



As máquinas herdarão nosso preconceito?

Ygor Canalli

Professor de Desenvolvimento de Sistemas

Colégio Pedro II

Duque de Caxias

Agenda

1. Exemplos do mundo real
2. Legislação internacional
3. O problema no Brasil
4. Inteligência Artificial vs. Aprendizado de Máquina
5. Causas da injustiça
6. A solução simplista
7. Desafios
8. Como posso contribuir?

Exemplos do mundo real

As máquinas herdarão nosso preconceito?¹

O que fazer quando os dados utilizados na inteligência artificial carregam consigo discrepâncias sociais e preconceito?

¹*Outras Palavras As máquinas herdarão nosso preconceito?*
https://outraspalavras.net/crise-civilizatoria/as-maquinas-herdarao-nossos-preconceitos/?utm_source=newsletter&utm_medium=email&utm_campaign=27_2&utm_term=2019-02-28.

- O *COMPAS* é um estimador de risco de reincidência criminal desenvolvido pela empresa *Northpointe* e utilizada pelo sistema de justiça de alguns estados americanos²
- A agência independente de jornalismo investigativo ProPublica comparou as previsões com a reincidência efetiva após dois anos.
- As análises foram publicadas em um artigo³ de grande repercussão

²*Northpointe COMPAS risk assesment tool.*

<https://wisconsin.northpointesuite.com/Production/Login.aspx>.

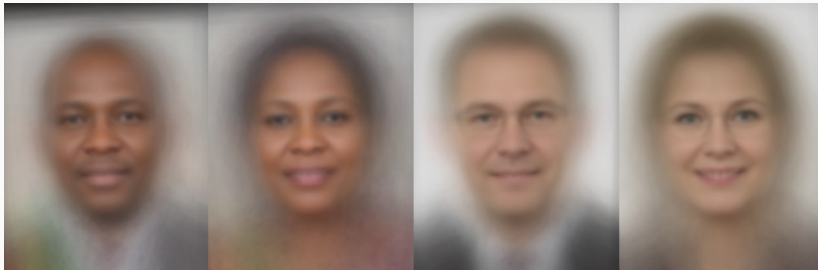
³*ProPublica How We Analyzed the COMPAS Recidivism Algorithm.*

<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.

As análises conduzidas pelo ProPublica revelam que:

- Réus **negros** tem seu risco de **reincidência superestimado**, sendo erroneamente classificados como de alto risco quase duas vezes mais que brancos (45% vs. 23%)
- Réus **brancos** tem seu risco de **reincidência subestimado**, sendo erroneamente classificados como de baixo risco quase duas vezes mais que negros (48% vs. 28%)
- Réus negros tem 45% mais chance de serem classificados como de alto risco de reincidência
- Quando se trata de crimes violentos, réus negros tem 77% mais chance de serem classificados como de alto risco de reincidência



















Um estudo semelhante publicado pelo Projeto Gender Shades⁴ revela um problema similar na previsão de gênero a partir de fotografias.



⁴Gender Shades. <http://gendershades.org/overview.html>.

Gender Shades

Grandes ferramentas de reconhecimento facial erram muito mais para pessoas negras:

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0% 	79.2% 	100% 	98.3% 	20.8% 
 FACE++	99.3% 	65.5% 	99.2% 	94.0% 	33.8% 
 IBM	88.0% 	65.3% 	99.7% 	92.9% 	34.4% 

- Word2Vec⁵ é uma ferramenta desenvolvida pelo Google que utiliza redes neurais para transformar palavras em vetores, uma espécie de linguagem semântica universal
- Tecnologias como esta permitem que ferramentas como o Google Translate traduza de um idioma para o outro passando por uma linguagem intermediária universal
- As línguas não são manualmente descritas, com gramática e dicionário, as redes neurais aprendem a partir de textos da web

⁵Tomas Mikolov et al. “Distributed Representations of Words and Phrases and their Compositionality”. In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges et al. Curran Associates, Inc., 2013, pp. 3111–3119. URL: <http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>.

Homem está para programador assim como mulher está para dona de casa?

- Um estudo conjunto da Universidade de Boston e Microsoft⁶, publicado em uma revista científica especializada em processamento de linguagem natural, mostra discrepâncias de uma rede neural treinada em um conjunto textual de notícias.

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}$$

⁶Tolga Bolukbasi et al. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Advances in Neural Information Processing Systems* 29. Ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 4349–4357. URL: <http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf>.

Gender stereotype *she-he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

Gender appropriate *she-he* analogies

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Extreme *she*

1. homemaker
2. nurse
3. receptionist
4. librarian
5. socialite
6. hairdresser
7. nanny
8. bookkeeper
9. stylist
10. housekeeper

Extreme *he*

1. maestro
2. skipper
3. protege
4. philosopher
5. captain
6. architect
7. financier
8. warrior
9. broadcaster
10. magician

- Recomendação de conteúdo com viés em plataformas de *streaming*
- Taxas de entrega diferenciadas em bairros desprivilegiados
- Acesso reduzido ao crédito para grupos demográficos
- Ações incoerentes de segurança pública

Legislação internacional

Trecho do Recital 71 da Regulação Geral de Proteção de Dados da União Europeia⁷:

A fim de garantir um tratamento justo e transparente em relação ao titular dos dados, levando em consideração as circunstâncias e o contexto específicos em que os dados pessoais são processados, o responsável pelo tratamento deve usar procedimentos matemáticos ou estatísticos adequados...

Dentre outras coisas, a legislação europeia exige direito à explicação de decisões tomadas a partir de dados.

Não se restringe aos países da União Europeia, mas a qualquer um que retenha dados de cidadãos europeus.

⁷European Union Law General Data Protection Regulation.
<https://eur-lex.europa.eu/eli/reg/2016/679/oj>.

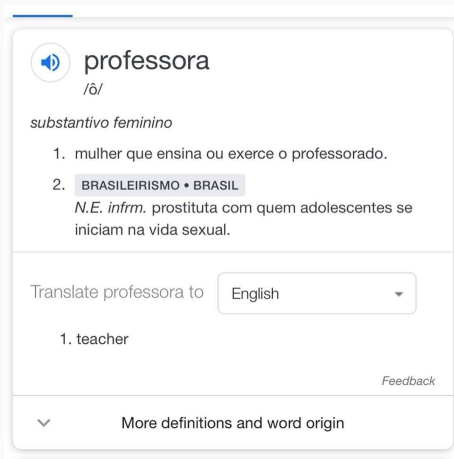
O problema no Brasil

Mas porque todos estes exemplos são em inglês?

- Além do problema em si, faltam estudos em casos no Brasil e em português

E no Brasil?

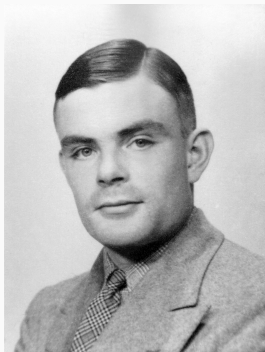
Mas não faltam problemas⁸...



A screenshot of the Google Translate interface. At the top, the word "professora" is entered, with a speaker icon and the phonetic transcription "/ô/". Below this, it is identified as a "substantivo feminino". Two definitions are listed: 1. "mulher que ensina ou exerce o professorado." and 2. "BRASILEIRISMO • BRASIL N.E. *infrm.* prostituta com quem adolescentes se iniciam na vida sexual." The second definition is highlighted with a grey background. Below the definitions, there is a section for translation: "Translate professora to" followed by a dropdown menu set to "English". The translation "1. teacher" is shown. At the bottom, there is a "Feedback" link and a button labeled "More definitions and word origin" with a downward arrow.

⁸Olhar Digital Google remove definição de professora como 'prostituta' no dicionário.
<https://olhardigital.com.br/noticia/google-remove-definicao-de-professora-como-prostituta-no-dicionario/91987>.

Inteligência Artificial vs. Aprendizado de Máquina



A pergunta original, "As máquinas podem pensar?" Acredito ser sem sentido demais para merecer discussão. Não obstante, acredito que no final do século o uso de palavras e opiniões educadas em geral tenha mudado tanto que seremos capazes de falar sobre o pensamento de máquinas sem esperar ser contrariados.

Allan Turing⁹

⁹Alan Turing *Computing machinery and intelligence*.
<http://www.calculamus.org/lect/08szt-intel/tur-paper-local.html>.

Aurélien Géron

O aprendizado de máquina é a ciência (e arte) da programação de computadores para que eles possam aprender com dados¹⁰.

Antonio Gulli e Sujit Pal

Aprendizado de Máquina é uma subárea da grande área de Inteligência Artificial que se concentra em conferir aos computadores a habilidade aprender sem a necessidade de ser explicitamente programados¹¹.

¹⁰Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 1st. O'Reilly Media, Inc., 2017. ISBN: 1491962291, 9781491962299.

¹¹Antonio Gulli and Sujit Pal. *Deep Learning with Keras*. Packt Publishing, 2017. ISBN: 1787128423, 9781787128422.

Causas da injustiça

Causas da injustiça

- O fenômeno social
 - Desigualdade e comportamento existem no espaço amostral
 - Ex.: Mulheres ganham menos que homens para desempenhar a mesma função
- Viés de amostragem
 - Frequentemente os dados não são representativos da população
 - Ex.: Um conjunto de dados pode conter muito menos negros do que na demografia
- Aquisição imprecisa de dados
 - As informações contidas nos dados podem não refletir a realidade
 - Ex.: Um árabe pode ser incorretamente registrado como islamista somente pela etnia
- Problemas algorítmicos
 - Discrepâncias causadas por erros técnicos do estimador

A solução simplista

- Chamamos os atributos passíveis de injustiça e preconceito de *atributo sensível*
- Ex.: raça, nacionalidade, credo, gênero, idade
- Porque não removemos os atributos sensíveis dos dados a serem aprendidos?

O efeito redlining

- Se removermos os atributos sensíveis, outros preditivos indiretos podem ser encontrados nos dados
- Ex.: O CEP de residência pode ser um preditivo indireto da raça
- Este efeito conhecido como *redlining*¹² pode agravar os impactos negativos

¹²Gregory D. Squires. “Racial Profiling, Insurance Style: Insurance Redlining and the Uneven Development of Metropolitan Areas”. In: *Journal of Urban Affairs* 25.4 (2003), pp. 391–410.

Desafios

- Este assunto tem sido alvo de pesquisas científicas formais desde o ano de 2008^{13,14}
- Após pouco mais de uma década de pesquisa há algum progresso
- Recentemente o assunto ganhou grande na comunidade acadêmica e o número de pesquisas cresceu¹⁵

¹³Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. “Discrimination-aware data mining”. In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08* (2008), p. 560.

¹⁴Faisal Kamiran and Toon Calders. “Classifying without discriminating”. In: *2009 2nd International Conference on Computer, Control and Communication, IC4 2009* (2009).

¹⁵Ninareh Mehrabi et al. *A Survey on Bias and Fairness in Machine Learning*. Tech. rep. arXiv: 1908.09635v2.

- Ainda não existe toda a ciência necessária para lidar apropriadamente com o problema
- O fenômeno social por si só causa a desigualdade, como aprender algo diferente?
- Existe uma grande dificuldade na obtenção de dados representativos e de qualidade
- Conscientização dos pesquisadores e profissionais da área

- Como definir e medir justiça?
- Ex.:
 - Homens e mulheres devem ter tratamento igual no regime de progressão de pena
 - Mulheres reincidem menos que homens, dados os menos antecedentes
- O que é justiça neste caso?
- Para conseguir avanços relevantes é necessário quantificar a justiça
- Para quantificar é necessário definir
- A definição deve ser cuidadosamente feita de acordo com cada situação

Como posso contribuir?

Como posso contribuir?

- Se informando
- Debatendo e mobilizando sua comunidade
- Atuando profissionalmente de maneira cuidadosa nas área de computação, engenharia e estatística
- Se tornando um pesquisador do assunto. Há muito o que ser feito!

Obrigado!