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

The basic model:

\[
\log \left( \frac{P(y = 1 \mid \mathbf{x})}{P(y = 0 \mid \mathbf{x})} \right) = w_0 + \sum_{i=1}^{m} w_i \cdot x_i
\]

\[
= w_0 + \mathbf{w}^\top \mathbf{x},
\]

where

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

- \( \mathbf{w} = (w_1, w_2, \ldots, w_m)^\top \in \mathbb{R}^m \) 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:

\[
P(y = 1 \mid x) = \left(1 + \exp(-w_0 - \mathbf{w}^\top x)\right)^{-1}
\]

\[
P(y = 0 \mid x) = 1 - P(y = 1 \mid x)
\]

- Predictions are typically made using the following decision rule:

\[
\hat{y} = \begin{cases} 
0 & \text{if } P(y = 1 \mid x) < 1/2 \\
1 & \text{if } P(y = 1 \mid x) \geq 1/2 
\end{cases}
\]
The parameters of the model (bias, regression coefficients) can be obtained through **Maximum Likelihood (ML)** estimation.

Given a sample of i.i.d. data

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

the likelihood function is given by

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

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

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

with

$$\theta^{(i)}(w_0, \mathbf{w}) = \left( 1 + \exp\left( -w_0 - \mathbf{w}^\top \mathbf{x}^{(i)} \right) \right)^{-1}$$
Logistic regression is very popular and widely used in practice. It is comprehensible and easy to interpret, especially since the influence of each variable can easily be captured from the model:

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

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
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!
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!
It can be shown that, by choosing the parameters in a proper way, logistic regression is indeed a special case of Choquistic regression.
Interpretation of choquistic regression as a two-stage process:

1. A (latent) utility degree \( u = C_\mu(x) \in [0, 1] \) is determined by the Choquet integral.
2. A discrete choice is made by thresholding \( u \) "probabilistically" at \( \beta \).

Probabilistic thresholding:

\[
P(y = 1) = \frac{1}{1 + \exp\left(-\gamma (C_\mu(x) - \beta)\right)}
\]

- Precision of the model determines \( \gamma = \infty \).
- Utility threshold determines \( \gamma = 0 \).
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 $\mu$ is defined as follows:

$$C_\mu(f) = \sum_{i=1}^{m} (f(c_i) - f(c_{i-1})) \cdot \mu(A_{(i)}) ,$$

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.
The fuzzy measure $\mu$ 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 $\{\text{smoking, age}\}$ 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.
We need to identify the following model parameters:

- the fuzzy measure $\mu$
- the utility threshold $\beta$
- the precision parameter $\gamma$

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:

$$C_\mu(f) = \sum_{T \subseteq C} m(T) \times \min_{c_i \in T} f(c_i).$$
ML estimation leads to a constrained optimization problem:

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

subject to:

- $$0 \leq \beta \leq 1$$ conditions on utility threshold and precision
- $$0 < \gamma$$
- $$\sum_{T \subseteq C} m(T) = 1$$ normalization and monotonicity of the fuzzy measure
- $$\sum_{B \subseteq A \setminus \{c_i\}} m(B \cup \{c_i\}) \geq 0 \quad \forall A \subseteq C, \forall c_i \in C$$

→ 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

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

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)

- **ERA**
  - $\gamma = 8.3$

- **CEV**
  - $\gamma = 69.2$

- **LEV**
  - $\gamma = 15.8$

- **ESL**
  - $\gamma = 396.8$