Máquinas Preconceituosas: Machine Learning e Ciências Sociais

O avanço das técnicas de aprendizagem de máquina e ciência de dados tem aflorado com vários aplicativos e algoritmos utilizados para reconhecimento de faces, casamento de padrões para construção de perfis de usuários, entre outras coisas. Os algoritmos estão cada vez mais interpretando comportamentos e dados sobre as pessoas. Porém, por mexer com pessoas, esses algoritmos acabam tendendo a ficar preconceituosas. Esses "bugs" de discriminação tem vários motivos, às vezes nos dados coletados para treinamento, ou por rotulação feita após a coleta. Isso pode levar a vários problemas éticos, de discriminação e acabar tendo repercussões muito negativas.

Esse preconceito e/ou discriminação pode vir de várias maneiras para um dataset, pois os reflexos de um estudo da sociedade são esses. Nossa sociedade ainda é muito preconceituosa. Alguns motivos para que os resultados desses algoritmos terem essa tendência preconceituosa são: rótulos dados por seres humanos aos dados ou então dados de treinamento não suficientes ou tendenciosos.

Uma avaliação mais a fundo desses dados, pode descrever por exemplo que minorias, por ter uma menor quantidade de dados no conjunto, sejam menos relevantes em alguns aspectos, o que já deixa o algoritmo tendencioso e discriminativo. Por isso que para conjuntos de dados que analisam comportamentos de pessoas devem ser estudados em conjunto com um cientista social. Essa análise final dos dados, deve vir dos dois lados, para que não haja situações como essas de discriminação. Esses dados podem vir a causar danos aos resultados da análise do conjunto.

Em uma análise inicial sobre os danos causados por essa discriminação divide em dois principais tipos de danos causados: os danos de alocação e os danos de representação.

Os danos de alocação são as mais comuns associações com discriminação. O dano alocativo é quando um sistema aloca ou não aloca certo grupo para um oportunidade ou recurso. Por exemplo, aceitar ou não uma proposta de empréstimo em um banco para diferentes grupos de pessoas.

Os danos de alocação podem ser divididos em 5 tipos: Estereotipação, Reconhecimento, Difamação, Sub-representação e Exoneração.

A estereotipação acontece quando um sistema assume que um certo grupo sempre terá o resultado que a maioria terá. Como por exemplo: O google translate traduz palavras de línguas que não tem gênero com o gênero mais utilizado.

O reconhecimento acontece quando um grupo é apagado ou simplesmente invisível para um certo sistema. Como por exemplo um bug do Google Photos que colocava pessoas negras não como faces e sim como animais. Ou mesmo o reconhecimento de faces de câmeras Nikon que não reconhecia pessoas de tons de pele mais escuros.

A difamação acontece quando um sistema denigre certos grupos com rótulos, ou ofensas culturais. O exemplo citado anteriormente do Google Photos também é em parte difamação, assim como também acontece em sistema de auto sugestões ao sugerir coisas como "judeus devem".

A sub-representação e a exoneração acontece quando grupos são removidos ou sub-representados em uma área do sistema, o que leva a má interpretação dos dados pelo sistema. Como, por exemplo, em uma pesquisa de imagens por CEOs no google a maioria dos resultados, na primeira página, eram homens brancos, tendo somente uma mulher.

Já os dano de representação são quando um sistema reforça as subordinações de certos grupos fazendo uma alocação má distribuída de acordo com raça, gênero, religião, etc.

Algumas soluções propostas para evitar que os algoritmos tenha essa tendência de tornar-se preconceituoso envolvem análises mais a fundo e uma parceria profunda entre cientista social e cientista de dados.

Como geralmente os algoritmos têm a tendência de levar menos em conta ou até ignorar dados que tem pouca representatividade em um conjunto, um primeiro passo seria analisar mais a fundo esses subconjuntos e tentar casar padrões para essas minorias.

Além disso, deve-se haver um processo de análise dos modelos a serem usados em conjunto com um cientista social, tanto para predição quanto para explanação, e principalmente para análises exploratórias.

Modelos de predição são modelos que tentam a partir da base de dados prever comportamentos, os de explanação tentam responder porque os dados estão daquela maneira e já os exploratórios tentam casar novos padrões para os dados coletados.

Os exploratórios são os mais abstratos e, portanto, devem ter uma atenção para a área social, de maneira a avaliar e tratar os dados mesmo após a execução do algoritmo tratando os danos que podem vir se os dados estiverem enviesados.

Para melhorarmos o processamento e resultado desses dados devemos primeiramente ter concepção que haverá implicitamente um preconceito, já que ele existe na nossa sociedade, e que devemos desafiar os estereótipos de minorias, mesmo que dê mais trabalho. Em segundo lugar, se quisermos usar outros métodos para responsavelmente e justamente analisar esses dados precisamos de mais envolvimento interdisciplinar da comunidade como um todo para o melhor entendimento da ciência.