Choquistic Regression: Generalizing Logistic Regression Using the Choquet Integral

Ali Fallah Tehrani, Weiwei Cheng, Eyke Hüllermeier
Knowledge Engineering & Bioinformatics Lab
Department of Mathematics and Computer Science
Marburg University, Germany

Philipps Universität Marburg

Outline of the Talk

- **Contribution:**
  We introduce a new method for (probabilistic) **binary classification**, called **choquistic regression**, which generalizes conventional logistic regression and takes advantage of the **Choquet integral** as a flexible and expressive aggregation operator.

- **Outline:**

  (1) Background on **logistic regression**

  (2) Generalization to **choquistic regression**

  (3) First **experimental results**

Logistic Regression

- **Logistic regression** modifies linear regression for the purpose of predicting (probabilities of) **a binary class label** instead of real-valued responses.

- The basic model:

log-odds ratio →

$$\log\left(\frac{\mathbf{P}(y=1\,|\,\boldsymbol{x})}{\mathbf{P}(y=0\,|\,\boldsymbol{x})}\right) = w_0 + \sum_{i=1}^{m} w_i \cdot x_i$$

$$= w_0 + \boldsymbol{w}^\top \boldsymbol{x} \ ,$$

↑ linear function of predictor variables

where

– $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)^\top \in \mathbb{R}^m$ is an instance to be classified,

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

– $w_0 \in \mathbb{R}$ is a constant bias (the intercept).

Logistic Regression: Class Prediction

- Equivalently, this can be expressed in terms of **posterior probabilities**:

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

$$\mathbf{P}(y = 0 \,|\, \boldsymbol{x}) = 1 - \mathbf{P}(y = 1 \,|\, \boldsymbol{x})$$

- **Predictions** are typically made using the following decision rule:

$$\hat{y} = \begin{cases} 0 & \text{if } \mathbf{P}(y = 1 \,|\, \boldsymbol{x}) < 1/2 \\ 1 & \text{if } \mathbf{P}(y = 1 \,|\, \boldsymbol{x}) \geq 1/2 \end{cases}$$

Logistic Regression: Parameter Estimation

- The parameters of the model (bias, regression coefficients) can be obtained through **Maximum Likelihood (ML)** estimation.
- Given a sample of i.i.d. data

$$\mathcal{D} = \left\{ (\boldsymbol{x}^{(i)}, y^{(i)}) \right\}_{i=1}^{n} \subset (\mathbb{R}^m \times \{0,1\})^n \quad ,$$

the likelihood function is given by

$$\prod_{i=1}^{n} \mathbf{P}\left(y = y^{(i)} \,|\, \boldsymbol{x}^{(i)} \right) \quad ,$$

and the **ML estimate** is the maximizer of (the log of) this function:

$$(\hat{w}_0, \hat{\boldsymbol{w}}) = \arg \max_{(w_0, \boldsymbol{w})} \sum_{i=1}^{n} y^{(i)} \log \theta^{(i)}(w_0, \boldsymbol{w}) + (1 - y^{(i)}) \log \left(1 - \theta^{(i)}(w_0, \boldsymbol{w}) \right)$$

with

$$\theta^{(i)}(w_0, \boldsymbol{w}) = \left(1 + \exp(-w_0 - \boldsymbol{w}^\top \boldsymbol{x}^{(i)}) \right)^{-1}$$

Logistic Regression: Important Features

- Logistic regression is very popular and widely used in practice.
- It is **comprehesible** and easy to **interpret**, especially since the influence of each variable can easily be captured from the model:

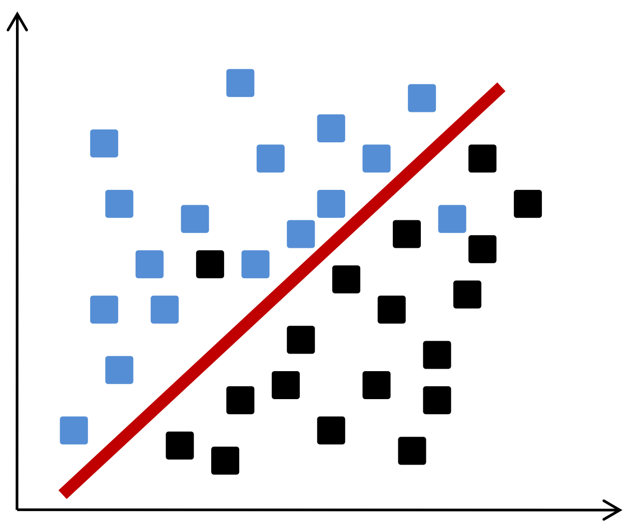
$$\log\left(\frac{\mathbf{P}(y=1\,|\,\boldsymbol{x})}{\mathbf{P}(y=0\,|\,\boldsymbol{x})}\right) = w_0 + \boxed{w_1} \cdot x_1 + w_2 \cdot x_2 + \ldots + w_m \cdot x_m$$

↑ direction and strength of influence of the first variable on the log-odds ratio (probability of positive class)

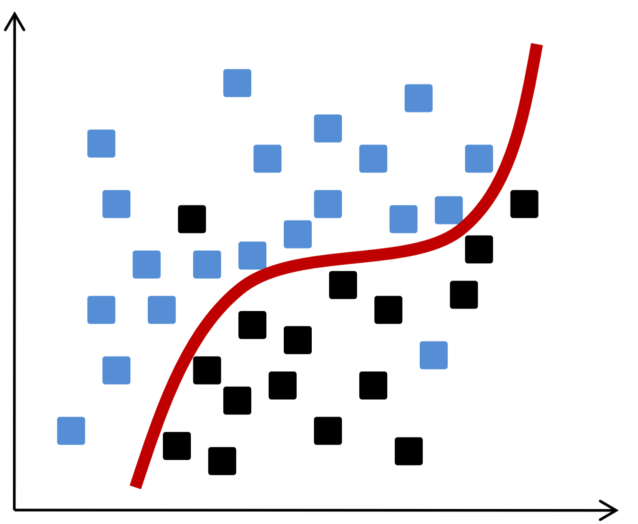
- Moreover, **monotonicity** can easily be assured by fixing the sign of regression coefficients: If a variable increases, then the probability of the positive class must only increase (decrease)!
  → *this is crucial in many applications (e.g., medicine)*
  → *violation of monotonicity may often lead to the refusal of a model*

From Logistic to Choquistic Regression

- A disadvantage of logistic regression is a **lack of flexibility**:
  In many applications, the assumption of a **linear dependency** (between predictor variables and log-odds ratio), and hence a **linear decision boundary** in the instance space, is not valid!



linear decision boundary



nonlinear decision boundary

From Logistic to Choquistic Regression

- A disadvantage of logistic regression is a **lack of flexibility**:
  In many applications, the assumption of a **linear dependency** (between predictor variables and log-odds ratio), and hence a **linear decision boundary** in the instance space, is not valid!

- Key question addressed in this paper:

How to increase the flexibility of logistic regression without losing its advantages of interpretability and monotonicity?

- Our general idea is to replace the linear model by the **Choquet integral** as a more flexible operator for aggregating the input attributes!

From Logistic to Choquistic Regression

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

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

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

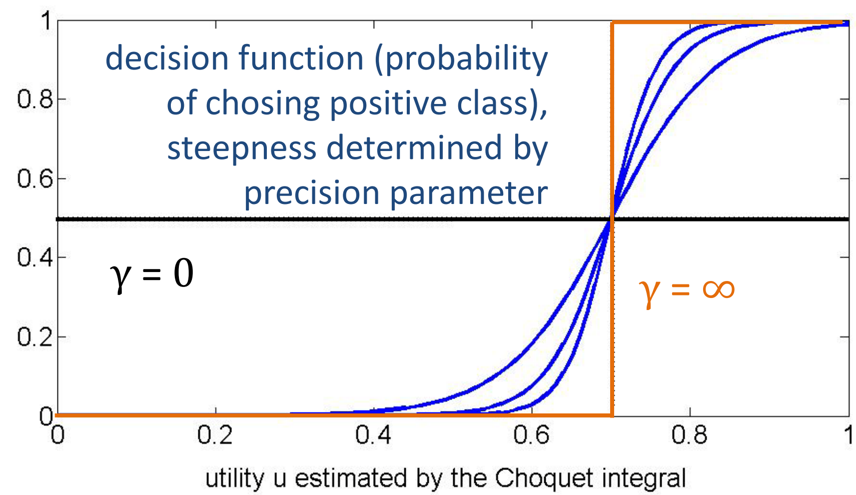
(2) a discrete choice is made by thresholding u "probabilistically" at β

Probabilistic 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)



Discrete Choquet Integral: A Brief Reminder

A **fuzzy measure** on $C = \{c_1, c_2, \ldots, c_m\}$ is a set function $\mu : 2^C \to [0,1]$ which is

- monotonic: $\mu(A) \leq \mu(B)$ for $A \subseteq B \subseteq C$
- normalized: $\mu(\emptyset) = 0$ and $\mu(C) = 1$

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.

Choquistic Regression: Interpretation

- The fuzzy 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 this 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 **intercation index** can be used, respectively, to quantify the importance of individual and the interaction between different variables.

- **Monotonicity** is obviously assured by the Choquet integral, too.

Choquistic Regression: Parameter Estimation

- We need to identify the following model parameters:
  - the fuzzy measure μ
  - the utility threshold β
  - the precision parameter γ
- The fuzzy 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*
- Again, 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 fuzzy measure

→ solution with sequential quadratic programming

Experimental Evaluation

- Experimental comparison with monotone logistic regression
- Collection of data sets for which monotonicity is a plausible assumption
- Classification error determined by means of cross validation

data set	logistic	choquistic
ESL	0.0621 ± 0.0096	**0.0547** ± 0.0105
ERA	0.2849 ± 0.0140	**0.2756** ± 0.0170
LEV	0.1669 ± 0.0134	**0.1340** ± 0.0115
DBS	**0.1443** ± 0.0371	0.1560 ± 0.0405
CPU	0.0400 ± 0.0093	**0.0119** ± 0.0138
CEV	0.1883 ± 0.0066	**0.0346** ± 0.0076
CYD-1	0.1254 ± 0.0074	**0.0729** ± 0.0066
CYD-2	0.2004 ± 0.0091	**0.0717** ± 0.0078
CYD-3	0.1512 ± 0.0238	**0.0762** ± 0.0163
CYD-4	0.1289 ± 0.0253	**0.0496** ± 0.0201
CYD-5	0.1242 ± 0.0099	**0.0204** ± 0.0057
CYD-6	0.1604 ± 0.0085	**0.0383** ± 0.0083
CYD-7	0.1958 ± 0.0207	**0.0646** ± 0.0089

Main results

- Choquistic regression achieves consistent gains
- Higher interaction between variables tends to come with higher gain

Conclusions & Outlook

- We introduced a new method called **choquistic regression**, a generalization of conventional logistic regression for **binary classification**.
- **Choquistic regression**
  - combines **probabilistic modeling** underlying logistic regression with the advantages of the **Choquet integral as a flexible aggregation operator**, notably its capability to capture **interactions between predictor variables**;
  - thereby, it becomes possible to **increase flexibility** while preserving core features of logistic regression, namely **interpretability and monotonicity**.
- First **experimental results** confirm advantages of choquistic regression in terms of predictive accuracy.
- **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?

Back up (Influence of precision parameter)