Ética e Aprendizado de Máquina: Análise de um Mapeamento Conceitual

Palavras-Chave: Inteligência Artificial, Ética, Aprendizado de Máquina

Autores/as:
Murilo Gonçalves – IC/UNICAMP
Prof. Romis Attux – DCA/FEEC/UNICAMP
Coautora:
Marina de Menezes Lima – FEEC/UNICAMP

INTRODUÇÃO:

Ao longo da última década, observa-se um aumento na confiança em algoritmos capazes de tomar decisões. Bilhões de pessoas utilizam ferramentas que empregam inteligência artificial (IA) em seu cotidiano. Tais ferramentas sugerem palavras, respondem clientes de e-commerce e até mesmo decidem se um cidadão está apto a receber um empréstimo. Essa confiança pode ser explicada, parcialmente, pelos notáveis avanços no campo de aprendizado de máquina (ML, do inglês *machine learning*), que têm levado a resultados impressionantes em classificação de imagens, sons, dentre outros dados. Entretanto, quando esses modelos de ML tomam decisões que envolvem dilemas éticos e têm consequências na vida das pessoas, é fundamental lidar com as implicações sociais e de justiça. Por essa razão, têm surgido estudos na área de Equidade ou *Fairness*, concomitantemente ao rápido desenvolvimento técnico da inteligência artificial. Neste projeto, partimos do referencial de [1], no qual os autores propõem um mapeamento conceitual dos problemas éticos comuns encontrados quando cabe a algoritmos tomar decisões de grande impacto para setores da sociedade.

DESENVOLVIMENTO:

Iniciamos nosso projeto com estudos nas áreas de probabilidade e teoria da informação. Em sequência, estudamos a área de IA, mais especificamente de uma perspectiva de ML, tendo por

base o material [2]. Passamos por todos os modelos clássicos como regressão linear, regressão logística, redes neurais, árvores binárias e florestas aleatórias. Além disso, aprofundamo-nos nos conceitos de erro quadrático médio, gradiente descendente, *overfitting, underfitting*, e ainda diferentes tipos de erros e procedimentos de regularização. Durante essa fase, aplicamos nossos novos conhecimentos em alguns *datasets* livres do repositório UCI [3], em particular o *Abalone Dataset, Wine Quality Dataset e Concrete Compressive Strength Dataset*.

Após a análise de todos os tópicos técnicos pertinentes, passamos a ler e discutir artigos e textos relacionados aos fundamentos da ética e sociologia da técnica, em particular [4] e [5]. Esse estudo nos forneceu uma visão interessante sobre como as tecnologias podem afetar aspectos sociais, políticos e econômicos com ou sem intenção explícita da parte de quem as aplica.

Munidos de todos os conceitos necessários pertencentes às duas grandes áreas que permeiam nosso projeto, finalmente iniciamos as discussões de *fairness,* que trata dos vieses que estão naturalmente presentes modelos de ML, quando não são construídos ou tratados com o devido cuidado. Essa discussão engloba não só os aspectos técnicos – como as diferenças entre correlação e causalidade – mas também discussões de cunho profundamente ético, como a própria definição de “justiça". Discutimos casos da vida real, com o auxílio dos exemplos presentes em [6], que trazem exemplos de como a falta de preocupação com os vieses de algoritmos prejudica grupos sociais, principalmente aqueles que já são desfavorecidos. Com essa introdução ao assunto, estudamos, com uma visão técnica, o artigo de Mitteldstadt et al. [1]. Seu mapeamento busca nomear universalmente os problemas éticos mais comuns encontrados quando algoritmos tomam decisões. São eles:

1) Evidência Inconclusiva: métodos de aprendizado de máquina trabalham com um grau de incerteza e as correlações por eles exploradas, muitas vezes, não permitem inferir dependências causais.

2) Evidência Inescrutável: os algoritmos são, muitas vezes, opacos, o que não permite que se compreenda como eles utilizam a informação disponível.

3) Evidência Desencaminhada: os dados são um limite inexorável para a qualidade da máquina – problemas com os dados se refletem em problemas de desempenho.

4) Resultados Injustos: algoritmos podem desrespeitar o que a sociedade considera “justo” (fair), prejudicando membros de certos setores da sociedade.

5) Efeitos Transformativos: a operação de algoritmos pode afetar a maneira pela qual as pessoas categorizam os dados da realidade.

6) Rastreabilidade: algoritmos herdam desafios éticos de técnicas e bases de dados, o que faz com que comportamentos questionáveis sejam difíceis de avaliar.

Além desse mapeamento ético, existem, como forma de quantificar alguns tipos de vieses, diversas métricas estatísticas, como Paridade Demográfica, Probabilidades Igualadas e Oportunidades Iguais [8]. Essas se baseiam na ideia de que um preditor não pode alterar seu resultado com base em qual grupo, protegido ou não, ou quais características sensíveis um indivíduo apresenta. Tais métricas são importantes, apesar de limitadas, como forma de modelar matematicamente algumas iniquidades comuns.

Com esse entendimento de conceitos relevantes em *fairness*, começamos a analisar as formas de mitigar os vieses de preditores baseados em ML. As principais intervenções utilizadas para tal são divididas em três grupos, a depender do momento em que são aplicadas: no pré-processamento, durante o treino ou no pós-processamento. Em geral, técnicas de pré-processamento são utilizadas em conjunto com outras, já que, apesar de apresentarem pouca melhora na diminuição dos vieses, não possuem muitos efeitos negativos na acurácia do resultado. Técnicas de pós-processamento costumam ser simples de serem implementadas, já que não dependem dos detalhes de implementação do algoritmo, mas possuem uma performance média inferior às técnicas durante o treino. Não obstante, é impossível criar um método generalista, já que as intervenções específicas que otimizam a equidade de um preditor dependem da natureza do problema e do que se é considera “equidade”, portanto cada caso deve ser analisado para a escolha da técnica pertinente.

Para fixar tudo o que foi visto até aqui, construímos um modelo de *Random* Forest e aplicamos intervenções no A*dult dataset*, que possui como objetivo utilizar características de um certo indivíduo para fornecer uma previsão da renda anual como maior ou menor do que cinquenta mil dólares. Analisamos os atributos sensíveis do *dataset* como “sexo”. Percebemos que existe grande viés, já o modelo tende a classificar as mulheres como como pertencentes do grupo de menor renda, enquanto que o oposto ocorre com os homens. Considerando o contexto social, esse viés é esperado, apesar de prejudicial.

Ainda trabalhando nesse *dataset*, utilizamos a biblioteca Fairlearn para analisar as métricas de Paridade Demográfica e Probabilidades Igualadas, que evidenciaram a presença de vieses não só no atributo “sexo”, mas também na etnia e país de origem dos dos indivíduos. Após estudar os algoritmos de mitigação de vieses mais conhecidos e analisar nosso caso, optamos por aplicar o algoritmo de Reponderação de dados [8], que é um método de pré-processamento que busca melhorar a paridade demográfica do conjunto de dados.

Por fim, associamos o mapeamento de [1] com os conceitos de ML e *fairness* estudados, levando em consideração a análise do *Adult dataset* e focando nos problemas de Resultados Injustos,

Evidências Desencaminhadas, Efeitos Transformativos e Rastreabilidade, que intersectam com as técnicas empregadas no nosso estudo de caso.

CONCLUSÃO:

Concluímos que é possível organizar de forma técnica e satisfatória grande parte dos problemas que existem atualmente nos modelos que utilizam ML para predição, o que significa que é possível analisá-los a fundo e buscar medidas de correção. Além disso, intervenção com métodos de mitigação de viés que existem atualmente se mostraram bastante eficazes em ajudar a tornar as predições mais justas, com a desvantagem, menos relevante de um ponto de vista ético, de diminuir sua acurácia. Sobretudo, é importante frisar que a questão da ética em ciências de dados é de extrema importância e não pode ser negligenciada por governos e empresas que empregam preditores baseados em ML. A discussão acerca de como esses modelos devem ser tratados e como serão inseridos na sociedade deve envolver pesquisadores não apenas técnicos, mas também cientistas sociais, filósofos e especialistas em humanidades, sendo a única forma de assegurar que essas novas tecnologias não trarão prejuízos para os diferentes indivíduos e grupos minoritários que compõem a população mundial.

BIBLIOGRAFIA

[1] B. D. MITTELSTADT, P. ALLO, S. WACHTER M. TADDEO, L. FLORIDI. **The Ethics of Algorithms: Mapping the Debate.** Big Data & Society, 3(2):1–21, 2016.

[2] L. BOCCATO, R. ATTUX, Notas de aula de IA006. https://www.dca.fee.unicamp.br/~lboccato/ia0062s2019.html.

[3] UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/index.php.

[4] L. WINNER. **Modern Technology: Problem or Opportunity?** Daedalus, 109(1):121–136,1980.

[5] S. ZUBOFF. **Big Other: Surveillance Capitalism and the Prospects of an Information Civilization.** Journal of Information Technology, 30:75–89, 2015.

[6] S. BAROCAS, M. HARDT, and A. NARAYANAN. **Fairness and Machine Learning**. fairml-book.org, 2019. http://www.fairmlbook.org.

[7] S. VERMA, J. RUBIN. **Fairness Definitions Explained**. 2018 ACM/IEEE International Workshop on Software Fairness. Pages 1–7, 05 2018.

[8] F. KAMIRAN, T. CALDERS. **Data Pre-Processing Techniques for Classification without Discrimination**. Knowlegde and Information Systems, 33, 10, 2011.