

Vieses em sistemas de machine learning

Lucas F. Zampar¹

¹Departamento de Ciências Exatas e Tecnológicas – Universidade Federal do Amapá
(UNIFAP)
68903-329– Macapá – AP – Brasil

lucas.26.zampar@gmail.com

Abstract. *The present work aims to present the problem of bias in the development of machine learning systems.*

Resumo. *O presente trabalho visa apresentar a problemática de vieses no desenvolvimento de sistemas de machine learning.*

1. Introdução

Os algoritmos de ML (machine learning) têm ganhado destaque ultimamente pela aplicação em diferentes áreas, bem como a crescente presença nas atividades cotidianas. Por exemplo, algoritmos do tipo podem ser encontrados nos sistemas de recomendações personalizadas das plataformas de *streaming*, de reconhecimento da fala nos *smartphones*, de reconhecimento facial nas redes sociais, etc.

No entanto, apesar dos numerosos benefícios, ainda há muitas preocupações referentes a tais tecnologias. Dentre elas, é possível citar a presença de viés, termo empregado por diversas áreas, porém com sentidos diferentes. Nesse contexto, Suresh e Guttag (2021) empregaram a palavra relacionando-a como fonte de dano potencial durante o ciclo de vida de um sistema de ML, desenvolvendo uma framework que identifica sete vieses.

2. Vieses comuns em Machine Learning

Neste trabalho, serão explorados quatro deles (histórico, de representação, de medida e de avaliação) de modo a apresentar a problemática do campo de forma geral. Além disso, para cada um deles, será apresentado ao menos um trabalho relacionado.

2.1 Viés histórico

O viés histórico ocorre na própria geração dos dados, uma vez que a sociedade que os produz é enviesada. Assim, mesmo modelos treinados com dados perfeitamente mensurados e amostrados podem causar danos a um determinado grupo, como o reforçamento de estereótipos.

Nesse contexto, há ampla documentação expondo o enviesamento de gênero por parte das aplicações comerciais de tradução automática, dado que conferem pronomes a profissões e atividades de acordo com atribuição histórica em muitos casos. Por exemplo, Fitria (2021) demonstrou esse comportamento no Google Translate ao traduzir

expressões neutras em relação ao gênero do indonésio para o inglês, obtendo como resultado a associação de profissões como doutor, engenheiro e presidente a homens, já as mulheres foram associadas a enfermeiras, professoras e cuidadoras de casa.

Indonesian	English
dia seorang dokter	he is a doctor
dia seorang perawat	she's a nurse
dia seorang insinyur	he's an engineer
dia seorang guru	she is a teacher
dia seorang pengacara	he's a lawyer
dia seorang penari	she's a dancer
dia seorang president	he's a president
dia seorang pembantu rumah tangga	she's a housekeeper
dia sedang membaca	he is reading
dia sedang mencuci	she was washing
dia sedang belajar	he is studying
dia sedang bersih-bersih	she is cleaning
dia sedang makan	he's eating
dia sedang masak	she is cooking
dia tidak menikah	he is not married
dia belum menikah	she is not married yet
dia berteriak	he is screaming
dia menangis	she cried

Figura 1 - Exemplo de enviesamento de gênero nas traduções do Google Translate.

Fonte: Fitria (2021)

2.2 Viés de representação

O viés de representação surge quando há sub-representação em parte dos dados empregados durante o treinamento. Dessa forma, a capacidade de generalização do modelo pode ser comprometida para esse conjunto.

Nesse sentido, a falta de representatividade em datasets pode ser notada no estudo de Shankar et al. (2017), que avaliaram o geodiversidade de dois datasets públicos de imagens amplamente empregados: OpenImage e ImageNet. Das imagens analisadas no primeiro, quase 60% eram situadas nos seis países mais representativos da América do Norte e Europa. Já para o segundo, 45% delas eram localizadas nos Estados Unidos.

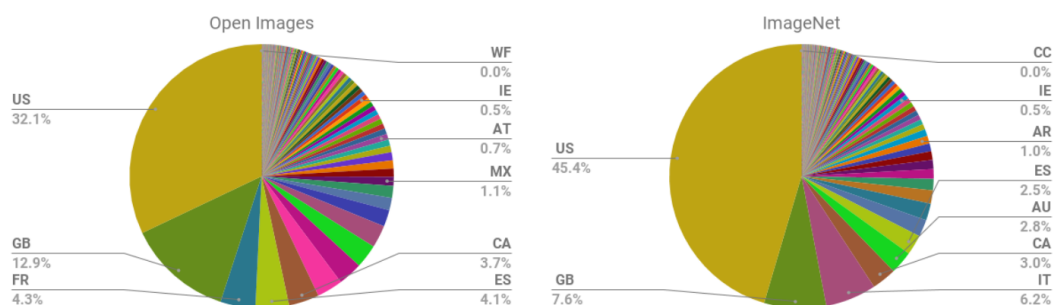


Figura 2 - Geodiversidade em datasets públicos de imagens.

Fonte: Shankar et al. (2017)

As consequências do viés de representação podem ser observadas no trabalho de DeVries et al. (2019), que analisaram a acurácia de seis grandes sistemas de reconhecimento de objetos em relação a um dataset geograficamente diverso. Dessa forma, eles demonstraram que a acurácia média dos sistemas cai conforme a renda mensal média dos países das imagens, indicando a falta de representatividade para países emergentes.

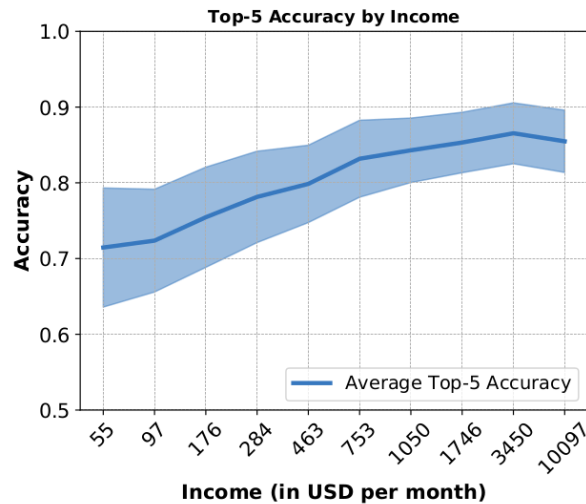


Figura 3 - Relação da acurácia média dos sistemas analisados com a renda mensal média dos países.

Fonte: DeVries et al. (2019)

2.3 Viés de medida

O viés de medida está presente quando as medidas, como características e rótulos, empregadas em um problema de predição refletem pobremente o que se deseja medir ou são geradas de forma diferente entre grupos.

Nesse contexto, no sistema judiciário, modelos que avaliam réus geralmente incluem número de prisões como medida do risco oferecido. Porém, tais medidas podem ser geradas desproporcionalmente entre grupos, dado que determinadas comunidades são mais policiadas do que outras. Por exemplo, Angwin et al. (2016) analisaram a acurácia do sistema COMPAS, utilizado por diversas cortes norte-americanas para avaliar o risco de reincidência dos réus. A partir disso, concluíram que a taxa de falsos positivos era significativamente maior para réus negros (44.9%) do que brancos (23.5%).

Prediction Fails Differently for Black Defendants		
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Figura 4 -A taxa de falsos positivos é maior para réus negros.

Fonte: Angwin et al. (2016)

2.4 Viés de avaliação

O viés de avaliação acontece quando o dataset de teste ou de benchmark sub-representa parte dos dados aos quais o modelo pode ser exposto futuramente. Dessa forma, o modelo pode aparentar um bom desempenho, porém para apenas alguns conjuntos de dados.

Nesse sentido, Buolamwini e Gebru (2018) analisaram a representatividade de mulheres negras em dois datasets de análise facial amplamente empregados para benchmark: Adience e IJB-A. No primeiro, constataram que elas constituíam apenas 7.4%; no segundo, 4.4%. Além disso, no mesmo estudo, verificaram a acurácia de três sistemas comerciais de reconhecimento de gênero das empresas Microsoft, IBM e Face++. Para todos, as maiores taxas de erros se concentraram em mulheres negras, enquanto as menores em homens brancos.

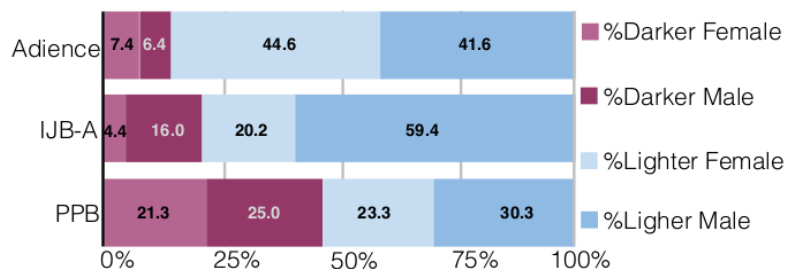




Figura 5 -Falta de representatividade de mulheres negras em datasets de benchmark.

Fonte: Buolamwini e Gebru (2018)

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0%	79.2%	100%	98.3%	20.8%
 FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
 IBM	88.0%	65.3%	99.7%	92.9%	34.4%



© MIT Media Lab

Figura 6 - Os desempenhos de três sistemas de reconhecimento de gênero são os piores para faces de mulheres negras.

Fonte: Buolamwini e Gebru (2018)

3. Conclusão

Uma vez que os sistemas de machine learning estão cada vez mais presentes nas tomadas de decisões que afetam a vida das pessoas em diferentes níveis, o debate sobre a presença de vieses neles é extramente importante. Nesse contexto, o desenvolvimento de modelos deve ser norteado a partir de uma perspectiva ética, analisando como eles podem afetar a vida de diferentes pessoas. Dessa forma, os riscos de danos a determinadas parcelas da sociedade, especialmente minorias, pode ser reduzido, promovendo assim sistemas justos.

Referências

SURESH, Harini; GUTTAG, John. A framework for understanding sources of harm throughout the machine learning life cycle. In: Equity and Access in Algorithms, Mechanisms, and Optimization. 2021. p. 1-9.

FITRIA, Tira Nur. Gender Bias in Translation Using Google Translate: Problems and Solution. Language Circle: Journal of Language and Literature, v. 15, n. 2, 2021.

SHANKAR, Shreya et al. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. arXiv preprint arXiv:1711.08536, 2017.

DE VRIES, Terrance et al. Does object recognition work for everyone?. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019. p. 52-59.

ANGWIN, Julia et al. Machine bias. There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. 23 maio 2016. Disponível em: <<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>>. Acesso em: 08 fev. 2022.

BUOLAMWINI, Joy; GEBRU, Timnit. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. PMLR, 2018. p. 77-91.