Explicabilidade em Aprendizado de Máquina utilizando o algoritmo NICE (Nearest Instance Counterfactual Explanations)

Arthur Veloso Kuahara [1] **Daniel Souza de Campos** [1] **Renato Polanczyk Resende** [1]

Abstract

Explicabilidade em inteligência artificial consiste em utilizar métodos contrafactuais para que um algoritmo possa tentar explicar o processo de decisão por trás do seu veredito. Nesse documento, iremos utilizar o algoritmo NICE sobre um data set de gêmeos (TWINS) para entender o motivo da sobrevivência ou não de uma criança e quais fatores mais contribuíram para isso.

1. O Problema dos Gêmeos

O data set escolhido para análise no trabalho consiste em dados de nascimentos de gêmeos nos EUA, de 1989 até 1991, chamado de TWINS. Os dados são detalhados, contendo informações sobre qual bebê nasceu primeiro, o mês de nascimento, presença de fatores de risco nos pais, informações gerais da mãe como idade e etnia, indicadores socioeconômicos da região, etc.

Este data set já foi mencionado em outros *papers* (Guo et al., 2018)(Cheng et al., 2020) como sendo um data set padrão recomendado para aprender sobre efeitos causais. Ele é formado por triplas da forma (X, t, y) em que X são as *features*, t é o tratamento e y é a saída(Guo et al., 2018).

Os dados originais podem ser obtidos em https://github.com/AMLab-Amsterdam/CEVAE/blob/master/datasets/TWINS/.

Dessa forma, o objetivo do trabalho é entender melhor quais seriam as características necessárias que um bebê que morreu após o parto deveria ter para que ele não viesse a óbito. Isso poderia auxiliar a equipe médica na tomada de ações que influenciem na predição de um modelo e, então, salvar a vida de um bebê.

*Equal contribution [1]Departmento de Ciência da Computação, Universidade Federal de Minas Gerais, Belo Horizonte, Brasil. Correspondence to: Arthur Veloso Kuahara <arthur.kuahara@gmail.com>, Daniel Souza de Campos <daniel1@dcc.ufmg.br>, Renato Polanczyk Resende <renato.resende13@gmail.com>.

2. O algoritmo NICE

Para realizar a análise de contrafactuais no data set, foi escolhido o algoritmo NICE. Nearest Instance Counterfactual Explanations (NICE) é um algoritmo proposto para gerar explicações contrafactuais para dados tabulares heterogêneos(Brughmans et al., 2021). Seu código está disponível em https://github.com/dbrughmans/nice.

2.1. Explicações Contrafactuais

O termo 'Explicações Contrafactuais' faz referência a um cenário hipotético, nos moldes de "Se um evento A não tivesse acontecido, então um outro evento B também não teria acontecido". Por exemplo, se uma pessoa não tivesse atendido ao telefone, então ela não teria conversado com quem a ligou.

No contexto de machine learning, a ação consistiria na mudança das features utilizadas pelo modelo e o novo resultado final seria o segundo evento. Assim, é possível modificar parâmetros passados ao algoritmo para verificar se as chances de um dos gêmeos sobreviver é maior ou menor do que a chance obtida anteriormente.

2.2. Funcionamento do algoritmo NICE

O algoritmo NICE tenta encontrar o número mínimo de alterações nas *features* de uma entrada x_0 de modo a classificar o resultado final x_c em outra banda com base nos dados de treino disponíveis.

1. O primeiro passo é encontrar o vizinho de categoria diferente x_n mais próximo (*Nearest Unlike Neighbor*) a x_0, ou seja, em que $f(x_0) \neq f(x_n)$, presente nos dados de treino. Além disso, iniciamos o contrafactual objetivo como sendo igual a x_0 ($x_c = x_0$).

2. O segundo passo é identificar uma lista L_f de *features* nos quais x_c e x_n divergem.

3. Em terceiro, serão gerados vetores de *features* híbridos entre x_c atual e x_n chamados de $x_{h,i}$ a cada iteração i. Nessas iterações, para cada *feature* de nome $L_f[j]$ divergente em L_f, formamos um vetor $x_{h,i}$ exatamente igual ao x_c atual mas com o valor da *feature* $x_c[L_f[j]]$

assumindo o valor de $x_n[L_f[j]]$, ou seja, $x_c[L_f[j]] = x_n[L_f[j]]$.

4. Para cada um desses vetores gerados, calcula-se um valor de recompensa R. Salvamos o vetor híbrido com tal maior valor de recompensa: $x_{h,i}$. Esse vetor possui a *feature* $L_f[j]$ como a que foi alterada.

5. Define-se $x_c = x_{h,i}$. Compara-se a predição do modelo $f(x_c)$ com $f(x_0)$. Se forem diferentes, então o contrafactual foi encontrado, portanto, retorna-se x_c. Se não, retira-se $L_f[j]$ de L_f e recomeça o processo a partir do passo 3.

Como outro exemplo na vida real, podemos considerar a negação de um empréstimo, por exemplo. Qual seria a menor mudança possível no perfil do solicitante para que o empréstimo fosse aprovado de acordo com dados já presentes na base de treino?

3. Outros trabalhos na área

Trabalhos nessa área estão presentes em todos os países devido principalmente a uma maior demanda na melhora da saúde neonatal. Regiões de maior taxa de mortalidade infantil devido a condições socioeconômicas extremas inclusive usam, muitas vezes, *features* peculiares, como o nível de higiene dos banheiros e o nível de educação da mãe. Artigos nessa linha foram publicados por (Mfateneza et al., 2022), (Brahma & Mukherjee, 2022) e (Batista et al., 2021) por exemplo.

Além dessas *features* convencionais e inusitadas, também é possível utilizar dados em formato de *time-series* com *Deep Learning* para obter bons resultados (Feng et al., 2021). Outros estudos que utilizam de *Machine Learning* para predição de mortalidade neonatal são (Sheikhtaheri et al., 2021) e (Mangold et al., 2021).

Além disso, o estudo de gêmeos em si é muito interessante em termos de condições genéticas herdadas pelos pais. Através deles, podemos ter *insights* mais aprofundados em respeito às relações causais de certas *features*, tendo uma relação genética bem definida entre os recém-nascidos.

References

Batista, A. F. M., Diniz, C. S. G., Bonilha, E. A., Kawachi, I., and Chiavegatto Filho, A. D. P. Neonatal mortality prediction with routinely collected data: a machine learning approach. *BMC Pediatr.*, 21(1):322, July 2021.

Brahma, D. and Mukherjee, D. Using machine learning to target neonatal and infant mortality. https://www.ideasforindia.in/topics/macroeconomics/using-machine-learning-to-target-neonatal-and-infant-mortality.html, 2022. Acessado: 26-09-2022.

Brughmans, D., Leyman, P., and Martens, D. NICE: An algorithm for nearest instance counterfactual explanations. 2021.

Cheng, L., Guo, R., Moraffah, R., Candan, K. S., Raglin, A., and Liu, H. A practical data repository for causal learning with big data. In *Benchmarking, Measuring, and Optimizing*, Lecture notes in computer science, pp. 234–248. Springer International Publishing, Cham, 2020.

Feng, J., Lee, J., Vesoulis, Z. A., and Li, F. Predicting mortality risk for preterm infants using deep learning models with time-series vital sign data. *NPJ Digit. Med.*, 4(1):108, July 2021.

Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. A survey of learning causality with data: Problems and methods. 2018.

Mangold, C., Zoretic, S., Thallapureddy, K., Moreira, A., Chorath, K., and Moreira, A. Machine learning models for predicting neonatal mortality: A systematic review. *Neonatology*, 118(4):394–405, July 2021.

Mfateneza, E., Rutayisire, P. C., Biracyaza, E., Musafiri, S., and Mpabuka, W. G. Application of machine learning methods for predicting infant mortality in rwanda: analysis of rwanda demographic health survey 2014-15 dataset. *BMC Pregnancy Childbirth*, 22(1):388, May 2022.

Sheikhtaheri, A., Zarkesh, M. R., Moradi, R., and Kermani, F. Prediction of neonatal deaths in NICUs: development and validation of machine learning models. *BMC Med. Inform. Decis. Mak.*, 21(1):131, April 2021.