Algoritmos e Discriminação - Resumo

Trabalho da cadeira de Introdução à Ciência dos Dados

Escrito por:

Alexsandro Vítor Serafim de Carvalho - avsc

Jeffson Carneiro Silva Simões - jcss3

Introdução

Este trabalho resume o conteúdo de 4 artigos, sendo os 2 primeiros jornalísticos e os 2 últimos científicos, que tratam de enviesamento e discriminação na tomada de decisões. Os artigos lidos foram:

- **Can an Algorithm Hire Better Than a Human?** Claire Cain Miller. The New York Times, 2015. https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html
- **When Algorithms Discriminate** Claire Cain Miller. The New York Times, 2015. https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html
- **Economic Models of (Algorithmic) Discrimination.** Bryce W. Goodman NIPS Symposium on Machine Learning and the Law. http://www.mlandthelaw.org/papers/goodman2.pdf
- **A survey on measuring indirect discrimination in machine learning.** https://arxiv.org/pdf/1511.00148.pdf

Resumo

Seres humanos são passíveis de tomar decisões baseadas em preferências e preconceitos, e por isso, nem sempre tomam as decisões corretas ou eficientes. Com o objetivo de aumentar a eficiência dessas escolhas, tem sido proposta a automatização das mesmas através de algoritmos preditivos. Esses algoritmos têm sido aplicados na contratação de funcionários e na avaliação de pedidos de empréstimos.

O uso desses algoritmos é promovido como mais eficiente em relação a tempo e custo. Além disso, devido às possíveis consequências de decisões enviesadas, como a redução de oportunidades de trabalho e o aumento do encarceramento de grupos sociais minoritários, eles também são propostos como soluções para este problema, devido a sua objetividade e ignorância a respeito dos preconceitos humanos.

Porém é possível que essa tecnologia, ao invés de resolver o problema, se torne parte dele. Algoritmos de aprendizagem são desenvolvidos para replicar as avaliações de amostras anteriores, e do ponto de vista de um, qualquer variável presente nas amostras é válida para influenciar seus resultados. Isso torna sua ignorância, mencionada anteriormente, uma faca de dois gumes. Se os rótulos de um conjunto de amostras foram influenciados por critérios de raça ou sexo, por exemplo, uma IA replicando essas classificações também replicará essas influências, sem levar em consideração as implicações sociais de tomar decisões baseando se neles.

Comparadas com as decisões de seres humanos, as decisões tomadas por algoritmos podem discriminar de forma mais sistemática e em maior escala. Desde vagas de empregos de altos salários aparecendo apenas para homens a pesquisas sobre esses cargos subrepresentando mulheres. No exemplo dado no 2º artigo, houveram 11% de mulheres na busca contra 27% na realidade, sendo o 1º uma boneca Barbie. Fazendo um novo teste durante a produção desse texto, a proporção de mulheres se mantém baixa, embora o primeiro resultado feminino seja o 5º, e não seja uma boneca.

Modelos computacionais também podem discriminar devido a falta de informação ou representatividade. Quando um grupo é sub-representado na base de dados, as amostras pertencentes a ele geram resultados mais incertos. Por exemplo: Um banco, com a intenção de evitar prejuízos com empréstimos arriscados, poderia lidar com essa incerteza penalizando a avaliação final de amostras com classificações mais incertas, o que dificultaria o acesso desse grupo minoritário a crédito. Uma outra possibilidade é a de que o algoritmo seja desconsiderado ao lidar

com esses grupos, e o banco voltaria a recorrer a estereótipos e preconceitos para a tomada de decisão.

Portanto, formas de descobrir a discriminação gerada pelos algoritmos e métodos utilizados para prevenir modelos discriminatórios são de extrema importância na criação dos modelos preditivos.

A descoberta da discriminação visa encontrar padrões discriminatórios nos dados usando métodos de mineração de dados. A abordagem de mineração de dados para a descoberta da discriminação tipicamente minera as regras de associação e classificação dos dados e, em seguida, avalia essas regras em termos de discriminação potencial.

Já na prevenção o objetivo é ter um modelo (regras de decisão) que obedeça a restrições de não discriminação, que normalmente estão diretamente relacionadas à medida de discriminação selecionada.

No contexto da aprendizagem automática, a não discriminação pode ser definida através de dois pontos :

- Pessoas que são semelhantes em termos de características não protegidas devem receber previsões semelhantes.
- Diferenças nas previsões entre grupos de pessoas só pode ser tão grandes quanto justificado por características não protegidas.

As características referenciadas como protegidas são pontos que podem levar a discriminação quando utilizados para previsões, por isso os algoritmos devem levar elas em conta ao fazerem suas predições, utilizando diferentes tipos de medidas de discriminação para tratar os dados de maneira justa e imparcial, afim de garantir que diferentes grupos sociais e demográficos possam ser tratados de forma justa e realmente imparcial.