São José dos Campos - SP
LV SBPO
LV Simpósio Brasileiro de Pesquisa Operacional
São José dos Campos, SP - 6 a 9 de novembro de 2023

Uma nova proposta para a regressão logística com base na integral de Choquet

Guilherme Dean Pelegrina, Leonardo Tomazeli Duarte
Faculdade de Ciências Aplicadas, Universidade Estadual de Campinas
R. Pedro Zaccaria, 1300, Limeira - SP, 13484-350
guidean@unicamp.br, leonardo.duarte@fca.unicamp.br

Michel Grabisch
Centre d'Économie de la Sorbonne, Université Paris I Panthéon-Sorbonne
Maison des Sciences Économiques, 106-112 bd de l'Hôpital, 75647 Paris Cedex 13
Michel.Grabisch@univ-paris1.fr

RESUMO

A regressão logística é uma ferramenta importante em aprendizado de máquina. Ela pode ser vista como uma extensão da regressão linear para problemas de classificação. De fato, a formulação da regressão logística também usa um agregador linear, porém, dentro da função sigmoide. Como o agregador não permite modelar interações entre atributos, o que muitas vezes é necessário em problemas práticos, trabalhos na literatura vêm propondo o uso da integral de Choquet dentro da sigmoide. No entanto, esses trabalhos definem a integral de Choquet com base em uma capacidade, o que dificulta a etapa de treinamento devido à inclusão de diversas restrições ao modelo. Neste artigo, propomos uma nova formulação da integral de Choquet, baseada em um jogo. Neste contexto, eliminamos as restrições do modelo e etapa de treinamento se torna similar ao da regressão logística. Avaliamos nossa proposta a partir de experimentos preliminares em vários dados reais usados na literatura.

PALAVRAS CHAVE. Regressão logística, Integral de Choquet, Aprendizado de máquina.

IC – Inteligência Computacional, EST&MP – Estatística e Modelos Probabilísticos

ABSTRACT

Logistic regression is an important tool in machine learning. It can be seen as an extension of the linear regression to classification problems. Indeed, the logistic regression formulation also uses a linear aggregator, however, within the sigmoid function. As this aggregator does not allow modeling interactions between attributes, which is often required in practical problems, some works in the literature proposed the use of the Choquet integral within the sigmoid. However, these works define the Choquet integral based on a capability, which brings a difficulty into the training step due to the inclusion of several constraints to the model. In this paper, we propose a new formulation of the Choquet integral, based on a game. In this context, we eliminate model constraints and the training step becomes similar to the logistic regression. We evaluated our proposal based on preliminary experiments on several real datasets used in the literature.

KEYWORDS. Logistic regression, Choquet integral, Machine learning.

IC - Computational Intelligence, EST&MP – Statistics and Probabilistic Models

DOI: 10.59254/sbpo-2023-175136
http://live-proceedings.pantheonsite.io/p/175136

1. Introdução

A adoção de modelos automatizados, baseados no aprendizado de máquina, vem sendo cada vez mais frequente nos dias atuais. São inúmeras aplicações com resultados práticos, tanto no contexto empresarial [Lee e Shin, 2020] quanto no cotidiano das pessoas [Sarker, 2021]. Junto a este crescente uso, vem a preocupação em se adotar modelos simples, os quais são da fácil implementação e que não exigem tanto esforço computacional. Via de regra, modelos complexos, como as redes neurais de arquitetura profunda [LeCun et al., 2015], atingem uma acurácia elevada em tarefas de regressão ou classificação. No entanto, esse modelos são de difícil interpretação e exigem um grande esforço computacional para o treinamento do algoritmo. Dessa forma, modelos mais simples surgem como opção devido à facilidade de uso.

Um dos modelos mais simples de aprendizado de máquina usados em tarefas de classificação é a regressão logística [Kleinbaum e Klein, 2002]. Em resumo, podemos definir a regressão logística como uma extensão da regressão linear para tarefas de classificação. Ao passo que a regressão linear agrega as variáveis dependentes a partir de uma função linear, a regressão logística usa a função sigmoide para modelar a probabilidade da ocorrência de um determinado evento. A função linear é, então, introduzida dentro da função sigmoide. Dessa forma, a implementação do modelo, assim como a obtenção dos parâmetros é relativamente simples comparado a modelos mais complexos em aprendizado de máquina.

Embora a regressão logística venha sendo utilizada em problemas práticos, a acurácia do modelo é limitada pelo agregador linear embutido na função sigmoide. Com o objetivo de aumentar a flexibilidade do modelo de regressão logística sem alterar sua essência, Tehrani et al. [2011] propuseram o uso da integral de Choquet [Choquet, 1954] como função de agregação dentro da sigmoide. Os parâmetros da integral de Choquet, chamados de coeficientes de capacidade, permitem modelar interações entre atributos. Embora a integral de Choquet seja uma função não linear, podemos defini-la de maneira que ela seja linear nos parâmetros e, então, a função custo fica similar à da regressão logística tradicional. Mas, modelando interações entre atributos, ganha-se no incremento da flexibilidade do modelo.

Um aspecto interessante no modelo proposto por Tehrani et al. [2011], chamado de *Choquistic Regression*, é que a integral de Choquet é definida em termos dos coeficientes de capacidade, os quais precisam respeitar restrições de monotonicidade e duas condições de contorno. No entanto, como discutido na literatura [Grabisch, 2016; Pelegrina et al., 2022], a integral de Choquet pode ser definida em relação a um jogo (cooperativo) ao invés da capacidade. Com isso, tanto as restrições de monotonicidade quanto uma das condição de contorno não são mais necessárias, o que facilita ainda mais a aprendizagem dos parâmetros.

Neste contexto, este artigo propõe uma formulação da regressão logística com base na integral de Choquet na qual os parâmetros são definidos a partir de um jogo cooperativo. Além disso, exploramos o conceito da jogos 2-aditivos, o que reduz o número de parâmetros a serem aprendidos. Nessa formulação, obtemos um modelo de regressão logística mais generalista em comparação ao modelo clássico com o agregador linear e que possui maior flexibilidade que o proposto por Tehrani et al. [2011], uma vez que eliminamos restrições acerca dos parâmetros. A partir de experimentos numéricos em diversos dados reais usados na literatura, verificamos o ganho na acurácia a partir da nossa proposta em comparação com a obtida pela regressão logística clássica.

As próximas seções deste artigo se dividem da seguinte forma. Na Seção 2, apresentamos as notações usadas nesse artigo acerca do problema de classificação. A Seção 3 discute os aspectos teóricos usados em nossa proposta. Mais precisamente, descrevemos o modelo de regressão logística, definimos a integral de Choquet e mostramos como ambos são combinados para

DOI: 10.59254/sbpo-2023-175136
http://live-proceedings.pantheonsite.io/p/175136

lidar com problemas de classificação em aprendizado de máquina. Nossa proposta é apresentada na Seção 4. Os experimentos numéricos preliminares usados para avaliar nossa proposta se encontram na Seção 5. Por fim, apresentamos as considerações finais e perspectivas futuras na Seção 6.

2. O problema de classificação

Neste artigo, abordamos o problema de classificação. Seja $\mathbf{X} \in \mathbb{R}^{n \times m}$ o conjunto de n amostras (objetos, pessoas, projetos, etc.) as quais são caracterizadas por m atributos. O objetivo é treinar um modelo de aprendizado de máquina capaz de classificar cada amostra em uma classe específica, dado um certo número fixo de classes. Representamos cada amostra por $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,m}]$ e sua classe associada por y_i, $i = 1, \ldots, n$. Suponha o caso da classificação binária[1], ou seja, quando as amostras são classificadas na classe 0 ou na classe 1. Além disso, considere $\mathbf{y} \in \{0, 1\}^n$ como o vetor que contém as classes das amostras em $\mathbf{X}$.

A fim de construir o modelo de classificação, primeiramente, definimos a função $\hat{y}_i = g(\mathbf{x}_i)$ que indicará qual a classe que cada amostra será classificada. O intuito é ajustar os parâmetros de $g(\cdot)$ a fim de minimizar o erro entre as classes preditas $\hat{y}_i$ e as classes verdadeiras y_i, $i = 1, \ldots, n$. Seja $\beta = [\beta_0, \beta_1, \ldots, \beta_p]$ os parâmetros de $g(\cdot)$. Procuramos, então, resolver o seguinte problema de otimização:

$$\min_{\beta_0, \beta_1, \ldots, \beta_p} \sum_{i=1}^{n} \mathcal{L}\left(\hat{y}_i, y_i\right), \tag{1}$$

onde $\mathcal{L}(\hat{y}_i, y_i)$ representa uma medida de dissimilaridade entre $\hat{y}_i$ e y_i.

Há diversos modelos de classificação e medidas de dissimilaridades que podem ser usados no problema de classificação [Kotsiantis, 2007]. Neste estudo, abordaremos a regressão logística, a qual usa a entropia cruzada como medida de dissimilaridade. Discutiremos esse método na sequência.

3. Fundamentação teórica

Nesta seção, discutimos brevemente os aspectos teóricos sobre a regressão logística e a integral de Choquet. Além disso, apresentamos o modelo *Choquistic Regression*, proposto por Tehrani et al. [2011].

3.1. Regressão logística

Como mencionado na Seção 1, a regressão logística surge como uma opção simples, do ponto de vista computacional e de implementação, para tratar problemas de classificação. A ideia consiste em estimar a probabilidade que cada amostra $\mathbf{x}_i$ seja classificada como a classe 1 (relembre que estamos no caso da classificação binária). Para isso, consideramos a função sigmoide, o que leva à seguinte expressão:

$$\hat{y}_i = P(y_i = 1) = \frac{1}{1 + \exp\left(-\left(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_m x_{i,m}\right)\right)}. \tag{2}$$

Note que o agregador linear $f(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_m x_{i,m}$ se encontra dentro da exponenciação. Quanto maior (resp., quanto menor) $f(\mathbf{x}_i)$, maior (resp., menor) a chance da amostra $\mathbf{x}_i$ ser classificada como sendo da classe 1.

Para ajustar os parâmetros da regressão logística, usamos a entropia cruzada como medida de dissimilaridade. Claramente, no caso da classificação binária, se $\hat{y}_i = P(y_i = 1)$ representa a probabilidade da amostra $\mathbf{x}_i$ ser classificada como sendo da classe 1, $P(y_i = 0) = 1 - P(y_i =$

[1] Embora adotamos a classificação binária para ilustrar a nossa proposta, podemos facilmente estender para o caso multiclasse.

DOI: 10.59254/sbpo-2023-175136
http://live-proceedings.pantheonsite.io/p/175136

$1) = 1 - \hat{y}_i$ representa a probabilidade da amostra $\mathbf{x}_i$ ser classificada como sendo da classe 0. Essa observação ajuda a definir a medida de dissimilaridade baseada na entropia cruzada como

$$\mathcal{L}_{EC}(\hat{y}_i, y_i) = -\left[y_i \log \hat{y}_i + (1-y)\log\left(1-\hat{y}_i\right)\right]. \tag{3}$$

Por fim, na regressão logística, lidamos com o seguinte modelo de otimização:

$$\min_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{i=1}^{n} -\left[y_i \log \hat{y}_i + (1-y)\log\left(1-\hat{y}_i\right)\right], \tag{4}$$

onde $\hat{y}_i$ é a função sigmoide definida na Equação (2).

3.2. Integral de Choquet e o conceito de jogos k-aditivos

Muito utilizada em tomada de decisão multicritério [Grabisch, 1996; Grabisch e Labreuche, 2010], a integral de Choquet [Choquet, 1954] consiste em uma função de agregação não linear que permite modelar interações entre atributos. Ela é definida como

$$f_{CI}(\mu, \mathbf{x}_i) = \sum_{i=1}^{m} (x_{i,(j)} - x_{i,(j-1)})\mu(\{(j), \ldots, (m)\}), \tag{5}$$

onde $x_{i,(j)}$ indica uma permutação dos índices j tal que $x_{i,(1)} \leq x_{i,(2)} \leq \ldots \leq x_{i,(m)}$ (com $x_{i,(0)} = 0$) e a capacidade $\mu(\cdot)$, definida no conjunto $M = \{1, \ldots, m\}$ de m atributos, consiste em uma função de conjunto $\mu : 2^M \to \mathbb{R}_+$ satisfazendo os seguintes axiomas:

- $A \subseteq B \subseteq M \Rightarrow \mu(A) \leq \mu(B) \leq \mu(M)$ (monotonicidade),
- $\mu(\emptyset) = 0$ e $\mu(M) = 1$ (condições de contorno).

Os parâmetros do modelo ($\mu(A)$ - coeficientes de capacidade) são, então, definidos para cada coalizão de atributos A. Um aspecto interessante associado a esses coeficientes é que eles permitem uma representação em diversos outros domínios [Grabisch, 1997]. Por exemplo, a transformação de Möbius em $\mu(\cdot)$, dada por

$$\eta^{\mu}(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|}\, \mu(B), \;\; \forall A \subseteq M, \tag{6}$$

onde $|A|$ indica a cardinalidade do subconjunto A, permite uma relação linear entre os parâmetros $\eta^{\mu}(A)$ e os coeficientes de capacidade $\mu(A)$. A partir de $\eta^{\mu}(A)$, para todo $A \subseteq M$, recuperamos os coeficientes de capacidade a partir de seguinte expressão:

$$\mu(A) = \sum_{B \subseteq A} \eta^{\mu}(B). \tag{7}$$

Com base na transformação descrita na Equação (7), podemos reescrever a integral de Choquet como

$$f_{CI}(\eta^{\mu}, \mathbf{x}_i) = \sum_{A \subseteq M} \eta^{\mu}(A) \bigwedge_{j \in A} x_{i,j}, \tag{8}$$

onde $\bigwedge$ representa o operador de mínimo.

A representação da capacidade através da transformação de Möbius facilita a definição da integral de Choquet com base no conceito de jogos k-aditivos. Em teoria dos jogos cooperativos,

DOI: 10.59254/sbpo-2023-175136
http://live-proceedings.pantheonsite.io/p/175136

um jogo é dito k-aditivo se $\eta^{\mu}(A) = 0$ para todas as coalizões A tal que $|A| > k$. Dessa forma, ao considerar um jogo 2-aditivo, por exemplo, reduzimos o número de parâmetros do modelo de 2^m para $m(m+1)/2$. Ou seja, ao invés do número de parâmetros crescer exponencialmente com o número de atributos, ele passa a crescer de maneira polinomial. Isso ajuda significativamente na etapa de aprendizagem, uma vez que o número de parâmetros a serem estimados reduz consideravelmente. Claramente, o modelo também reduz a flexibilidade em modelar todas as interações entre atributos. No entanto, modelar interações entre pares de atributos já traz uma grande vantagem em relação ao agregador linear.

Ao considerarmos o modelo 2-aditivo, a integral de Choquet pode ser definida como

$$f_{CI}^{2add}(\eta^{\mu}, \mathbf{x}_i) = \sum_{j \in M} x_{i,j}\mu(i) + \sum_{j,j' \in M} \left(\mu(j,j') - \mu(j) - \mu(j')\right) \wedge \left(x_{i,j}, x_{i,j'}\right). \tag{9}$$

3.3. *Choquistic Regression*

Na literatura [Tehrani et al., 2011], a integral de Choquet definida a partir de uma capacidade (ou sua transformação para o domínio de Möbius) vem sendo adotada a fim de permitir modelar interações entre atributos no modelo de regressão logística. O modelo proposto por Tehrani et al. [2011], chamado *Choquistic Regression*, consiste em substituir o agregador linear $f(\cdot)$ na regressão logística pela integral de Choquet $f_{CI}(\cdot)$ (ou $f_{CI}^{2add}(\cdot)$ se assumirmos uma capacidade 2-aditiva). Com isso, as probabilidades associadas às amostras $\mathbf{x}_i$ são dadas por

$$\hat{y}_i = P(y_i = 1) = \frac{1}{1 + \exp\left(-\gamma\left(f_{CI}(\eta^{\mu}, \mathbf{x}_i) - \beta_0\right)\right)}. \tag{10}$$

onde γ é uma constante que controla a convergência do modelo (veja [Tehrani et al., 2011] para maiores detalhes). Com isso, o modelo de otimização a ser resolvido é o seguinte:

$$\begin{aligned} \min_{\beta_0, \mu(A), A \subseteq M} \sum_{i=1}^{n} - \Bigg[& y_i \log \frac{1}{1 + \exp\left(-\gamma\left(f_{CI}(\eta^{\mu}, \mathbf{x}_i) - \beta_0\right)\right)} \\ & + (1-y) \log\left(1 - \frac{1}{1 + \exp\left(-\gamma\left(f_{CI}(\eta^{\mu}, \mathbf{x}_i) - \beta_0\right)\right)}\right) \Bigg]. \end{aligned} \tag{11}$$

$$\begin{aligned} \text{sujeito à } & \mu(A) \leq \mu(B), \forall A \subseteq B \\ & \mu(\emptyset) = 0 \\ & \mu(M) = 1 \end{aligned}$$

No entanto, como apontado pelos autores, ao assumir a integral de Choquet definida pelos coeficientes de capacidade, as condições de contorno ($\mu(\emptyset) = 0$ e $\mu(M) = 1$) e as restrições de monotonicidade (se $A \subseteq B \subseteq M$, $\mu(A) \leq \mu(B) \leq \mu(M)$), o problema de otimização tem muitas restrições, o que pode impor uma dificuldade computacional ao resolvê-lo. A fim de suavizar tal inconveniente, modelos heurísticos foram propostos para estimar os coeficientes de capacidade com menor esforço computacional [Tehrani, 2021].

Um outro aspecto no modelo desenvolvido por Tehrani et al. [2011] é que, ao considerar as restrições de monotonicidade, é necessário corrigir a direção de maximização dos atributos. Ou seja, caso o aumento no valor de um determinado atributo contribui para que a amostra seja classificada como pertencente à classe 0, esse atributo deve ser normalizado a fim de garantir que, quanto maior seu valor, maior a chance da amostra ser classificada como pertencente à classe 1.

DOI: 10.59254/sbpo-2023-175136
http://live-proceedings.pantheonsite.io/p/175136

4. Modelo proposto

Essa seção discorre sobre nossa proposta para generalizar o uso da integral de Choquet na regressão logística a partir da sua definição com base em um jogo. É conhecido na literatura [Grabisch, 2016; Pelegrina et al., 2022] que a integral de Choquet pode ser definida a partir de um jogo υ ao invés da capacidade μ. O interessante, nesse caso, é que ao se adotar o jogo υ, não é mais necessário respeitar as restrições de monotonicidade. Além disso, a condição de limitante superior para o jogo (o que era $\mu(M) = 1$ no caso da capacidade) também é desnecessária, uma vez que não restringimos a valor do jogo quando todos os atributos estão na coalizão. Sendo assim, a única condição que permanece no modelo é $\upsilon(\emptyset) = 0$, a qual pode ser facilmente incluída na integral de Choquet sem a necessidade de considerá-la como uma restrição do modelo. De fato, como pode ser visto[2] na Equação (7), $\upsilon(\emptyset) = 0$ implica que $\eta^{\upsilon}(\emptyset) = 0$. Dessa forma, basta retirarmos $\eta^{\upsilon}(\emptyset)$ da integral de Choquet (e, consequentemente, do modelo de aprendizagem) para que essa condição seja satisfeita.

Com as adaptações mencionadas acima acerca da formulação da integral de Choquet a partir de um jogo, o modelo de otimização que lidamos em nossa proposta é o seguinte:

$$\min_{\beta_0,\upsilon(A),A\subseteq M} \sum_{i=1}^{n} -\Bigg[y_i \log \frac{1}{1+\exp\left(-\gamma\left(f_{CI}(\eta^{\upsilon},\mathbf{x}_i)-\beta_0\right)\right)} + (1-y)\log\left(1-\frac{1}{1+\exp\left(-\gamma\left(f_{CI}(\eta^{\upsilon},\mathbf{x}_i)-\beta_0\right)\right)}\right)\Bigg], \tag{12}$$

onde

$$f_{CI}(\eta^{\upsilon},\mathbf{x}_i) = \sum_{A\subseteq M, A\neq\emptyset} \eta^{\upsilon}(A) \bigwedge_{j\in A} x_{i,j}, \tag{13}$$

ou, no caso do modelo 2-aditivo,

$$f_{CI}^{2add}(\eta^{\upsilon},\mathbf{x}_i) = \sum_{j\in M} x_{i,j}\upsilon(i) + \sum_{j,j'\in M} \left(\upsilon(j,j') - \upsilon(j) - \upsilon(j')\right) \wedge \left(x_{i,j}, x_{i,j'}\right). \tag{14}$$

Com as adaptações mencionadas acima, ao retirar as restrições do modelo de otimização, a nossa proposta fica mais simples de ser resolvida em comparação com o modelo proposto por Tehrani et al. [2011]. De fato, a resolução do problema (12) se assemelha à resolução da regressão logística tradicional (com mais parâmetros, claramente), uma vez que o agregador dentro da sigmoide é linear nos parâmetros. Assim, em relação ao método de resolução, podemos adotar a mesma estratégia que a aplicada à regressão logística tradicional (como comentaremos na próxima seção, usaremos implementações em Python disponíveis no *scikit-learn*). Além disso, sem as restrições de monotonicidade, os parâmetros ficam livres para assumirem valores tanto positivos quanto negativos. Isso também facilita o uso do modelo, uma vez que não há mais a necessidade em corrigir a direção de maximização dos atributos.

5. Experimentos e resultados preliminares

Com o intuito de avaliar a nossa proposta frente ao modelo clássico da regressão logística[3], aplicamos ambas as abordagens nos problemas de classificação definidos pelos seguintes conjuntos

[2] A transformação de Möbius é válida tanto em relação a uma capacidade μ quanto a um jogo υ.

[3] Vale ressaltar que não comparamos com o modelo proposto por [Tehrani et al., 2011] uma vez que o espaço de busca do *Choquistic Regression* é um subconjunto do espaço de busca do nosso modelo. Ou seja, a solução do *Choquistic Regression* também é uma solução factível do nosso modelo e, portanto, nossa solução generaliza o uso da integral de Choquet como função de agregação na regressão logística.

DOI: 10.59254/sbpo-2023-175136
http://live-proceedings.pantheonsite.io/p/175136

de dados: *Banknote authentication* (BANK), *Blood transfusion* (BLOOD), *Rice species* (RICE) e *Diabetes PIMA* (DIAB). A Tabela 1 descreve os números de amostras, atributos e classes de cada conjunto de dados.

Tabela 1: Descrição dos conjuntos de dados.

Conjunto de dados	**# de amostras**	**# de atributos**	**# de classes**
BANK [Lohweg et al., 2009]	1372	4	2
BLOOD [Yeh et al., 2009]	748	4	2
RICE [Cinar e Koklu, 2019]	3810	7	2
DIAB [Smith et al., 1988]	768	8	2

Visto que a resolução do problema de otimização em nossa proposta se assemelha ao da regressão logística, usamos a implementação *LogisticRegression* do *scikit-learn*, em Python, para encontrar os parâmetros dos seguintes modelos:

- LR: Regressão logística clássica, com agregador linear,
- CR: Regressão logística com a integral de Choquet como agregador,
- CR^{2add}: Regressão logística com a integral de Choquet 2-aditiva como agregador.

Além disso, para cada modelo de regressão logística, avaliamos o uso de diferentes *solvers* disponíveis no *scikit-learn*. São eles: *lbfgs*, *newton-cg*, *sag* e *saga*.

Para cada conjunto de dados, conduzimos um treinamento com validação cruzada (com 5 *folds*) e calculamos a acurácia média tanto na etapa de treinamento quanto nos dados de teste. Com base em um experimento com 50 repetições, as acurácias médias alcançadas nas etapas de treinamento e teste são apresentadas nas Tabelas 2 e 3, respectivamente.

Tabela 2: Acurácia nos dados de treinamento.

Conjunto de dados	***Solver***	LR	CR	CR^{2add}
BANK	lbfgs	99.1	**100**	**100**
	newton-cg	99.1	**100**	**100**
	sag	99.01	**100**	99.99
	saga	98.98	**100**	99.9
BLOOD	lbfgs	77.39	**77.86**	**77.86**
	newton-cg	77.39	**77.86**	**77.86**
	sag	77.39	**77.83**	77.81
	saga	77.39	**77.82**	77.7
RICE	lbfgs	93.11	**93.27**	93.16
	newton-cg	93.11	**93.48**	93.21
	sag	93.18	**93.24**	93.02
	saga	93.15	**93.22**	92.93
DIAB	lbfgs	77.99	**83.62**	80.59
	newton-cg	77.99	**83.71**	80.60
	sag	77.99	**83.49**	80.30
	saga	77.99	**83.42**	80.08

DOI: 10.59254/sbpo-2023-175136
http://live-proceedings.pantheonsite.io/p/175136

Tabela 3: Acurácia nos dados de teste.

Conjunto de dados	***Solver***	LR	CR	CR^{2add}
BANK	lbfgs	98.98	**99.80**	99.65
	newton-cg	98.98	99.80	**99.87**
	sag	98.88	**99.81**	99.71
	saga	98.88	**99.79**	99.51
BLOOD	lbfgs	77.18	77.33	**77.41**
	newton-cg	77.18	77.33	**77.41**
	sag	77.19	77.35	**77.47**
	saga	77.18	**77.39**	**77.39**
RICE	lbfgs	**93.03**	92.62	92.85
	newton-cg	**93.03**	92.7	92.89
	sag	**93.1**	92.74	92.85
	saga	**93.07**	92.79	92.78
DIAB	lbfgs	**77.25**	74.35	76.92
	newton-cg	**77.25**	74.38	76.92
	sag	77.25	75.39	**77.33**
	saga	77.24	75.74	**77.38**

Para os resultados na etapa de treinamento, as melhores acurácias foram obtidas a partir do modelo com a integral de Choquet sem considerações adicionais acerca dos parâmetros. Em alguns casos (para os dados BANK e BLOOD), o modelo 2-aditivo resultou em desempenhos iguais ao modelo sem essa condição. Esse resultado superior ao da regressão logística clássica já era esperado, uma vez que com a integral de Choquet temos mais flexibilidade para ajustar o classificador ao modelar interações entre atributos.

Em relação às acurácias obtidas ne etapa de teste, temos resultados mistos. Para os dados BANK e BLOOD, os modelos com a integral de Choquet alcançaram as melhores acurácias. No entanto, para os dados RICE, o melhor desempenho foi da regressão logística clássica. Isso pode ser explicado pelo número de atributos nesse conjunto de dados. Como há mais atributos que os dados BANK e BLOOD, nossa hipótese é que houve uma sobreajuste na etapa de treinamento e, portanto, o classificador ficou especialista para os dados de treinamento e não tão generalista para os dados de teste. Isso também pode ser observado nos dados DIAB, onde para alguns *solvers* a regressão logística tradicional foi melhor que os modelos com a integral de Choquet.

6. Considerações finais

Abordamos neste trabalho o uso da regressão logística como ferramenta para tratar problemas de classificação em aprendizado de máquina. Se por um lado a regressão logística é de fácil implementação e com baixo esforço computacional, sua versão clássica se limita a agregar o valor dos atributos de maneira linear. Isso pode reduzir a eficácia do modelo quando há interações entre atributos.

Neste artigo, propusemos usar a integral de Choquet como agregador na regressão logística. Nossa ideia se baseou em trabalhos da literatura que também considera o emprego da integral de Choquet, porém, definida a partir de uma capacidade. Nesses casos, há a inclusão de restrição no modelo de otimização, o que aumenta o esforço computacional na etapa de aprendizagem. Em nossa proposta, definimos a integral de Choquet com base em um jogo cooperativo. Dessa forma,

DOI: 10.59254/sbpo-2023-175136
http://live-proceedings.pantheonsite.io/p/175136

evitamos a inclusão de restrições ao modelo e, consequentemente, a resolução do problema de otimização se assemelha ao da regressão logística clássica.

Avaliamos nossa proposta em quatro conjunto de dados frequentemente utilizados na literatura. Os resultados preliminares encontrados ressaltam o interesse em se adotar agregadores como a integral de Choquet, a qual possibilita modelar interações entre atributos e permite uma maior flexibilização ao ajustar o classificador. Isso pode ser verificado nas acurácias superiores encontradas na etapa de treinamento. Porém, na etapa de teste e para os dados com maior número de atributos, a regressão logística clássica se mostrou superior para alguns *solvers* empregados. Esse resultado sugere que houve um sobreajuste na etapa de treinamento e, portanto, os modelos com a integral de Choquet não atingiram uma boa capacidade de generalização.

A fim de evitar o problema de sobreajuste, temos como perspectivas futuras empregar a regularização nos parâmetros ao resolver o problema de otimização. Com isso, esperamos que o modelo proposto não fique especialista nos dados de treinamento e pouco generalista para os de teste. Além disso, almejamos formular o modelo de regressão logística com base em outras funções de agregação que levam em conta interações entre atributos, como o modelo multilinear [Owen, 1972; Pelegrina et al., 2020]. Por fim, também temos como perspectiva avaliar o uso do método proposto em problemas de classificação multiclasse.

Agradecimentos

Os autores gostariam de agradecer aos processos n. 2020/09838-0 (BI0S - Brazilian Institute of Data Science), 2020/10572-5 e 2021/11086-0, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Referências

Choquet, G. (1954). Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295.

Cinar, I. e Koklu, M. (2019). Classification of rice varieties using artificial intelligence methods. *International Journal of Intelligent Systems and Applications in Engineering*, 7(3):188–194.

Grabisch, M. (1996). The application of fuzzy integrals in multicriteria decision making. *European Journal of Operational Research*, 89:445–456.

Grabisch, M. (1997). Alternative representations of discrete fuzzy measures for decision making. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, 5(5):587–607.

Grabisch, M. (2016). *Set Functions, games and capacities in decision making*. Springer International Publishing.

Grabisch, M. e Labreuche, C. (2010). A decade of application of the choquet and sugeno integrals in multi-criteria decision aid. *Annals of Operations Research*, 175:247–286.

Kleinbaum, D. G. e Klein, M. (2002). *Logistic regression*. Springer-Verlag, New York.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 31:249–268.

LeCun, Y., Bengio, Y., e Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

DOI: 10.59254/sbpo-2023-175136
http://live-proceedings.pantheonsite.io/p/175136

Lee, I. e Shin, Y. J. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons*, 63(2):157–170.

Lohweg, V., Gillich, E., e Schaede, J. (2009). New concepts in banknote authentication.

Owen, G. (1972). Multilinear extensions of games. *Management Science*, Part 2, 18(5):64–79.

Pelegrina, G. D., Duarte, L. T., Grabisch, M., e Romano, J. M. T. (2020). The multilinear model in multicriteria decision making: The case of 2-additive capacities and contributions to parameter identification. *European Journal of Operational Research*, 282:945–956.

Pelegrina, G. D., Duarte, L. T., e Grabisch, M. (2022). A k-additive Choquet integral-based approach to approximate the SHAP values for local interpretability in machine learning. *ArXiv ID: 2208.11362*. URL `http://arxiv.org/abs/2211.02166`.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3):160.

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., e Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Applications in Medical Care*, p. 261–265. American Medical Informatics Association.

Tehrani, A. F., Cheng, W., e Hüllermeier, E. (2011). Choquistic regression: Generalizing logistic regression using the Choquet integral. In *Proceedings of the 7th Conference of the European Society for Fuzzy Logic and Technology and French Days on Fuzzy Logic and Applications (EUSFLAT - LFA 2011)*, p. 868–875, Aix-les-Bains, France. Atlantis Press.

Tehrani, A. F. (2021). Heuristics-based learning approach for choquistic regression models. *Pattern Recognition Letters*, 149:137–142.

Yeh, I. C., Yang, K. J., e Ting, T. M. (2009). Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems with Applications*, 36:5866–5871.

DOI: 10.59254/sbpo-2023-175136
http://live-proceedings.pantheonsite.io/p/175136