20 lições sobre preconceito contra sistemas de aprendizado de máquina

Este artigo teve como objetivo apresentar diferentes tendências de sistemas de aprendizagem de máquinas e maneiras de lidar com possíveis problemas que possam surgir no processo de aprendizado. O uso de sistemas com aprendizagem de máquinas está gerando um novo ecossistema repleto de técnicas e infraestrutura, porém o aprendizado absoluto, dependendo de sua plena capacidade, ainda está no começo. Esse novo ecossistema é muito interessante e bastante relevante quando se observa as possibilidades de atividades que podem ser executadas através dessas técnicas. No entanto, existem problemas realmente preocupantes. Se pode citar como exemplo formas de preconceito, estereotipagem e determinação injusta.

O interesse em estudar sobre o viés de aprendizagem de máquinas vem crescendo exponencialmente nos últimos 3 anos. Esse número mais que dobrou no último ano ao se perceber que essa tendência pode ser aplicada em diferentes âmbitos e grupos. Tal interesse tem como objetivo o entendimento de justiça nesse ecossistema, onde tendência pode ser considerado como desvio que produz algum tipo de dano. O conjunto de dados utilizado para o treinamento de tais sistemas nunca são totalmente neutros e nem sempre é possível neutralizá-los, logo, não existe um jeito de resolver todos os danos de polarização nesses meios. Existem várias técnicas para resolver o problema de enviesamento do conjunto de treinamento. Dentre elas estão o *scrubbing to neutral* (esfregando no neutro) e *demographic sampling* (amostragem demográfica), entretanto todos são enviesados pelo "neutro".

Esse viés pode se referir a coisas diferentes a depender do contexto que está inserido, todavia, é possível se ter uma definição genérica que diz: "'Viés' é um desvio que produz um tipo de dano". O viés, normalmente, vem através do conjunto de treinamento. Contudo, se esse conjunto estiver tendencioso, incompleto, ou inclinado, dessa forma o viés acabará sempre penso de acordo com a inclinação. Esses conjuntos de dados têm origem através da rotulagem humana, que é uma outra maneira pelas quais preconceitos e suposições culturais podem se arrastar e resultam na exclusão ou super representação da subpopulação. O viés estrutural é uma questão social e uma questão técnica, e, se não formos capazes de considerá-lo sócio-técnicos, esse tipo de preconceito será um problema para a aprendizagem de máquina.

Há dois tipos de danos que são causados por causa da tendência: os danos de alocação e os danos de representação - o primeiro tem uma visão economicamente orientada, enquanto o outro é mais cultural -. O dano de alocação se dá pela retenção de grupos, é quantificável e trata de questões de justiça - e justiça em transações específicas e discretas - , enquanto o dano de representação se dá pela representação difusa dos seres humanos e da sociedade. Os danos de alocação ainda podem ser classificados em 5 tipos: estereótipos, reconhecimento, difamação, subrepresentação e exnomeação.

Pelo fato de que as questões de polarização podem permanecer e se manifestar de diversas e novas maneiras, devemos entender que a classificação não é apenas uma questão de técnica mas também é uma questão social onde o viés tem consequências reais para as pessoas que estão sendo classificadas.

O viés do sistema é cada vez mais impactante devido ao seu grande uso na atualidade. Algoritmos demonstrando preconceito no sistema podem causar um grande prejuízo na vida humana. Geralmente, quando isso acontece é porque a lista de categoria dos dados é muito limitada ou quando se é utilizado dados pessoais inválidos. Algumas dessas consequências podem parecer hipotéticas e de longo prazo enquanto outras, são imediatas e diretas.

O artigo fala sobre o crescimento do uso de aprendizagem de máquinas e os problemas que podem ocorrer se não forem tomados os devidos cuidados; enfatiza o cuidado que deve ser tomado ao se estabelecer o conjunto treinamento, pois o enviesamento do sistema começa pelos dados que serão usados e, se esses são tendenciosos para a classe A ou para a classe B, isso significa que o seu sistema estará mais propenso a favorecer ou a prejudicar uma determinada classe. Da mesma forma que a aprendizagem de máquina pode ser de grande ajuda em qualquer área, um simples erro no conjunto de treinamento pode causar um efeito catastrófico na vida de alguém.