Ordinal Choquistic Regression

Ali Fallah Tehrani and Eyke Hüllermeier

Department of Mathematics and Computer Science
University of Marburg, Germany
{fallah, eyke}@mathematik.uni-marburg.de

Abstract

We propose an extension of choquistic regression from the case of binary to the case of ordinal classification. Choquistic regression itself has been introduced recently as a generalization of conventional logistic regression. The basic idea of this method is to replace the linear function of predictor variables in the logistic regression model by the Choquet integral. Thus, it becomes possible to capture nonlinear dependencies and interactions among predictor variables while preserving two important properties of logistic regression, namely the comprehensibility of the model and the possibility to ensure its monotonicity in individual predictors. In experimental studies, choquistic regression consistently improves upon standard logistic regression in terms of predictive accuracy, especially when being combined with a novel regularization technique that prevents from exceeding the required level of nonadditivity.

Keywords: logistic regression, ordinal classification, Choquet integral, monotone classification, attribute interaction

1. Introduction

Logistic regression is a well-established statistical method for (probabilistic) classification [1]. Its popularity is due to a number of appealing properties, including the following ones:

- Since the model, at least in its basic form, is essentially *linear* in the input attributes, it is easily comprehensible. In particular, the strength of influence of each predictor is directly reflected by the corresponding regression coefficient.
- The model behaves *monotone* in each predictor variable; in the binary case, for example, this means that an increase of the value of the variable can only increase (decrease) the probability of the positive class.

Both of the above points, comprehensibility and monotonicity, are important prerequisites for the acceptance of a model by a domain expert. Indeed, in many cases, monotonicity is a very natural requirement. In a medical context, for example, tobacco consumption is expected to increase the probability of cancer, and each model violating this property will not be considered as trustworthy.

Nevertheless, the linearity of a logistic regression model is of course a strong restriction from a learning point of view. Quite often, the response variable (output) depends on the predictor variables (inputs) in a *nonlinear* way. In previous work [2,3], we therefore proposed an extension of logistic regression that allows for modeling nonlinear relationships between input and output variables while preserving the aforementioned advantages of the approach, namely comprehensibility and monotonicity. Roughly speaking, the basic idea of our approach, called "choquistic regression", is to replace the linear function in the logistic regression model by the Choquet integral.

Choquistic regression as proposed in [2,3] is restricted to the dichotomous case, i.e., to the case of classification with two classes. From a decision making point of view, this means that, based on a set of criteria, alternatives are simply classified as "positive" or "negative" ("good" or "bad"). In this paper, we generalize choquistic regression to the polychotomous case or, more specifically, the case of ordinal classification, where instances (alternatives) can be assigned to one among several ordered categories. For example, based on criteria such as the number of citations, scientific journals might be categorized as A^*, A, B or C [4].

The rest of this paper is organized as follows. In the next section, we give a brief introduction to the problem of (ordinal) classification. In Section 3, we recall the basics of (ordinal) logistic regression. The discrete Choquet integral is briefly recalled in Section 4. Ordinal choquistic regression is then introduced in Section 5. Experimental results are presented in Section 6, prior to concluding the paper in Section 7.

2. Problem setting

We consider the problem of classification, that is, predicting the value of an output (response) variable $y \in \mathcal{Y}$ given the values of a set of input attributes (predictors) $x_i \in \mathcal{X}_i$, $i = 1, \ldots, m$. The vector

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

is called an *instance*, and $\mathcal{X}$ the *instance space*. In binary classification, $\mathcal{Y} = \{0, 1\}$ consists of only two classes, typically called the negative (0) and the positive (1) class. In ordinal classification,

$$\mathcal{Y} = \{y_1, y_2, \ldots, y_K\} \tag{1}$$

consists of K classes that are assumed to be ordered: $y_1 < y_2 < \ldots < y_K$.

The goal is to learn a classifier $\mathcal{L} : \mathcal{X} \to \mathcal{Y}$ from a given set of training data

$$\mathcal{D} = \left\{ (\boldsymbol{x}^{(i)}, y^{(i)}) \right\}_{i=1}^{n} \subset (\mathcal{X} \times \mathcal{Y})^n. \tag{2}$$

The data $\mathcal{D}$ is supposed to be an *i.i.d.* (independent and identically distributed) sample generated by an underlying (though unknown) probability measure $\mathbf{P}_{XY}$ on $\mathcal{X} \times \mathcal{Y}$. A common goal, then, is to induce a classifier with minimal risk, where the risk $R(\mathcal{L})$ of a classifier $\mathcal{L}$ is defined in terms of its expected loss:

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

where $\ell(\cdot)$ is a loss function penalizing incorrect predictions. In binary classification, the most commonly used loss is the simple 0/1 loss given by

$$\ell(\hat{y}, y) = \begin{cases} 0 \text{ if } \hat{y} = y \\ 1 \text{ if } \hat{y} \neq y \end{cases} \ . \tag{3}$$

Although this loss can of course also be used in ordinal classification, it does not take the order of the classes (1) into account; therefore, one often uses the L_1 loss instead:[1]

$$\ell(\hat{y}, y) = |\operatorname{ind}(\hat{y}) - \operatorname{ind}(y)| \ , \tag{4}$$

with $\operatorname{ind}(y_j) = j$ for all $y_j \in \mathcal{Y}$.

3. Background on logistic regression

3.1. The binary case

In the binary case, logistic regression models the probability of the positive class (and hence of the negative class) as a linear (affine) function of the input attributes. More specifically, since a linear function does not necessarily produce values in the unit interval, the response is defined as a generalized linear model, namely in terms of the logarithm of the probability ratio:

$$\log \left(\frac{\mathbf{P}(y = 1 \,|\, \boldsymbol{x})}{\mathbf{P}(y = 0 \,|\, \boldsymbol{x})} \right) = \beta_0 + \boldsymbol{w}^\top \boldsymbol{x} \ , \tag{5}$$

where $\boldsymbol{w} = (w_1, w_2, \ldots, w_m)^\top \in \mathbb{R}^m$ is a vector of regression coefficients and $\beta_0 \in \mathbb{R}$ 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

$$\mathbf{P}(y = 1 \,|\, \boldsymbol{x}) = \frac{1}{1 + \exp(-\beta_0 - \boldsymbol{w}^\top \boldsymbol{x})} \ . \tag{6}$$

[1] Note that this loss implicitly assumes an equal "distance" between the ordinal categories, which is clearly arguable.

3.2. Ordinal logistic regression

Now, consider the case of ordinal classification, where we are given K ordered classes $\{y_1, \ldots, y_K\}$. The idea of ordinal logistic regression is to reduce the corresponding classification problem to the binary case while taking into account (and actually exploiting) the class order. To this end, it models a probability ratio similar to (5), but this time for the *cumulative distribution*:

$$\log \left(\frac{\pi_k(\boldsymbol{x})}{1 - \pi_k(\boldsymbol{x})} \right) = \beta_k + \boldsymbol{w}^\top \boldsymbol{x} \tag{7}$$

for $k \in [K-1] = \{1, \ldots, K-1\}$, where

$$\pi_k(\boldsymbol{x}) = \mathbf{P}(y \leq y_k \,|\, \boldsymbol{x})$$

is the (conditional) probability that the class y observed for $\boldsymbol{x}$ is at most y_k; correspondingly,

$$1 - \pi_k(\boldsymbol{x}) = \mathbf{P}(y > y_k \,|\, \boldsymbol{x})$$

is the probability that the class y is larger than y_k. Obviously, the left-hand side in (7) is non-decreasing in k. Therefore, since the right-hand side only differs in the intercepts (thresholds) β_k, we need to impose the condition

$$\beta_1 \leq \beta_2 \leq \cdots \leq \beta_{K-1} \ .$$

From (7), one derives

$$\begin{aligned} \pi_k(\boldsymbol{x}) &= \mathbf{P}(y \leq y_k \,|\, \boldsymbol{x}) \\ &= \frac{\exp(\beta_k) \exp(\boldsymbol{w}^\top \boldsymbol{x})}{1 + \exp(\beta_k) \exp(\boldsymbol{w}^\top \boldsymbol{x})} \end{aligned}$$

Moreover, exploiting the definition of the cumulative distribution, the class probabilities can be derived as

$$\begin{aligned} \mathbf{P}(y = y_k \,|\, \boldsymbol{x}) &= \mathbf{P}(y \leq y_k \,|\, \boldsymbol{x}) - \mathbf{P}(y \leq y_{k-1} \,|\, \boldsymbol{x}) \\ &= \pi_k(\boldsymbol{x}) - \pi_{k-1}(\boldsymbol{x}) \end{aligned}$$

for $k \in [K]$ (where $\pi_K(\boldsymbol{x}) = 1$ and $\pi_0(\boldsymbol{x}) = 0$ by definition).

Given a set of training data (2), the estimation of the parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{K-1})$ and $\boldsymbol{w}$ is then accomplished through maximum likelihood estimation, i.e., by maximizing the log-likelihood

$$\begin{aligned} l(\beta, \boldsymbol{w}) &= \sum_{i=1}^{n} \log \mathbf{P}\left(y^{(i)} \,|\, \boldsymbol{x}^{(i)} \right) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{I}(y^{(i)} = y_k) \, \log \left(\pi_k(\boldsymbol{x}) - \pi_{k-1}(\boldsymbol{x}) \right) \end{aligned}$$

4. The discrete Choquet integral

In this section, we recall the basic definition of the Choquet integral and related notions. The first definition of the Choquet integral for additive measures is due to Vitali [5]. For the general case of a capacity (i.e., a non-additive measure or fuzzy measure),

it was later on introduced by Choquet [6]. Yager proposed a generalized version in [7].

Let $C = \{c_1, \ldots, c_m\}$ be a finite set and $\mu : 2^C \to [0, 1]$ a measure. For each $A \subseteq C$, the value $\mu(A)$ can be interpreted as the weight or, say, the importance of the set of elements A. 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 the "context" A.

This lack of expressivity motivates the use of non-additive measures, also called capacities or fuzzy measures, which are simply normalized and monotone but not necessarily additive [8]:

$$\begin{aligned} &\mu(\emptyset) = 0,\ \mu(C) = 1 \\ &\mu(A) \leq \mu(B) \text{ for all } A \subseteq B \subseteq C \end{aligned} \tag{8}$$

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

$$\mu(B) = \sum_{A \subseteq B} \boldsymbol{m}_\mu(A) \tag{9}$$

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

$$\boldsymbol{m}_\mu(A) = \sum_{B \subseteq A} (-1)^{|A|-|B|} \mu(B) \ . \tag{10}$$

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. In particular, as can be seen from (9), it means that a measure μ can formally be specified by significantly fewer than 2^m values, which are needed in the general case.

Suppose the "criteria" $c_i \in C$ are simply considered as binary features, which are either present or absent in a set A. Mathematically, $\mu(A)$ can then 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*

$$\mathcal{C}_\mu(f) = \sum_{i=1}^{m} w_i \cdot f(c_i) = \sum_{i=1}^{m} \mu(\{c_i\}) \cdot f(c_i) \ , \tag{11}$$

which is a natural aggregation operator in this case. A non-trivial question, however, is how to generalize (11) 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 [6].

In the discrete case, the Choquet integral is formally defined as follows:

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

where $(\cdot)$ is a permutation of $[m]$ 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 $A_{(i)} = \{c_{(i)}, \ldots, c_{(m)}\}$. In terms of the Möbius transform of μ, the Choquet integral can also be expressed as follows:

$$\mathcal{C}_\mu(f) = \sum_{T \subseteq C} \boldsymbol{m}(T) \cdot \min_{i \in T} f(c_i) \tag{12}$$

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

5. Ordinal choquistic regression

5.1. The choquistic model

In order to model non-linear 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 (affine) linear function $\boldsymbol{x} \mapsto \beta_k + \boldsymbol{w}^\top \boldsymbol{x}$ in (7) by the Choquet integral. More specifically, we propose the following model

$$\log\left(\frac{\pi_k(\boldsymbol{x})}{1 - \pi_k(\boldsymbol{x})}\right) = \gamma\Big(\mathcal{C}_\mu(f_{\boldsymbol{x}}) - \beta_k\Big) \tag{13}$$

where $\mathcal{C}_\mu(f_{\boldsymbol{x}})$ is the Choquet integral (with respect to the measure μ) of the evaluation function

$$f_{\boldsymbol{x}} : \{c_1, \ldots, c_m\} \to \mathbb{R}_+$$

that maps each attribute c_i to a value $x_i = f_{\boldsymbol{x}}(c_i)$; $\gamma \geq 0$ and $\beta_1, \ldots, \beta_{K-1}$ are real constants such that $0 = \beta_1 \leq \beta_2 \leq \cdots \leq \beta_{K-1} \leq 1$.

The value of c_i is normalized in order to turn each predictor variable into a criterion, i.e., a "the higher the better" attribute, and to assure commensurability between the criteria [9]. A simple transformation, that we shall also employ in our experimental studies, is given by the mapping

$$z_i = F_i^{-1}(x_i) \ , \tag{14}$$

where F_i is the cumulative distribution function $x \mapsto \mathbf{P}(X_i \leq x)$. Of course, since this function is in general not known, it has to be replaced by an estimate $\hat{F}_i$; to this end, we simply adopt the empirical distribution of the training data:

$$\hat{F}_i(x) = \#\big\{ (x_1, \ldots, x_m) \in \mathcal{D} \,|\, x_i \leq x \big\}$$

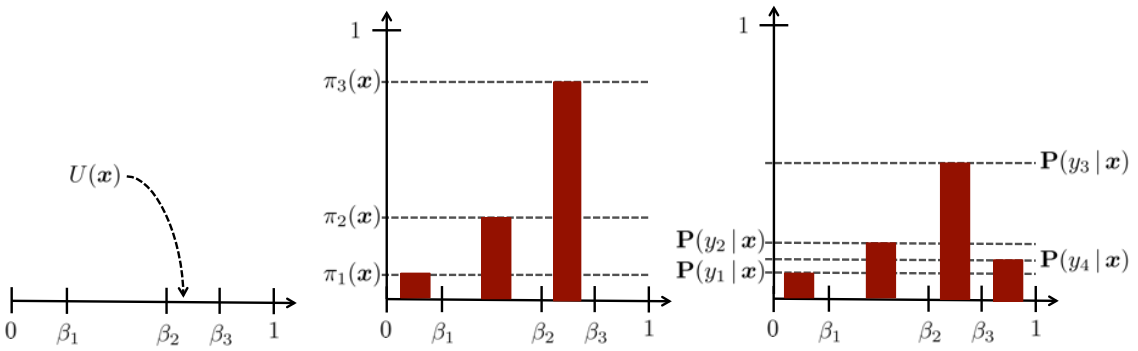

Figure 1: Illustration of the ordinal choquistic regression model for $\mathcal{Y} = \{y_1, y_2, y_3, y_4\}$: Class assignment via hard thresholding (left) versus probabilistic classification. The cumulative distribution $y_k \mapsto \pi_k(\boldsymbol{x})$ is shown in the middle, the probability distribution $y_k \mapsto \mathbf{P}(y = y_k \,|\, \boldsymbol{x}) = \pi_k(\boldsymbol{x}) - \pi_{k-1}(\boldsymbol{x})$ on the right.

The model (13) can be seen as a discrete choice process consisting of two steps: The first step consists of an assessment of the input $\boldsymbol{x}$ in terms of a utility degree

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

Then, in a second step, this utility degree is compared with the thresholds β_k, which become increasingly demanding. The question whether or not the alternative should be classified as "at least as good as y_k" depends on whether or not $U(\boldsymbol{x})$ exceeds the threshold β_k.

The second step is a probabilistic version of a (hard) thresholding procedure (see Figure 1): Instead of simply assigning $\boldsymbol{x}$ to the class $y_{k\bullet}$ such that $\beta_{k\bullet-1} \leq U(\boldsymbol{x}) < \beta_{k\bullet}$, each class has a certain probability. The level of determination of the decision is specified by γ, which serves as a scaling parameter: The larger γ, the more peaked the probability distribution becomes; in the limit $\gamma \to \infty$, the class $y_{k\bullet}$ will be chosen deterministically.

5.2. Parameter estimation

The model (13) has several degrees of freedom: The fuzzy measure μ (Möbius transform $\boldsymbol{m} = \boldsymbol{m}_\mu$) determines the (latent) utility function, while the utility thresholds $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{K-1})$ 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 (2). 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

$$\begin{aligned} l(\boldsymbol{m}, \gamma, \boldsymbol{\beta}) &= \log \mathbf{P}\Big(\mathcal{D} \,|\, \boldsymbol{m}, \boldsymbol{\beta}, \gamma\Big) \\ &= \log \prod_{i=1}^{n} \mathbf{P}\left(y^{(i)} \,|\, \boldsymbol{x}^{(i)}; \boldsymbol{m}, \beta, \gamma\right) \qquad (15) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{I}\left(y^{(i)} = y_k\right) \log \pi_k^*(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{\beta}, \gamma), \end{aligned}$$

where

$$\begin{aligned} \pi_k^*(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{\beta}, \gamma) = &\frac{\exp(-\gamma\beta_k)\exp(\gamma\mathcal{C}_\mu(f_{\boldsymbol{x}}))}{1 + \exp(-\gamma\beta_k)\exp(\gamma\mathcal{C}_\mu(f_{\boldsymbol{x}}))} \\ &- \frac{\exp(-\gamma\beta_{k-1})\exp(\gamma\mathcal{C}_\mu(f_{\boldsymbol{x}}))}{1 + \exp(-\gamma\beta_{k-1})\exp(\gamma\mathcal{C}_\mu(f_{\boldsymbol{x}}))} \end{aligned}$$

One can verify that (15) is convex with respect to $\boldsymbol{m}, \gamma$, and β. In principle, maximization of the log-likelihood can hence 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, namely the maximization of (15) under the following conditions (recall that $C = \{c_1, \ldots, c_m\}$ denotes the set of predictor variables):

$$\begin{aligned} &0 \leq \beta_1 \leq \beta_2 \leq \cdots \leq \beta_{K-1} \leq 1 \\ &0 < \gamma \\ &\sum_{T \subseteq C} \boldsymbol{m}(T) = 1 \\ &\sum_{B \subseteq A \setminus \{c_i\}} \boldsymbol{m}(B \cup \{c_i\}) \geq 0 \quad \forall A \subseteq C, c_i \in C \end{aligned}$$

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.

5.3. Regularization

Since the (full) Choquet integral is a complex and highly nonlinear function with many degrees of freedom, ML estimation obviously comes with the danger of over-fitting the data. Therefore, a kind of regularization of the estimation process is clearly advisable from a learning point of view. Apart from improving generalization performance, regularization may also serve the purpose of obtaining simpler models that are smaller in size.

Regularization is typically done by adding a complexity term to the objective function (in our case the log-likelihood (15)), the influence of which is controlled by a regularization parameter ρ. Thus, the idea is to penalize overly complex models or, stated differently, to find a compromise between the complexity of a model and its fit to the data. The key question, of course, it how to quantify the notion of "complexity" of a model.

In our approach to ordinal choquistic regression, adding a regularization term of the form

$$-\rho \sum_{A \subseteq C} g(|A|)\, |\boldsymbol{m}(A)| \tag{16}$$

to the objective function (15) turned out to produce good results. Defining $g(\cdot)$ as a strictly increasing function, this term encourages $\boldsymbol{m}(A) = 0$ for larger subsets of criteria A; in other words, it encourages a restriction to measures with a low level of nonadditivity. We note that (16) can be seen as a specific instance of the idea of "hierarchical regularization" as introduced in [10], with a hierarchy on the power set 2^C defined through subset cardinality (i.e., the first level of the hierarchy are the singletons $\{c_i\}$, the second level the two-subsets $\{c_i, c_j\}$, etc.).

data set	#instances	m	K
ESL	488	4	9
ERA	1000	4	9
MPG	398	8	7
LEV	1000	4	5
CEV	1728	6	4
CYD 1 (Blue 26)	120	3	5
CYD 2 (Brown 1)	120	3	5
CYD 3 (Blue 56)	120	3	5
CYD 4 (Red 60)	120	3	5
CYD 5 (Yellow 7)	120	3	5
CYD 6 (Yellow 23)	120	3	5
CYD 7 (Mixture)	120	3	5

Table 1: Data sets and their properties (number of instances, number of attributes (m), number of classes (K).

5.4. Prediction

Once a choquistic model (13) has been learned on a given set of training data, it can be used to predict the class of a new query instance $\boldsymbol{x} \in \mathcal{X}$. This prediction, however, is not straightforward, since (13) does not produce a class prediction directly. Instead, it maps $\boldsymbol{x}$ to a probability distribution

$$\Big(\mathbf{P}(y_1 \,|\, \boldsymbol{x}), \ldots, \mathbf{P}(y_K \,|\, \boldsymbol{x})\Big) \in [0,1]^{\mathcal{Y}},$$

from which a class prediction has to be derived. The most obvious prediction, of course, is the mode of this distribution:

$$\hat{y} = \arg\max \Big\{ \mathbf{P}(y_k \,|\, \boldsymbol{x}) \,|\, y_k \in \mathcal{Y} \Big\} \tag{17}$$

Indeed, this prediction minimizes the risk with respect to the 0/1 loss (3). The risk minimizer with respect to the L_1 loss (4), however, is the median of the distribution:

$$\hat{y} = \arg\operatorname{med} \Big(\mathbf{P}(y_1 \,|\, \boldsymbol{x}), \ldots, \mathbf{P}(y_K \,|\, \boldsymbol{x})\Big) \tag{18}$$

6. Experimental evaluation

Experimentally, we compared our generalized variant (13) to the standard version (5) of logistic regression on several benchmark data sets. What we expect, of course, is an improved predictive accuracy thanks to the increased flexibility of choquistic regression, namely its ability to capture nonlinear dependencies between predictor variables and response. It should be noted, however, that such an improvement, despite being plausible, is not self-evident. In fact, if the true underlying dependency is indeed linear or close to linear, then standard logistic regression will be the model of choice, whereas choquistic regression may tend to overfit the training data and hence generalize worse.

The following methods were included in the experiments:

- Ordinal logistic regression (OLR). Actually, OLR was implemented as ordinal choquistic regression restricted to the case of an additive measure. These two variants are essentially equivalent, except for the fact that the latter incorporates monotonicity constraints. Given that our data sets (see below) are monotone, this constraint is clearly reasonable.
- Two variants of ordinal choquistic regression as introduced in Section 5, namely with (OCR+R) and without (OCR) hierarchical regularization (16).

6.1. Data

Even though the topic of monotone classification is receiving more and more interest in the machine learning community [11–15], benchmark data for this problem is not as abundant as for conventional classification; for our purpose, the class attribute also needs to be polychotomous and ordered, which restricts the choice of data sets even further. Nevertheless, we managed to collect a number of suitable benchmark data sets, mostly from the UCI repository[2] and the WEKA machine learning toolbox.[3] In what follows, we give a brief description of each of these data sets; an overview of the data sets together with their main properties is given in Table 1.

- Employee Selection (ESL): This data set contains profiles of applicants for certain industrial jobs. The values of the four input attributes were determined by psychologists based upon

[2] http://archive.ics.uci.edu/ml/
[3] http://www.cs.waikato.ac.nz/ml/weka/

data set	OLR	OCR	OCR+R
ESL	.3094±.0325(1)	.3504±0.0939(3)	.3361±.0427(2)
ERA	1.2520±.0393(1)	1.2770±0.0279(3)	1.2617±.0292(2)
MPG	.3667±.0314(1)	.3761±0.0318(2)	.3830±.0419(3)
LEV	.4264±.0148(3)	.4184±0.0187(1)	.4224±.0242(2)
CEV	.2310±.0075(3)	.1097±0.0361(1)	.1097±.0361(1)
CYD–1	.3167±.0441(3)	.1778±0.0536(2)	.1611±.0509(1)
CYD–2	.7722±.0712(3)	.3500±0.0810(2)	.3472±.0885(1)
CYD–3	.4667±.0471(3)	.2722±0.0360(2)	.2694±.0386(1)
CYD–4	.5133±.0414(3)	.2833±0.0583(2)	.2783±.0634(1)
CYD–5	.3100±.0465(3)	.2633±0.0477(2)	.2500±.0373(1)
CYD–6	.5083±.0874(3)	.2556±0.0750(2)	.2500±.0667(1)
CYD–7	.7150±.0541(3)	.3867±0.0628(2)	.3850±.0739(1)
ESL	.3400±.0504(1)	.3488±.0464(3)	.3456± .0184(2)
ERA	1.2824±.0648(2)	1.292±.0552(3)	1.2712±.0384(1)
MPG	.3365±.0375(3)	.3105±.0335(2)	.3045±.0310(1)
LEV	.4372±.0344(3)	.4164±.0140(1)	.4204±.0148(2)
CEV	.2205±.0096(3)	.1203±.0291(2)	.1137±.0246(1)
CYD–1	.3479±.0490(3)	.1952±.0498(2)	.1896±.0493(1)
CYD–2	.8167±.1017(3)	.3483±.0644(2)	.3425±.0698(1)
CYD–3	.4167±.0786(3)	.2700±.0375(1)	.2733±.0425(2)
CYD–4	.4633±.0576(3)	.3000±.0437(2)	.2933±.0432(1)
CYD–5	.3067±.0562(3)	.2833±.0360(2)	.2724±.0410(1)
CYD–6	.5583±.0748(3)	.2867±.0461(2)	.2783±.0409(1)
CYD–7	.7711±.0727(3)	.3289±.0682(1)	.3380±.0610(2)

Table 2: Average L_1 loss ± standard deviation (in brackets the rank). The results above refer to the median predictor (18), the results below to the mode predictor (17).

psychometric test results and interviews with the candidates. The output is an overall score on an ordinal scale between 1 and 9, corresponding to the degree of suitability of each candidate to this type of job.

- Employee Rejection/Acceptance (ERA): This data set originates from an academic decision-making experiment. The input attributes are features of a candidate such as past experience, verbal skills, etc., and the output is the subjective judgment of a decision-maker, measured on an ordinal scale from 1 to 9, to which degree he or she tends to accept the applicant for the job.
- Auto MPG (MPG): This data set was used in the 1983 American Statistical Association Exposition. The data is about the city-cycle fuel consumption of cars in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes (cylinders, displacement, horsepower, weight, acceleration, model year, origin). We removed incomplete instances and discretized the numerical output (mpg) into 7 ordinal classes using equi-with binning.
- Lecturers Evaluation (LEV): This data set contains examples of anonymous lecturer evaluations, taken at the end of MBA courses. Students were asked to score their lecturers according to four attributes such as oral skills and contribution to their professional/general knowledge. The output was a total evaluation of each lecturer's performance, measured on an ordinal scale from 0 to 4.
- Car Evaluation (CEV): This data set contains 6 attributes describing a car, namely, buying price, price of the maintenance, number of doors, capacity in terms of persons to carry, the size of luggage boot, estimated safety of the car. The output is the overall evaluation of the car: unacceptable, acceptable, good, very good.
- Color Yield (CYD): Finally, we took data from an industrial polyester dyeing process that was also analyzed in [16]. Here, the output variable is the color yield, which has been measured as a function of three important factors: disperse dyes concentration, temperature and time of dyeing. Corresponding experiments have been made for 7 different colors, giving rise to 7 data sets. Each of these data sets was discretized by equi-frequency binning of the color yield.

6.2. Experimental setup and results

We use an experimental setup that randomly splits the data into two parts, one half for training and one half for testing. The model induced from training data is then evaluated on the test data. This procedure is repeated 100 times, and the results are averaged.

The function $g(\cdot)$ in the regularization term (16)

data set	temperature	time	concentration
CYD–1	0.6147	0.1342	0.2511
CYD–2	0.4612	0.0878	0.4510
CYD–3	0.3935	0.1191	0.4874
CYD–4	0.4473	0.1051	0.4476
CYD–5	0.2477	0.0330	0.7193
CYD–6	0.4238	0.1146	0.4616
CYD–7	0.4939	0.0311	0.4749

Table 3: Shapley values of criteria for the color data.

was defined as $g(k) = k^\alpha$. Thus, two hyperparameters need to be tuned for OCR+R, namely ρ and α. This tuning was done by searching the grid

$$(\alpha, \rho) \in \{2, 4, 6, 8\} \times \{10^{-4}, 10^{-3}, \ldots, 10^4\}$$

and evaluating parameter combinations by means of a (nested) cross validation on the training data.

Table 2 provides a summary of the results in terms of the average L_1 loss (4). As can be seen, OCR often achieves clear improvements over OLR, especially on those data sets for which the response is known to depend on the predictors in a nonlinear way (e.g., CEV and CYD). Moreover, our regularization method is paying off, too, since the results of OCR+R are mostly even better than those of OCR.

As already mentioned, the Choquet integral offers interesting means to support model interpretation, notably measures of attribute importance and interaction [17]. As an example, we derived the Shapley value, which measures the importance of individual criteria, for temperature, time and concentration in the color data. These values, shown in Table 3, nicely agree with domain knowledge: temperature and concentration are more important than time, i.e., a high temperature and a high concentration have a stronger influence on the color yield than the time of dyeing. Moreover, comparing temperature and concentration, the former tends to be slightly more important than the latter for Mono Azo and Anthraquinone colors (Blue, Brown, Red), and vice versa for Diazo colors (Yellow).

7. Summary and conclusion

This paper is a continuation of previous work [2,3], in which we have advocated the use of the discrete Choquet integral in the context of classification with monotonicity constraints. More specifically, the idea of our approach, called choquistic regression, is to use the Choquet integral for representing a latent utility function in the logistic regression model. Thus, it becomes possible to capture nonlinear dependencies and interactions among predictor variables in a convenient way. Hitherto, choquistic regression was restricted to the case of binary classification. Here, we have proposed an extension of this method to the case of ordinal classification.

As already pointed out in [2,3], nonlinearity in logistic regression can of course also be incorporated in other ways, for example by using polynomials of higher degrees instead of linear functions in the logistic model. Then, however, some important properties of logistic regression may get lost. First, ensuring monotonicity may become quite difficult for nonlinear functions such as polynomials. Moreover, the comprehensibility of the (linear) logistic model might be affected. The Choquet integral is especially appealing from this point of view, as it offers measures of the importance of individual attributes (Shapley value) and the interaction among subsets of attributes.

In future work, our method ought to be compared, both conceptually and experimentally, with related work on using the Choquet integral in classification, such as [4,18]. Moreover, we plan to extend the empirical validation of choquistic regression by searching for further benchmark data for monotone ordinal classification as well as real applications.

Acknowledgments: This work was supported by the German Research Foundation (DFG). We thank Maryam Nasiri for providing us the color yield data.

References

[1] D. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, 2nd edition, 2000.

[2] A. Fallah Tehrani, W. Cheng, and E. Hüllermeier. Choquistic regression: Generalizing logistic regression using the Choquet integral. In S. Galichet, J. Montero, and G. Mauris, editors, *Proceedings EUSFLAT–2011, 7th International Conference of the European Society for Fuzzy Logic and Technology*, pages 868–875, Aix-les-Bains, France, 2011.

[3] A. Fallah Tehrani, W. Cheng, K. Dembczynski, and E. Hüllermeier. Learning monotone nonlinear models using the Choquet integral. *Machine Learning*, 89(1):183–211, 2012.

[4] G. Beliakov and S. James. Citation-based journal ranks: the use of fuzzy measures. *Fuzzy Sets and Systems*, 167(1):101–119, 2011.

[5] G. Vitali. Sulla definizione di integrale delle funzioni di una variabile. *Annali di Matematica Pura ed Applicata*, 2(1):111–121, 1925.

[6] G. Choquet. Theory of capacities. *Annales de l'institut Fourier*, 5:131–295, 1954.

[7] R. Yager. Generalized OWA aggregation operators. *Fuzzy Optimization and Decision Making*, 3(1):93–107, 2004.

[8] M. Sugeno. *Theory of Fuzzy Integrals and its Application*. PhD thesis, Tokyo Institute of Technology, 1974.

[9] F. Modave and M. Grabisch. 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, pages 164–171. Editions EDK, 1998.

[10] F. Bach. High-dimensional non-linear variable selection through hierarchical kernel learning, September 2009.

[11] A. Ben-David, L. Sterling, and Y. H. Pao. Learning and classification of monotonic ordinal concepts. *Computational Intelligence*, 5(1):45–49, 1989.

[12] R. Potharst and A. Feelders. Classification trees for problems with monotonicity constraints. *ACM SIGKDD Explorations Newsletter*, 4(1):1–10, 2002.

[13] W. Duivesteijn and A. Feelders. Nearest neighbour classification with monotonicity constraints. In *Machine Learning and Knowledge Discovery in Databases*, volume 5211 of *Lecture Notes in Computer Science*, pages 301–316. Springer, 2008.

[14] K. Dembczyński, W. Kotlowski, and R. Slowinski. Learning rule ensembles for ordinal classification with monotonicity constraints. *Fundamenta Informaticae*, 94(2):163–178, 2009.

[15] A. Feelders. Monotone relabeling in ordinal classification. In *Proceedings of the 10th IEEE International Conference on Data Mining*, pages 803–808. IEEE Computer Society, 2010.

[16] M. Nasiri. Fuzzy regression modeling of colour yield in dyeing polyester with disperse dyes. Master's thesis, Textile Engineering Department, Isfahan University of Technology, 2003.

[17] M. Grabisch. The representation of importance and interaction of features by fuzzy measures. *Pattern Recognition Letters*, 17(6):567–575, 1996.

[18] S. Angilella, S. Greco, and B. Matarazzo. 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*, pages 1194–1199. IFSA/EUSFLAT, 2009.