Diretrizes para Melhorar a Imparcialidade de Classificadores

Diego Minatel[1], **Nícolas Roque dos Santos**[1], **Alneu de Andrade Lopes**[1]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo

diegominatel@gmail.com.br, nrsantos@usp.br, alneu@icmc.usp.br

Abstract. *One essential aspect of machine learning solutions is ensuring that their decisions do not perpetuate discrimination, whether by favoring or disadvantaging any sociodemographic group. In this context, this paper provides guidelines to mitigate such undesirable situations in classification tasks and promotes machine learning that prioritizes fairness in its outcomes.*

Resumo. *Um dos aspectos essenciais em soluções baseadas em aprendizado de máquina é garantir que suas decisões não perpetuem discriminação, seja favorecendo ou desfavorecendo qualquer grupo sociodemográfico. Nesse contexto, este trabalho fornece diretrizes para mitigar tais situações indesejadas em tarefas de classificação, com o objetivo de promover um aprendizado de máquina que priorize a imparcialidade em seus resultados.*

1. Introdução

Segundo [Commission et al. 2019], o desenvolvimento de soluções baseadas em inteligência artificial deve respeitar todas as leis e regulamentos aplicáveis, princípios e valores éticos, e considerar o ambiente social em que estão inseridas. Tais sistemas devem também evitar a propagação de preconceitos e apoiar a diversidade. Essas orientações estão alinhadas com a legislação brasileira, que, no inciso IV do artigo III de sua Constituição, estabelece: "promover o bem de todos, sem preconceitos de origem, raça, sexo, cor, idade e quaisquer outras formas de discriminação" [BRASIL 1988].

Portanto, sistemas de tomada de decisão automáticos que deliberam sobre cidadãos brasileiros devem ser projetados considerando essas questões para evitar o chamado impacto adverso, que ocorre quando os resultados favorecem ou prejudicam grupos sociodemográficos específicos. De acordo com [Barocas et al. 2023], muitas das decisões sensíveis envolvendo pessoas, nas quais esses sistemas são empregados, ocorrem em tarefas de classificação. Nesse contexto, diversas ferramentas foram desenvolvidas para identificar o impacto adverso e mitigar esse efeito nos classificadores, de forma a torná-los mais justos [Mehrabi et al. 2021].

Este trabalho mapeia boas práticas desenvolvidas pela literatura especializada e apresenta diretrizes para aprimorar a imparcialidade das predições feitas por classificadores. O objetivo é ressaltar a importância de incorporar conceitos de justiça no processo de aprendizado e promover soluções mais justas, que não perpetuem nenhuma forma de discriminação na sociedade brasileira, por meio de um guia acessível para desenvolvedores de soluções baseadas em aprendizado de máquina.

2. Diretrizes para indução de classificadores mais justos

Esta seção apresenta um conjunto de diretrizes, dividido em quatro tópicos, que auxiliam na indução de classificadores mais justos.

2.1. Coleta de dados

Ao contrário de outros tipos de dados, os dados sociais frequentemente contêm subjetividades. Esses dados funcionam como um espelho social, refletindo diversas formas de discriminação presentes na sociedade. Como o aprendizado de máquina é orientado por dados, classificadores treinados com dados discriminatórios tendem a reproduzir esse comportamento indesejado [Barocas et al. 2023].

Uma dessas formas é o que se denomina exemplos contaminados, nos quais a discriminação se manifesta na atribuição dos rótulos dos exemplos. Por exemplo, em um processo de seleção de candidatos realizado por recrutadores que julgam candidatos com base em gênero ou cor da pele, os dados ficam contaminados pelos preconceitos desses recrutadores. Outra forma é o viés populacional, que, devido às condições histórico-sociais, torna um grupo populacional menos representativo em determinadas situações [Mehrabi et al. 2021].

Portanto, a coleta de dados para aplicações que envolvem pessoas deve ser realizada com atenção a essas questões, para evitar que a discriminação esteja refletida diretamente nos dados e para garantir que a amostra seja suficientemente diversificada, representando de forma significativa todos os grupos populacionais. Outras questões devem ser consideradas, como qual parte do passado é representativa para a tarefa em questão e quais atributos abstraem essa tarefa de maneira eficaz para todos os grupos analisados.

2.2. Análise de justiça para grupos

A análise de justiça para grupos avalia a paridade de resultados entre diferentes grupos sociodemográficos, verificando se não há disparidades que favoreçam um grupo em detrimento de outro. Para orientar essa análise, foram adotados alguns conceitos de justiça, sendo dois dos principais a paridade demográfica e a igualdade de oportunidades. A paridade demográfica exige que a taxa de predições positivas seja a mesma para todos os grupos, o que é particularmente relevante em tarefas de seleção de candidatos. Por sua vez, a igualdade de oportunidades é alcançada quando há igualdade nas taxas de revocação entre os grupos [Barocas et al. 2023, Mehrabi et al. 2021].

Como a paridade de diversos desses conceitos é inviável, utiliza-se geralmente uma versão relaxada deles. Nessa versão, calcula-se a diferença ou a razão entre os resultados dos grupos e avalia-se o quão distante um classificador está do valor ideal para cada conceito. Além disso, é recomendável realizar uma análise semelhante para a métrica de desempenho preditivo adotada, como a acurácia, para identificar possíveis assimetrias em que o classificador apresenta precisão significativamente maior ao rotular exemplos pertencentes a um grupo específico.

O viés na avaliação desempenha um papel fundamental na propagação de efeitos discriminatórios pelos classificadores. Esse viés ocorre quando a avaliação do modelo não inclui a análise de justiça ou utiliza métricas inadequadas para o problema em questão. Para mitigar esse viés, é essencial realizar uma análise detalhada, considerando diferentes conceitos e granularidades de grupos em um conjunto de teste que seja verdadeiramente representativo [Suresh e Guttag 2021]. Classificadores que atingem valores próximos ao ideal em diversos conceitos de justiça tendem a ser mais imparciais; no entanto, é igualmente importante que o classificador apresente um bom desempenho preditivo. Como destacado por [Green e Hu 2018], cada análise de justiça deve ser adaptada

ao propósito específico da aplicação, uma vez que cada contexto possui objetivos e características únicas, impactando a vida das pessoas de diversas formas.

2.3. Incorporando conceitos de justiça no processo de aprendizado

A análise de justiça para grupos é uma ferramenta importante para identificar classificadores com comportamento indesejado. No entanto, para efetivamente melhorar a imparcialidade dos classificadores, diversos algoritmos foram propostos na literatura. Por exemplo, métodos de pré-processamento visam minimizar o viés inerente ao conjunto de treinamento, conforme discutido na Seção 2.1. Essa categoria de métodos apresenta abordagens diretas para mitigar problemas relacionados a exemplos contaminados e ao desbalanceamento de dados [Barocas et al. 2023].

Outra categoria inclui aqueles que incorporam conceitos de justiça no processo de aprendizado de máquina. Muitos desses métodos atuam na etapa de indução do modelo, modificando a função objetivo ou impondo restrições na otimização de algoritmos de classificação tradicionais, com o objetivo de auxiliar na convergência de uma parametrização que maximize a imparcialidade das predições. Outros métodos, por sua vez, operam no pós-processamento, visando ajustar as predições do classificador para aprimorar o senso de justiça do modelo treinado [Mehrabi et al. 2021].

Na busca por gerar classificadores mais justos, é possível combinar duas ou mais das técnicas mencionadas. Em algumas situações, devido à natureza do problema e do conjunto de dados, algoritmos de classificação tradicionais, como *Random Forest* e *Support Vector Machines*, podem fornecer soluções mais equitativas do que aqueles projetados especificamente para esse fim. Portanto, é altamente recomendável testar diversos algoritmos com uma grande variedade de configurações.

2.4. Seleção de modelo

A avaliação da imparcialidade em classificação pode abranger diversos conceitos de justiça, além das tradicionais métricas de desempenho preditivo. Dessa forma, é crucial selecionar o classificador que melhor equilibre desempenho e justiça entre os disponíveis. Uma abordagem comum é escolher o classificador com a maior pontuação na métrica de justiça empregada. Contudo, essa estratégia pode resultar na seleção de um modelo com baixo desempenho preditivo, o que não é ideal para a automação de tarefas.

Uma solução para essa questão é aplicar o conceito de multiplicidade, no qual um grupo de classificadores sub-ótimos com desempenhos semelhantes é inicialmente selecionado. Dentre esses modelos, escolhe-se aquele que apresenta o melhor desempenho na métrica de justiça adotada [Black et al. 2022]. Outra abordagem é utilizar medidas multicritério, como a *Multi-Criteria Performance Measure* (MCPM) [Parmezan et al. 2017], que integra a pontuação de várias métricas em uma única avaliação, permitindo a escolha do classificador que otimiza todos os critérios combinados. Adicionalmente, essas abordagens podem ser usadas em conjunto para aprimorar a seleção do modelo.

Outra prática que aprimora a seleção de modelo é a estratificação de dados por grupo e classe na divisão do conjunto de dados em treinamento, validação e teste. Essa estratificação garante que as métricas sejam calculadas em subconjuntos que refletem a distribuição real dos exemplos, o que contribui para a redução do viés de avaliação [Minatel et al. 2023a].

3. Discussão

As diretrizes apresentadas neste trabalho partem do pressuposto de que o viés discriminatório no processo de aprendizado de máquina é adicionado de forma inconsciente, sem a intenção dos desenvolvedores. Mesmo sem essa intenção, nos últimos anos foram noticiados diversos casos em que o aprendizado de máquina contribuiu para a propagação de efeitos discriminatórios na sociedade [Minatel et al. 2023b]. Isso reforça a importância de aplicar essas diretrizes no processo de treinamento de classificadores, principalmente a análise de justiça, na busca por soluções imparciais.

Para concluir esta análise, é importante que essa discussão transcenda o campo científico e tenha efeitos legislativos. Leis ou diretrizes devem estabelecer de forma objetiva os requisitos mínimos que os modelos de aprendizado de máquina devem atender ao apoiar a tomada de decisões envolvendo cidadãos brasileiros. Além disso, é essencial considerar outros aspectos desejáveis para esses sistemas, como a privacidade dos dados e a explicabilidade. Por fim, como salientado por [Green e Hu 2018], o aprendizado de máquina deve se alinhar aos desafios sociais para ajudar a construir um mundo mais justo.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Black, E., Raghavan, M., and Barocas, S. (2022). Model multiplicity: Opportunities, concerns, and solutions. In *ACM FAccT 2022*, pages 850–863.

BRASIL (1988). *Constituição da República Federativa do Brasil*. Brasília.

Commission, E., Directorate-General for Communications Networks, C., and Technology (2019). *Ethics guidelines for trustworthy AI*. Publications Office.

Green, B. and Hu, L. (2018). The myth in the methodology: Towards a recontextualization of fairness in machine learning.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Minatel, D., da Silva, A. C. M., dos Santos, N. R., Curi, M., Marcacini, R. M., and de Andrade Lopes, A. (2023a). Data stratification analysis on the propagation of discriminatory effects in binary classification. In *KDMiLe 2023*, pages 73–80. SBC.

Minatel, D., dos Santos, N. R., da Silva, A. C. M., Cúri, M., Marcacini, R. M., and Lopes, A. d. A. (2023b). Unfairness in machine learning for web systems applications. In *29th Brazilian Symposium on Multimedia and the Web*, pages 144–153.

Parmezan, A. R. S., Lee, H. D., and Wu, F. C. (2017). Metalearning for choosing feature selection algorithms in data mining: Proposal of a new framework. *Expert Syst. Appl.*, 75:1–24.

Suresh, H. and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *EAAMO'21*, pages 1–9.