Ética e Aprendizado de Máquina:

Implementação de um Algoritmo

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

Autores/as:

Marina de Menezes Lima – FEEC/UNICAMP

Prof. Romis Attux – DCA/FEEC/UNICAMP

Coautor:

Murilo Gonçalves – DCA/FEEC/UNICAMP

Introdução

Motivada pela obtenção de excelentes resultados em diversas áreas, a tomada de decisões por parte das máquinas é um fenômeno que se mostra cada vez mais intrínseco à sociedade e à tecnologia moderna. Por estar presente em diversos contextos, o aprendizado de máquina acaba tratando, direta ou indiretamente, de questões com implicações sociais, sendo extremamente importante que essas sejam trabalhadas de forma ética e justa. A garantia dessa justiça, porém, não é algo trivial a se atingir e envolve não somente questões técnicas de limitação das máquinas, mas também discussões sobre o conceito de justiça e igualdade, o que acaba por atrair a atenção de diversos profissionais. Tendo por base essa questão, portanto, este projeto visa trabalhar as questões de ética em aprendizado de máquina, se baseando no mapa conceitual proposto em [1], e desenvolvendo um algoritmo que visa mitigar as questões apresentadas por ele.

Desenvolvimento

O projeto foi iniciado com estudos dirigidos que começaram nas áreas de probabilidade e teoria da informação, seguindo para o estudo de conceitos sobre inteligência artificial. Na segunda parte, estudamos todos os modelos clássicos como regressão linear, regressão logística, redes neurais, árvores binárias e florestas aleatórias, assim como diversos tipos de erros e dificuldades que são encontrados nessa área, como o erro quadrático médio, gradiente descendente, *overfitting e underfitting* [2]. Para fixar os conceitos aprendidos trabalhamos com alguns *datasets* livres do repositório UCI [3], aplicando e analisando os processos de aprendizados desenvolvidos.

Para continuarmos nossa trajetória de estudos, passamos então para a discussão social que engloba os fundamentos da ética, trabalhando sobre artigos que discutem as formas como as tecnologias são moldadas por forças sociais e econômicas [3] e, ainda, a discussão acerca da fonte dos dados que são utilizados para o aprendizado de máquina e os interesses que existem por trás deles [4].

Concluídos assim, os estudos das áreas técnicas e sociais separadamente, partimos para as discussões sobre o nosso tema principal, *Fairness.* Este tema diz respeito à presença de vieses no aprendizado de máquina,

sendo ela um fenômeno que ocorre quando um algoritmo produz resultados que são sistematicamente prejudicados devido a suposições errôneas no processo de aprendizado de máquina, e suas implicações na justiça e equidade do algoritmo.

Esse viés, que pode decorrer de amostras que refletem as disparidades e distorções da sociedade e de recursos limitados e pouco informativos para grupos minoritários, apresenta desafios únicos que precisam ser compreendidos para revisar os resultados de maneira adequada e evitar que os dados tendenciosos do aprendizado de máquina afetem inesperadamente os resultados obtidos.

Para analisar o nível de justiça em algoritmos, diversas métricas como a Paridade Demográfica, Probabilidades Igualadas, Oportunidades Iguais, entre outras foram criadas se baseando em aspectos que englobam a definição da justiça e suas aplicações nos contextos sociais. Vale ressaltar que satisfazer todas as métricas simultaneamente é uma tarefa impossível, já que a imposição simultânea delas causa muitas restrições. Sendo assim, é importante avaliar o contexto e a aplicação em que as definições de justiça precisam ser usadas em consideração e usá-las de acordo com os objetivos que visam ser alcançados.

As três principais práticas que prometem diminuir os vieses presentes nos classificadores são o tratamento durante o pré-processamento, durante o treinamento e durante o pós-processamento. Cada uma apresenta diversas vantagens sobre as outras, apesar de possuírem lados negativos também. Assim, vale analisar o problema trabalhado e a métrica principal da análise para concluir qual processo será o mais adequado.

Dando sequência a toda a teoria trabalhada, associamos o mapeamento citado em [1] ao contexto de ML, discutindo os conceitos relevantes da área que corroboram com o autor e algumas das técnicas de *Fairness* que se propõem a ajudar cada um dos problemas levantados. Neste mapeamento, seis aspectos principais são trabalhados, sendo eles:

- 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.

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

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

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

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

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

Após a análise técnica do mapeamento, focamos na implementação de um algoritmo que visasse mitigar pelo menos um dos pontos apresentados. Como nosso estudo tem como tema principal a ética e a justiça no aprendizado de máquina, os Resultados Injustos, as Evidências Desencaminhadas, os Efeitos Transformativos e a Rastreabilidade foram os pontos de maior foco em nossa implementação.

Escolhemos trabalhar com o *Adult dataset* obtido em [5], que possui como objetivo principal o fornecimento de uma previsão da renda anual de um certo indivíduo como maior ou menor do que cinquenta mil dólares, com base em suas características. Começamos tratando o *dataset* e analisando se realmente existiam vieses presentes nos resultados, tomando o atributo ‘sexo’ como o sensível. A análise, que resultou nas matrizes de confusão da Tab. 1, denunciam que a maior quantidade de Falsos Negativos e menor quantidade de Verdadeiros Positivos para mulheres indica que o modelo tende a classificá-las como membros do grupo de

menor renda. Já com mais Falsos Positivos e menos Verdadeiros Negativos, os homens apresentam a tendência de serem classificados com rendas altas, como já era de se esperar quando levamos em conta o contexto histórico da sociedade, confirmando o viés já esperado.

		Real	
		Positivo	Negativo
Previsão	Positivo	0.6321	0.3678
	Negativo	0.0968	0.9032

(a) Característica 'homem'

		Positivo	Negativo
Previsão	Positivo	0.5332	0.4668
	Negativo	0.0229	0.9770

(b) Característica 'mulher'

Tabela 1: Matriz de confusão para as características atribuídas ao atributo 'sexo'

Em seguida, analisamos também as métricas de Paridade Demográfica e Probabilidades Igualadas através da biblioteca Fairnlearn. Verificamos ambas através da diferença e da proporção e para todos os casos os valores obtidos se apresentaram longe do ideal esperado. Assim, mostrou-se necessário aplicar um algoritmo de mitigação dos vieses apresentados, com o objetivo de tornarmos o nosso modelo mais justo.

Após estudar as diversas opções de algoritmos propostos por diversos pesquisadores e ter em mente a mitigação dos aspectos propostos por Mittelsatdt, escolhemos trabalhar com a Reponderação de dados proposta por Kamiran e Calders [6]. Ela se baseia em uma técnica de pré-processamento que realiza uma reponderação dos dados de treinamento, visando garantir uma paridade demográfica ideal. Este algoritmo, portanto, identifica os pontos sensíveis mal representados e eleva seus pesos, de forma com que tenham maior impacto no treinamento do modelo.

Mitigando diretamente o problema de Evidência Desencaminhada, tal processo faz com que, a questão acerca da qualidade dos dados apresentados seja atenuada, refletindo em uma diminuição dos problemas de desempenho e justiça da máquina. Além disso, tal tratamento influenciará também nos Resultados Injustos, já que diminui os vieses das decisões da máquina, e, consequentemente, nos Efeitos Transformativos.

Por fim, após aplicarmos tal algoritmo com o auxílio da biblioteca AI Fairness 360 [7], oferecida pela empresa IBM, foi possível verificar que a técnica utilizada realmente ajudou a tornar Paridade Demográfica mais próxima do ideal, reajustando o alcance de justiça em nosso código. Apesar de não mitigar tal problema em sua totalidade, o algoritmo pode ser combinado com outras intervenções para atingir a paridade demográfica total, já que se trata de uma etapa de pré-processamento.

Conclusão

Concluímos, então, que os problemas relacionados aos vieses presentes nas decisões de aprendizado de máquina podem ser analisados de uma forma técnica que permite tratá-los e diminuí-los. Os métodos de mitigação que são propostos atualmente se mostram 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.

Referências

[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 do Curso IA006*. https://www.dca.fee.unicamp.br/-lboccato/ia0062s2019.html

[3] L. WINNER. "Modern Technology: Problem or Opportunity?" Daedalus, 109(1):121-136, 1980.

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

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

[6] F. KAMIRAN, T. CALDERS. Data pre-processing techniques for classification without discrimination. Knowledge and Information Systems, 33, 10, 2011.

[7] Introducing AI Fairness 360, https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/.