Is it ethical to avoid error analysis?

O artigo tem por proposta discutir sobre as implicações éticas da não realização de análise de erro do modelo e resultados de algoritmos de aprendizagem de máquina, tendo como foco de estudo as práticas de falsificação dos resultados e discriminação de grupos específicos. Por fim, são levantados métodos de análise de erro para deep learning, dado que ele possui uma estrutura de difícil interpretação.

Os autores definem falsificação como sendo manipulação ou omissão de qualquer dado envolvido na pesquisa (materiais de pesquisa, imagens, equipamentos ou processos) de forma que o resultado final seja moldado como desejado. É argumentado que quando o pesquisador escolhe não realizar um trabalho minucioso, sem intensão de falsificar os resultados os dados, isso pode ser considerado antiético e não profissional, porém não seria enquadrado como uma fraude de pesquisa.

Um sistema não discriminativo é definido como um sistema que possui como saída o mesmo valor independente da presença de alguma característica não protegida (raça, gênero, etc.), e, caso isso não seja atendido, o pesquisador é capaz de justificar as razões por trás disso. Os autores argumentam que o pesquisador possui uma obrigação ética de garantir que seu algoritmo não possui viés discriminatório caso o mesmo for aplicado em um cenário relacionado à humanos.

São apresentados dois motivos dos algoritmos de deep learning serem tão difíceis de se analisar, a primeiro é que as grandes companhias não revelam o funcionamento de seus algoritmos devido às suas políticas de propriedade intelectual, uma solução levantada pelos autores seria o incentivo às publicações de código aberto. O segundo motivo é que os algoritmos não possuem uma lógica intuitiva e a representação de dados intermediários é de difícil compreensão, uma solução seria estudar a saída do algoritmo dada uma entrada bem especifica, de forma a se focar como o modelo se comporta se necessariamente saber quais são os mecanismos por trás.

Por fim, os autores concluem que quando o algoritmo utilizado é disponibilizado de forma open source e há um entendimento profundo de como ele funciona será produzida uma pesquisa de alta qualidade. Dessa forma, a análise de erros leva a 4 benefícios principais: melhor entendimento do algoritmo proposto, transparência, boa prestação de contas e avanço em pesquisas futuras.

A survey on measuring indirect discrimination in machine learning

O artigo tem por objetivo ser um guia completo sobre as diferentes formas de se medir discriminação em algoritmos de aprendizagem de máquina. Inicialmente é dado uma visão geral das questões de discriminação e do funcionamento dos algoritmos e posteriormente são apresentadas as várias métricas e seus cenários de aplicação.

Aprendizagem de máquina não discriminatório é uma área interseção entre ciência da computação, direito e ciência sociais, esta possui duas vertentes de pesquisa, descoberta e prevenção, ambas fazem uso de métricas de discriminação para suas análises. Essas métricas podem ser classificadas em 4 categorias: testes estatísticos, métricas absolutas, métricas condicionais e métricas estruturais.

Os testes estatísticos têm por objetivo verificar se há ou não presença de discriminação no modelo proposto. Podem ser teste de inclinação de regressão, por diferença de teste de médias e diferença em proporções para dois grupos ou vários grupos.

Medidas absolutas, possuem como função identificar a magnitude da diferença no tratamento entre grupos, estes determinados por características protegidas. Os autores citam algumas como: diferença média, diferença normalizada, área sob a curva, taxa de impacto, relação elift e relação de probabilidades.

Métricas condicionais tentam identificar o quanto da diferença entre dois grupos pode ser aplicada por outras características dos indivíduos que não são protegidas. As principais métricas são: diferença inexplicável, medida de propensão e relação de Belift.

Métricas estruturais são utilizadas para se identificar discriminação direta, aquela que segrega intencionalmente devido à origem racial ou étnica. São abordados os testes de situação e medida de consistência.

Para fim de comparação foram aplicadas as métricas à um conjunto de dados sintético. Analisando os resultados de desempenho obtidos os autores não recomendaram o uso das métricas baseadas em relação (taxa de impacto, relação elift e relação de probabilidades), dado a sua dificuldade de interpretação, por outro lado, foi recomendado o uso de métricas baseadas em diferenças (diferença média, diferença normalizada, etc.). Por fim, concluem que o uso de uma única técnica não é suficiente para identificar corretamente a presença de discriminação.