofushi and Soneda [22], is defined as follows:

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

This index ranges between -1 and 1 and indicates a positive (negative) interaction between criteria c_i and c_j if $I_{i,j} > 0$ ($I_{i,j} < 0$). The interaction index can also be expressed in terms of the Möbius transform:

$$I_{i,j} = \sum_{K \subseteq C \setminus \{c_i, c_j\}, |K| = k} \frac{1}{k+1} \boldsymbol{m}\Big(\{c_i, c_j\} \cup K\Big) \ .$$

Furthermore, as proposed by Grabisch [23], the definition of interaction can be extended to more than two criteria, i.e., to subsets $T \subseteq C$:

$$I_T = \sum_{k=0}^{m-|T|} \frac{1}{k+1} \sum_{K \subseteq C \setminus T, |K|=k} \boldsymbol{m}\Big(T \cup K\Big) \; .$$

3.3 The Choquet Integral

So far, the criteria c_i were simply considered as binary features, which are either present or absent. Mathematically, $\mu(A)$ can thus also be seen as an *integral* of the indicator function of A, namely the function f_A given by $f_A(c) = 1$ if $c \in A$ and $= 0$ otherwise. Now, suppose that $f : C \to \mathbb{R}_+$ is any non-negative function that assigns a *value* to each criterion c_i; for example, $f(c_i)$ might be the degree to which a candidate satisfies criterion c_i. An important question, then, is how to *aggregate* the evaluations of individual criteria, i.e., the values $f(c_i)$, into an overall evaluation, in which the criteria are properly weighted according to the measure μ. Mathematically, this overall evaluation can be considered as an integral $\mathcal{C}_\mu(f)$ of the function f with respect to the measure μ.

Indeed, if μ is an additive measure, the standard integral just corresponds to the *weighted mean*

$$\mathcal{C}_\mu(f) = \sum_{i=1}^{m} w_i \cdot f(c_i) = \sum_{i=1}^{m} \mu(\{c_i\}) \cdot f(c_i) \; , \tag{7}$$

which is a natural aggregation operator in this case. A non-trivial question, however, is how to generalize (7) in the case where μ is non-additive.

This question, namely how to define the integral of a function with respect to a non-additive measure (not necessarily restricted to the discrete case), is answered in a satisfactory way by the Choquet integral, which has first been proposed for additive measures by Vitali [24] and later on for non-additive measures by Choquet [25]. The point of departure of the Choquet integral is an alternative representation of the "area" under the function f, which, in the additive case, is a natural interpretation of the integral. Roughly speaking, this representation decomposes the area in a "horizontal" instead of a "vertical" manner, thereby making it amenable to a straightforward extension to the non-additive case. More specifically, note that the weighted mean can be expressed as follows:

$$\begin{aligned} \sum_{i=1}^{m} f(c_i) \cdot \mu(\{c_i\}) &= \sum_{i=1}^{m} \Big(f(c_{(i)}) - f(c_{(i-1)}\Big) \cdot \Big(\mu(\{c_{(i)}\}) + \ldots + \mu(\{c_{(n)}\})\Big) \\ &= \sum_{i=1}^{m} \Big(f(c_{(i)}) - f(c_{(i-1)}\Big) \cdot \mu\Big(A_{(i)}\Big) \; , \end{aligned}$$

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 $f(c_{(0)}) = 0$ by definition), and $A_{(i)} = \{c_{(i)}, \ldots, c_{(m)}\}$; see Fig. 1 as an illustration.

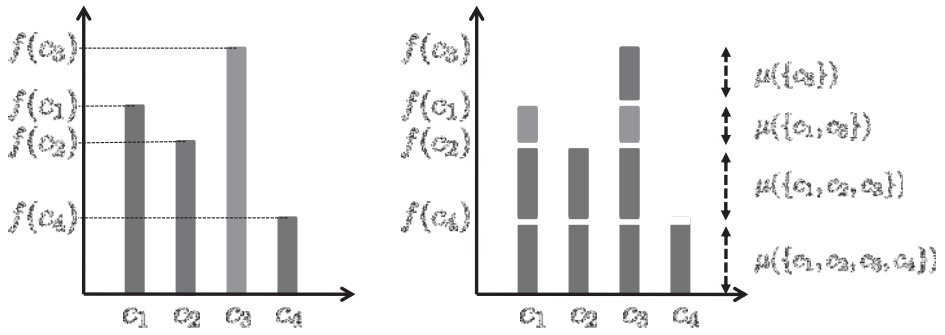

Fig. 1. Vertical (left) versus horizontal (right) integration. In the first case, the height of a single bar, $f(c_i)$, is multiplied with its "width" (the weight $\mu(\{c_i\})$), and these products are added. In the second case, the height of each horizontal section, $f(c_{(i)}) - f(c_{(i-1)})$, is multiplied with the corresponding "width" $\mu(A_{(i)})$.

Now, the key difference between the left and right-hand side of the above expression is that, whereas the measure μ is only evaluated on single elements c_i on the left, it is evaluated on *subsets* of elements on the right. Thus, the right-hand side suggests an immediate extension to the case of non-additive measures, namely the Choquet integral, which, in the discrete case, is formally defined as follows:

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

In terms of the Möbius transform of μ, the Choquet integral can also be expressed as follows:

$$\begin{aligned}
\mathcal{C}_\mu(f) &= \sum_{i=1}^{m} \big(f(c_{(i)}) - f(c_{(i-1)})\big) \cdot \mu(A_{(i)}) \\
&= \sum_{i=1}^{m} f(c_{(i)}) \cdot (\mu(A_{(i)}) - \mu(A_{(i+1)})) \\
&= \sum_{i=1}^{m} f(c_{(i)}) \sum_{R \subseteq T_{(i)}} \boldsymbol{m}(R) \\
&= \sum_{T \subseteq C} \boldsymbol{m}(T) \times \min_{i \in T} f(c_i) \qquad (8)
\end{aligned}$$

where $T_{(i)} = \big\{S \cup \{c_{(i)}\} \,|\, S \subset \{c_{(i+1)}, \ldots, c_{(m)}\}\big\}$.

4 Choquistic Regression

Consider the standard setting of binary classification, where the goal is to predict the value of an output (response) variable $y \in \mathcal{Y} = \{0, 1\}$ for a given instance

$$\boldsymbol{x} = (x_1, \ldots, x_m) \in \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_m$$

represented in terms of a feature vector. More specifically, the goal is to learn a classifier $\mathcal{L} : \mathcal{X} \to \mathcal{Y}$ from a given set of (i.i.d.) training data

$$\mathcal{D} = \left\{ (\boldsymbol{x}^{(i)}, y^{(i)}) \right\}_{i=1}^{n} \subset (\mathcal{X} \times \mathcal{Y})^n$$

so as to minimize the risk

$$R(\mathcal{L}) = \int_{\mathcal{X} \times \mathcal{Y}} \ell(\mathcal{L}(\boldsymbol{x}), y) \, d\mathbf{P}_{XY}(\boldsymbol{x}, y) \ ,$$

where $\ell(\cdot)$ is a loss function (e.g., the simple 0/1 loss given by $\ell(\hat{y}, y) = 0$ if $\hat{y} = y$ and $= 1$ if $\hat{y} \neq y$).

Logistic regression is a well-established statistical method for (probabilistic) classification [26]. Its popularity is due to a number of appealing properties, including monotonicity and comprehensibility: Since the model is essentially *linear* in the input attributes, the strength of influence of each predictor is directly reflected by the corresponding regression coefficient. Moreover, the influence of each attribute is *monotone* in the sense that an increase of the value of the attribute can only increase (decrease) the probability of the positive class.

Formally, the probability of the positive class (and hence of the negative class) is modeled as a generalized linear function of the input attributes, namely in terms of the logarithm of the probability ratio:

$$\log \left(\frac{\mathbf{P}(y = 1 \,|\, \boldsymbol{x})}{\mathbf{P}(y = 0 \,|\, \boldsymbol{x})} \right) = w_0 + \boldsymbol{w}^\top \boldsymbol{x} \ , \tag{9}$$

where $\boldsymbol{w} = (w_1, w_2, \ldots, w_m) \in \mathbb{R}^m$ is a vector of regression coefficients and $w_0 \in \mathbb{R}$ a constant bias (the intercept). A positive regression coefficient $w_i > 0$ means that an increase of the predictor variable x_i will increase the probability of a positive response, while a negative coefficient implies a decrease of this probability. Besides, the larger the absolute value $|w_i|$ of the regression coefficient, the stronger the influence of x_i.

Since $\mathbf{P}(y = 0 \,|\, \boldsymbol{x}) = 1 - \mathbf{P}(y = 1 \,|\, \boldsymbol{x})$, a simple calculation yields the posterior probability

$$\pi_l \stackrel{\text{df}}{=} \mathbf{P}(y = 1 \,|\, \boldsymbol{x}) = \Big(1 + \exp(-w_0 - \boldsymbol{w}^\top \boldsymbol{x}) \Big)^{-1} \ . \tag{10}$$

The logistic function $z \mapsto (1 + \exp(-z))^{-1}$, which has a sigmoidal shape, is a specific type of *link function*.

Needless to say, the linearity of the above model is a strong restriction from a learning point of view, and the possibility of interactions between predictor variables has of course also been noticed in the statistical literature [27]. A standard way to handle such interaction effects is to add interaction terms to the linear function of predictor variables, like in (2). As explained earlier, however, the aforementioned advantages of logistic regression will then be lost.

In the following, we therefore propose an extension of logistic regression that allows for modeling nonlinear relationships between input and output variables while preserving the advantages of comprehensibility and monotonicity.

4.1 The Choquistic Model

In order to model nonlinear dependencies between predictor variables and response, and to take interactions between predictors into account, we propose to extend the logistic regression model by replacing the linear function $\boldsymbol{x} \mapsto w_0 + \boldsymbol{w}^\top \boldsymbol{x}$ in (9) by the Choquet integral. More specifically, we propose the following model

$$\pi_c \stackrel{\text{df}}{=} \mathbf{P}(y = 1 \,|\, \boldsymbol{x}) = \Big(1 + \exp(-\gamma\,(\mathcal{C}_\mu(f_{\boldsymbol{x}}) - \beta))\Big)^{-1}, \tag{11}$$

where $\mathcal{C}_\mu(f_{\boldsymbol{x}})$ is the Choquet integral (with respect to the measure μ) of the function $f_{\boldsymbol{x}} : \{c_1, \ldots, c_m\} \to [0, 1]$ that maps each attribute c_i to a normalized value $x_i = f_{\boldsymbol{x}}(c_i) \in [0, 1]$; $\beta, \gamma \in \mathbb{R}$ are constants.

The normalization is meant to turn each predictor variable into a criterion, i.e., a "the higher the better" attribute, and to assure commensurability between the criteria [28]. A simple transformation, that we shall also employ in our experimental studies, is given by the mapping $z_i = (x_i - m_i)/(M_i - m_i)$, where m_i and M_i are lower and upper bounds for x_i (perhaps estimated from the data); if the influence of x_i is actually negative (i.e., $w_i < 0$), then the mapping $z_i = (M_i - x_i)/(M_i - m_i)$ is used instead.

In order to verify that our model (11) is a proper generalization of standard logistic regression, recall that the Choquet integral reduces to a weighted mean (7) in the special case of an additive measure μ. Moreover, consider any linear function $\boldsymbol{x} \mapsto g(\boldsymbol{x}) = w_0 + \boldsymbol{w}^\top \boldsymbol{x}$ with $\boldsymbol{w} = (w_1, \ldots, w_m)$. This function can also be written in the form

$$\begin{aligned} g(\boldsymbol{x}) &= w_0 + \sum_{i=1}^{m} (w_i p_i + |w_i|(M_i - m_i) z_i) \\ &= w_0 + \sum_{i=1}^{m} w_i p_i + \sum_{i=1}^{m} |w_i|(M_i - m_i) z_i \\ &= w_0' + \left(\sum_{i=1}^{m} u_i\right)^{-1} \sum_{i=1}^{m} u_i' z_i \\ &= \gamma \left(\sum_{i=1}^{m} u_i' z_i - \beta\right), \end{aligned}$$

where $p_i = m_i$ if $w_i \geq 0$ and $p_i = M_i$ if $w_i < 0$, $u_i = |w_i|(M_i - m_i)$, $\gamma = (\sum_{i=1}^m u_i)^{-1}$, $u_i' = u_i/\gamma$, $w_0' = w_0 + \sum_{i=1}^m w_i p_i$, $\beta = -w_0'/\gamma$. By definition, the u_i' are non-negative and sum up to 1, which means that $\sum_{i=1}^m u_i' z_i$ is a weighted mean of the z_i that can be represented by a Choquet integral.

Recalling the idea of "evaluating" an instance $\boldsymbol{x}$ in terms of a set of criteria, the model (11) can be seen as a two-step procedure: The first step consists of an assessment of $\boldsymbol{x}$ in terms of a (latent) utility degree

$$u = U(\boldsymbol{x}) = \mathcal{C}_\mu(f_{\boldsymbol{x}}) \in [0, 1].$$

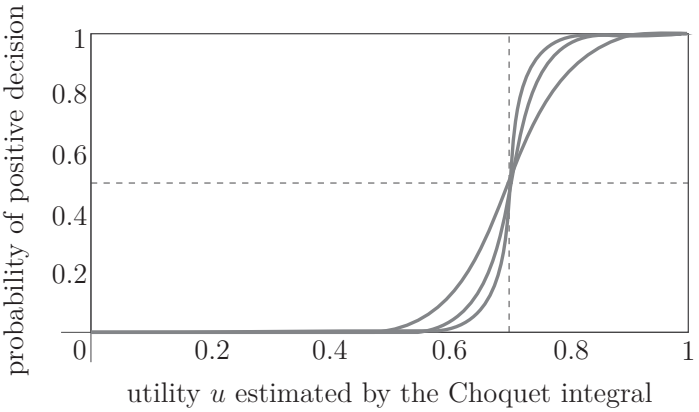

Fig. 2. Probability of a positive decision, $\mathbf{P}(y = 1 \,|\, \boldsymbol{x})$, as a function of the estimated degree of utility, $u = U(\boldsymbol{x})$, for a threshold $\beta = 0.7$ and different values of γ

Then, in a second step, a discrete choice (yes/no decision) is made on the basis of this utility. Roughly speaking, this is done through a "probabilistic thresholding" at the utility threshold β. If $U(\boldsymbol{x}) > \beta$, then the decision tends to be positive, whereas if $U(\boldsymbol{x}) < \beta$, it tends to be negative. The precision of this decision is determined by the parameter γ (see Fig. 2): For large γ, the decision function converges toward the step function $u \mapsto \mathbb{I}(u > \beta)$, jumping from 0 to 1 at β. For small γ, this function is smooth, and there is a certain probability to violate the threshold rule $u \mapsto \mathbb{I}(u > \beta)$. This might be due to the fact that, despite being important for decision making, some properties of the instances to be classified are not captured by the utility function. In that case, the utility $U(\boldsymbol{x})$, estimated on the basis of the given attributes, is not a perfect predictor for the decision eventually made. Thus, the parameter γ can also be seen as an indicator of the quality of the classification model.

4.2 Parameter Estimation

The model (11) has several degrees of freedom: The fuzzy measure μ (Möbius transform $\boldsymbol{m} = \boldsymbol{m}_\mu$) determines the (latent) utility function, while the utility threshold β and the scaling parameter γ determine the discrete choice model. The goal of learning is to identify these degrees of freedom on the basis of the training data $\mathcal{D}$. Like in the case of standard logistic regression, it is possible to harness the maximum likelihood (ML) principle for this purpose. The log-likelihood of the parameters can be written as

$$\begin{aligned} l(\boldsymbol{m}, \gamma, \beta) &= \log \mathbf{P}(\mathcal{D} \,|\, \boldsymbol{m}, \beta, \gamma) \\ &= \log \left(\prod_{i=1}^{n} \mathbf{P}(y^{(i)} \,|\, \boldsymbol{x}^{(i)}; \boldsymbol{m}, \beta, \gamma) \right) \\ &= \sum_{i=1}^{n} y^{(i)} \log \pi_c^{(i)} + \left(1 - y^{(i)}\right) \log \left(1 - \pi_c^{(i)}\right) . \end{aligned} \tag{12}$$

One easily verifies that (12) is convex with respect to $\boldsymbol{m}, \gamma$, and β. In principle, maximization of the log-likelihood can be accomplished by means of standard gradient-based optimization methods. However, since we have to assure that μ is a proper fuzzy measure and, hence, that $\boldsymbol{m}$ guarantees the corresponding monotonicity and boundary conditions, we actually need to solve a *constrained* optimization problem:

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

$$\text{s.t.} \quad \gamma > 0,\ 0 \leq \beta \leq 1,\ \sum_{T \subseteq C} \boldsymbol{m}(T) = 1, \ \text{ and}$$

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

A solution to this problem can be produced by standard solvers. Concretely, we used the `fmincon` function implemented in the optimization toolbox of Matlab. This method is based on a sequential quadratic programming approach.

Recall that, once the model has been identified, the importance of each attribute and the degree of interaction between groups of attributes can be derived from the Möbius transform $\boldsymbol{m}$; these are given, respectively, by the Shapley value and the interaction indexes as introduced in Section 3.2.

5 Experiments

5.1 Data Sets

Although the topic is receiving increasing interest in the machine learning community, benchmark data for monotone classification is by far not as abundant as for conventional classification. In total, we managed to collect 9 data sets from different sources, notably the UCI repository[2] and the WEKA machine learning framework [29], for which monotonicity in the input variables is a reasonable assumption; see Table 1 for a summary. All the data sets can be downloaded at our website[3]. Many of them have also been used in previous studies on monotone learning. Some of them have a numerical or ordered categorical output, however. These outputs were binarized by thresholding at the median. Moreover, all input attributes were normalized.

5.2 Methods

Since choquistic regression (CR) can be seen as an extension of standard logistic regression (LR), it is natural to compare these two methods. Essentially,

[2] `http://archive.ics.uci.edu/ml/`

[3] `http://www.uni-marburg.de/fb12/kebi/research/`

Table 1. Data sets and their properties

data set	#instances	#attributes	source
Den Bosch (DBS)	120	8	[30]
CPU	209	6	UCI
Breast Cancer (BCC)	286	9	UCI
Auto MPG	392	7	UCI
Employee Selection (ESL)	488	4	WEKA
Mammographic (MMG)	961	6	UCI
Employee Rejection/Acceptance (ERA)	1000	4	WEKA
Lecturers Evaluation (LEV)	1000	4	WEKA
Car Evaluation (CEV)	1728	6	UCI

this comparison should give an idea of the usefulness of an increased flexibility. On the other side, one may also ask for the usefulness of assuring monotonicity. Therefore, we additionally included two other extensions of LR, which are flexible but not necessarily monotone, namely kernel logistic regression (KLR) with polynomial and Gaussian kernels. The degree of the polynomial kernel was set to 2, so that it models low-level interactions of the features. The Gaussian kernel, on the other hand, is able to capture interactions of higher order. For each data set, the width parameter of the Gaussian kernel was selected from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$ in the most favorable way. Finally, we included a method which is both monotone and flexible, namely the MORE algorithm for learning rule ensembles under monotonicity constraints [6].

5.3 Results

Classification accuracy was measured in terms of 0/1 loss and determined by randomly splitting the data into two parts, one part for training and one part for testing. This was repeated 100 times, and the accuracy degrees were averaged.

A possible improvement of CR over its competitors, in terms of predictive accuracy, may be due to two reasons: First, in comparison to standard LR, it is more flexible and has the ability to capture nonlinear dependencies between input attributes. Second, in comparison to non-monotone learners, it takes background knowledge about the dependency between input and output variables into consideration.

Both aspects have to be put in perspective, however. First, regarding flexibility, it is clear that an improvement is unlikely unless additional flexibility is indeed needed. On the contrary, if the true underlying dependency is indeed a linear one, at least approximately, then standard logistic regression will be the model of choice, whereas CR may tend to overfit the training data and hence generalize worse. Regarding monotonicity, previous studies have indeed shown that improvements are possible, albeit of a small margin. In fact, upon closer examination, the benefit of enforcing monotonicity is not entirely obvious [31]. Moreover, the more extensive the training data, the smaller the improvement

Table 2. Classification performance in terms of the mean and standard deviation of 0/1 loss. From top to bottom: 20%, 50%, and 80% training data.

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)
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)
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)
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

tends to be. This is understandable, since background knowledge will lose importance with an increasing number of observations.

The results of the experiments are summarized in Table 2 and 3. As can be seen, CR compares quite favorably with the other approaches, especially with the non-monotone KLR methods. It also outperforms LR, at least for sufficiently extensive training data; if the amount of training data is small, however, LR is even better, probably because CR will then tend to overfit the data. Finally, CR also compares favorably with MORE, although the difference in terms of the average ranks is not statistically significant (the critical distance for the Nemenyi test at significance level 0.05 is 2.03).

In Fig. 3, a visualization of the (pairwise) interaction between attributes is shown for the car evaluation data, for which CR performs significantly better than LR. In this data set, the evaluation of a car (output attribute) depends on a number of criteria, namely (a) buying price, (b) price of the maintenance, (c) number of doors, (d) capacity in terms of persons to carry, (e) size of luggage boot, (f) safety of the car. These criteria form a natural hierarchy: (a) and (b) form a subgroup PRICE, whereas the other properties are of a TECHNICAL nature and can be further decomposed into COMFORT (c–e) and safety (f). Interestingly, the interaction in our model nicely agrees with this hierarchy: Interaction within each subgroup tends to be smaller (as can be seen from the

Table 3. Win statistics (number of data sets on which the first method was better than the second one) for 20%, 50%, and 80% training data

	CR	LR	KLR-ply	KLR-rbf	MORE
CR	–	2 | 6 | 9	7 | 7 | 9	7 | 9 | 9	7 | 5 | 6
LR	7 | 3 | 0	–	7 | 5 | 5	7 | 7 | 6	7 | 3 | 2
KLR-ply	2 | 2 | 0	2 | 4 | 4	–	5 | 5 | 6	7 | 4 | 5
KLR-rbf	2 | 0 | 0	2 | 2 | 3	4 | 4 | 3	–	6 | 2 | 2
MORE	2 | 4 | 3	2 | 6 | 7	2 | 5 | 4	3 | 7 | 7	–

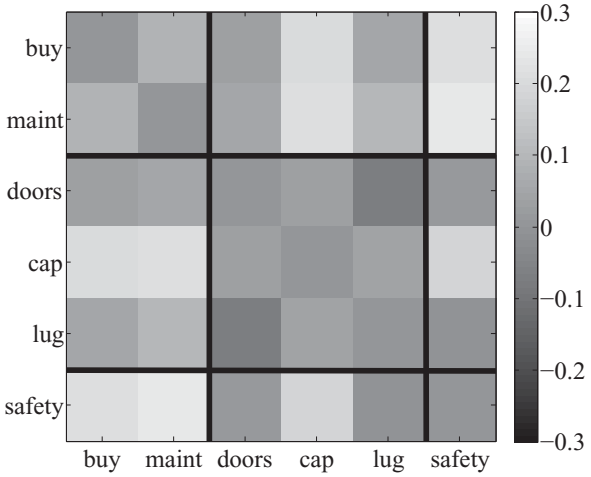

Fig. 3. Visualization of the interaction index for the car evaluation data (numerical values are shown in terms of level of gray, values on the diagonal are set to 0). Groups of related criteria are indicated by the black lines.

darker colors) than interaction between criteria from different subgroups, suggesting a kind of redundancy in the former and complementarity in the latter case.

6 Concluding Remarks

In this paper, we have advocated the use of the discrete Choquet integral as an aggregation operator in machine learning, especially in the context of learning monotone models. Apart from combining monotonicity and flexibility in a mathematically sound and elegant manner, the Choquet integral offers measures for quantifying the importance of individual predictor variables and the interaction between groups of variables, thereby providing important information about the relationship between independent and dependent variables.

As a concrete application, we have proposed a generalization of logistic regression, in which the Choquet integral is used for modeling the log odds of the positive class. First experimental studies have shown that this method, called choquistic regression, compares quite favorably with other methods. We like to

mention again, however, that an improvement in prediction accuracy should not necessarily be seen as the main goal of monotone learning. Instead, the adherence to monotonicity constraints is often an important prerequisite for the acceptance of a model by domain experts.

An interesting question to be addressed in future work concerns a possible restriction of the choquistic model to k-additive measures, for a suitable value of k. This may have two important advantages: First, it may prevent from overfitting the data in cases where the full flexibility of the Choquet integral is actually not needed. Second, since less parameters need to be identified, the computational complexity will be reduced, too. Of course, the key problem to be addressed in this regard concerns the question of how to choose k in the most favorable way.

Beyond that, the Choquet integral can of course be combined with other machine learning methods, and its use is not restricted to (binary) classification. We are quite convinced of its high potential in machine learning in general, and we are looking forward to exploring this potential in greater detail.

References

1. Ben-David, A., Sterling, L., Pao, Y.-H.: Learning and classification of monotonic ordinal concepts. Computational Intelligence 5(1), 45–49 (1989)
2. Duivesteijn, W., Feelders, A.: Nearest neighbour classification with monotonicity constraints. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part I. LNCS (LNAI), vol. 5211, pp. 301–316. Springer, Heidelberg (2008)
3. Sill, J.: Monotonic networks. In: Advances in Neural Information Processing Systems, pp. 661–667. The MIT Press, Denver (1998)
4. Ben-David, A.: Monotonicity maintenance in information-theoretic machine learning algorithms. Machine Learning 19, 29–43 (1995)
5. Potharst, R., Feelders, A.: Classification trees for problems with monotonicity constraints. ACM SIGKDD Explorations Newsletter 4(1), 1–10 (2002)
6. Dembczyński, K., Kotlowski, W., Slowinski, R.: Learning rule ensembles for ordinal classification with monotonicity constraints. Fundamenta Informaticae 94(2), 163–178 (2009)
7. Chandrasekaran, R., Ryu, Y., Jacob, V., Hong, S.: Isotonic separation. INFORMS Journal on Computing 17, 462–474 (2005)
8. Dembczyński, K., Kotłowski, W., Słowiński, R.: Additive preference model with piecewise linear components resulting from dominance-based rough set approximations. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Żurada, J.M. (eds.) ICAISC 2006. LNCS (LNAI), vol. 4029, pp. 499–508. Springer, Heidelberg (2006)
9. Feelders, A.: Monotone relabeling in ordinal classification. In: Proceedings of the 10th IEEE International Conference on Data Mining, pp. 803–808. IEEE Computer Society, Los Alamitos (2010)
10. Kotłowski, W., Dembczyński, K., Greco, S., Słowiński, R.: Stochastic dominance-based rough set model for ordinal classification. Information Sciences 178(21), 3989–4204 (2008)
11. Grabisch, M., Murofushi, T., Sugeno, M. (eds.): Fuzzy Measures and Integrals: Theory and Applications. Physica, Heidelberg (2000)

12. Grabisch, M.: Fuzzy integral in multicriteria decision making. Fuzzy Sets and Systems 69(3), 279–298 (1995)
13. Torra, V.: Learning aggregation operators for preference modeling. In: Preference Learning, pp. 317–333. Springer, Heidelberg (2011)
14. Torra, V., Narukawa, Y.: Modeling Decisions: Information Fusion and Aggregation Operators. Springer, Heidelberg (2007)
15. Grabisch, M.: Modelling data by the Choquet integral. In: Information Fusion in Data Mining, pp. 135–148. Springer, Heidelberg (2003)
16. Mori, T., Murofushi, T.: An analysis of evaluation model using fuzzy measure and the Choquet integral. In: Proceedings of the 5th Fuzzy System Symposium, pp. 207–212. Japan Society for Fuzzy Sets and Systems (1989)
17. Grabisch, M.: A new algorithm for identifying fuzzy measures and its application to pattern recognition. In: Proceedings of IEEE International Conference on Fuzzy Systems, vol. 1, pp. 145–150. IEEE, Los Alamitos (1995)
18. Angilella, S., Greco, S., Matarazzo, B.: Non-additive robust ordinal regression with Choquet integral, bipolar and level dependent Choquet integrals. In: Proceedings of the Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference, IFSA/EUSFLAT, pp. 1194–1199 (2009)
19. Beliakov, G., James, S.: Citation-based journal ranks: the use of fuzzy measures. Fuzzy Sets and Systems 167(1), 101–119 (2011)
20. Grabisch, M., Nicolas, J.-M.: Classification by fuzzy integral: performance and tests. Fuzzy Sets and Systems 65(2-3), 255–271 (1994)
21. Sugeno, M.: Theory of Fuzzy Integrals and its Application. PhD thesis, Tokyo Institute of Technology (1974)
22. Murofushi, T., Soneda, S.: Techniques for reading fuzzy measures (III): interaction index. In: Proceedings of the 9th Fuzzy Systems Symposium, pp. 693–696 (1993)
23. Grabisch, M.: k-order additive discrete fuzzy measures and their representation. Fuzzy Sets and Systems 92(2), 167–189 (1997)
24. Vitali, G.: Sulla definizione di integrale delle funzioni di una variabile. Annali di Matematica Pura ed Applicata 2(1), 111–121 (1925)
25. Choquet, G.: Theory of capacities. Annales de l'institut Fourier 5, 131–295 (1954)
26. Hosmer, D., Lemeshow, S.: Applied Logistic Regression, 2nd edn. Wiley, Chichester (2000)
27. Jaccard, J.: Interaction Effects in Logistic Regression. Saga Publications, Thousand Oaks (2001)
28. Modave, F., Grabisch, M.: Preference representation by a Choquet integral: commensurability hypothesis. In: Proceedings of the 7th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp. 164–171. Editions EDK (1998)
29. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter 11(1), 10–18 (2009)
30. Daniels, H., Kamp, B.: Applications of mlp networks to bond rating and house pricing. Neural Computation and Applications 8, 226–234 (1999)
31. Ben-David, A., Sterling, L., Tran, T.: Adding monotonicity to learning algorithms impair their accuracy. Expert Systems with Applications 36(3), 6627–6634 (2009)