Racistische Algoritmes

Data bias en ethiek in de informatica

Bram van den Boomen

November 2016

Racistische Algoritmes

Inleiding

In de Verenigde Staten wordt op dit moment een techniek gebruikt in de rechtszaal genaamd "risk assessment score". Feitelijk is dit een leer-algoritme wat heeft geleerd van eerdere zaken. Op basis van deze kennis geeft het algoritme de verdachte een waarschijnlijkheids-score van 1 tot 10 die uitdrukt hoe groot de kans is dat deze persoon opnieuw de fout in zal gaan.

Uit recent onderzoek van onder andere onderzoeksjournalisten van nieuwsmedium "ProPublica" blijkt nu dat het algoritme achter de risk-assessment software zwarte mensen systematisch een hogere score geeft dan blanke mensen. Dit onderzoek lichtte een aantal kernpunten toe die uit hun analyse van de data bleek: (Julia Angwin en Kirchner, 2016)

1. In meerdere gevallen werden zwarte mensen die terecht stonden voor bijvoorbeeld diefstal met een hogere risico-score gewaardeerd dan blanke mensen met een historie van geweldpleging en gewapende diefstal.

2. Analyse van een groot aantal scores in de staat Florida liet zien dat niet alleen de gemiddelde score van zwarten hoger lag dan die van blanken, maar dat de verdeling van de scores nauwelijks op elkaar leek. Scores van zwarte mensen waren uniform verdeeld over alle scores, elke score kwam ongeveer even vaak voor, terwijl bij blanke mensen lage scores veel vaker voorkwamen dan hoge scores.

3. Tot slot bleek uit cijfers over recidivisme, hetgeen wat de risk-assesment software poogt te beoordelen, dat de in het vorige punt genoemde verdeling precies omgekeerd zou moeten zijn. De data uit Florida liet zien dat type-I fouten vaker voorkomen bij zwarten, dat wil zeggen: Meer zwarte mensen gingen niet opnieuw de fout in ondanks hun hoge score (44,9%) dan blanke mensen (23,5%). Ook kwamen type-II fouten vaker voor bij blanken dan bij zwarten: meer blanke mensen gingen opnieuw de fout in ondanks hun hoge score (47.7%) dan zwarte mensen (28%).

Hoewel de uitkomst van dit onderzoek schokkend is, is deze niet verrassend. Het probleem in dit geval heet "Data-Bias". Dat wil zeggen: op het moment dat de data in een onderzoek niet representatief is voor de werkelijkheid is de kans groot dat de uitkomst van analyse van deze data op dezelfde misrepresentatie zal uitkomen als de data zelf. Nog makkelijker gezegd: als je alleen data van zwarte criminelen verzamelt kan je vanzelf concluderen dat alleen zwarte mensen crimineel zijn.

Machine-learning en Data-Bias

Hoewel Data-Bias hoofdzakelijk een probleem was uit de statistiek is het probleem bijna net zo sterk aanwezig in de sub-discipline van de informatica, machine-learning. Ook dit is niet verrassend, machine-learning gebruikt veel technieken uit de statistiek. Hoewel het probleem in de statistiek al vrij ernstig is, is het misschien nog wel meer aanwezig in het vakgebied van machine-learning. Dit komt naar mijn idee door de volgende punten:

1. Machine-learning is een erg complex vakgebied.

2. De technieken waarmee de data wordt geanalyseerd zijn vaak onbekend omdat de code vrijwel altijd geheim is.

3. De datasets waar machine-learning mee omgaat kunnen nagenoeg oneindig groot zijn.

Uit deze punten wordt duidelijk waar statistiek en machine-learning van elkaar verschillen. Tijdens de analyse van de data kan een statisticus herkennen dat er sprake is van data-bias door de uitkomst te bestuderen. Op het moment dat machine-learning de plaats van de statisticus inneemt is deze "check" er niet meer. Het systeem is ontworpen om alleen de conclusie te geven van zijn analyse. In het geval van de risk-assessment is het enige waarmee data-bias herkend kan worden een serie nummers tussen 1 en 10. De uitkomst van het algoritme is te eenvoudig om te kunnen analyseren (zonder uitvoerig onderzoek zoals ProPublica heeft gedaan), de input-data is te complex om handmatig te analyseren, de software is niet inzichtelijk, en waar de software inzichtelijk is, is deze alleen toegankelijk voor andere data-ingenieurs. Daarmee is de enige die data-bias nog makkelijk kan herkennen de data-ingenieur die het systeem ontwerpt en test.

Ethische machine-learning

Praktisch gezien zijn er twee mogelijkheden voor data-ingenieurs om discriminatie te voorkomen: Een eerste mogelijkheid is het ontwijken of bewerken van discriminatie-gevoelige data. Dit is in veel gevallen niet mogelijk omdat data elkaar beïnvloedt (Kamishima e.a., 2012). Door het ontwijken van bijvoorbeeld het datapunt "ras-achtergrond" bij risk-assessment zal de risk-assessment niet minder racistisch zijn. Omdat ras onder meer invloed heeft op woonplaats, inkomen, wereldbeeld, etcetera, zullen al deze kenmerken ook *identifiers* zijn voor ras. Het weglaten van deze data leidt ook tot het verlies van nuttige data en het toenemen van willekeur in de data (Ristanoski, Liu en Bailey, 2013). Deze aanpak heeft niet de voorkeur omdat hoewel het discriminatie in data kan verminderen, de data ook minder bruikbaar is.

Een andere manier is het aanpassen van machine-learning algoritmes. In deze hoek is al vooruitgang geboekt. Niet alleen zijn er technieken om "scheve" data te compenseren of anticiperen, er is ook mogelijkheid om de discriminerende *identifiers* te herkennen (Ristanoski, Liu en Bailey, 2013). In het paper "Discrimination aware classification for imbalanced datasets" worden verschillende technieken uiteen gezet en ook dit blijkt een complex probleem te zijn. Discriminatie in data komt niet alleen voor door oververtegenwoordiging, maar ook (vaker zelfs) door ondervertegenwoordiging. Ook kan discriminatie leiden tot benadelen van een kleine groep maar ook het bevoordelen van een grote groep. Daarnaast zal compenseren voor discriminatie vrijwel altijd leiden tot een grotere foutmarge in de data-analyse (Ristanoski, Liu en Bailey, 2013). Ook het artikel van Kamishima noemt een aantal vormen van bevooroordeling en noemt "Negative Legacy" (een negatieve bias in het algoritme door negatieve data), precies datgene wat er aan de hand is bij de risk-assessment case, als een van de meest problematische vormen van discriminatie om te herkennen en tegen te gaan. Volgens dit onderzoek is een relatief kleine eerlijke dataset nodig om dit probleem op te lossen door middel van "transfer learning" (Kamishima e.a., 2012). Deze techniek probeert een referentiekader te krijgen door te leren van een kleine dataset om vervolgens van de grote dataset te leren binnen het referentiekader van de kleine dataset.

Ook zie ik een ander voordeel van de uiteenzetting door deze onderzoeken. Het feit dat de technieken die nu gebruikt kunnen worden voor het compenseren van scheve en discriminatoire datasets betekent ook dat dit soort algoritmes discriminatie kunnen herkennen. Ik moet daarbij wel een scherpe kanttekening plaatsen dat met de huidige methodes slecht of geen onderscheid kan worden gemaakt tussen incomplete data, foute datavergaring en discriminatie. Ontwikkeling in deze richting zou technieken opleveren die niet alleen helpen met het ontwijken van discriminatie maar discriminatie ook actief te bestrijden.

Big Data processes codify the past, they do

> not invent the future. Doing that requires moral imagination, and that's something only humans can provide. (O'Neil, 2016)

Ethisch Programmeren

Dit neemt niet weg dat het zo blijft dat data-ingenieurs en programmeurs verantwoordelijk zijn voor deze ontwikkeling en dat het data-ingenieurs en programmeurs zijn die zullen moeten blijven waken voor discriminatie in machine learning. Er ligt dus een grote verantwoordelijkheid bij de ontwikkelaars van dergelijke software en dat is een probleem. Ten eerste zijn veel programmeurs niet voorbereid op het nadenken over vraagstukken van discriminatie of ethiek. De opleiding Informatica van de Universiteit Utrecht biedt bijvoorbeeld geen enkel vak wat toekomstige onderzoekers, projectleiders, programmeurs en data-ingenieurs voorbereidt op dit soort problematiek (Utrecht, 2016). Dit leidt tot gevallen zoals in Florida waar op het moment dat journalisten en ethici zich over de zaak buigen al veel schade is toegebracht. Ten tweede zijn veel programmeurs niet ten machte om iets te doen aan onethische praktijken als ze ze al herkennen of vindt men het niet hun plaats om er bezwaar over te maken. Een voorbeeld wat dit illustreert is het verhaal van Bill Sourour: hij schrijft in zijn blog "The code I'm still afraid of" hoe hij bij een vorige baan een website maakt die farmaceutische reclame-wetgeving omzeilt in opdracht van een farmaceutisch bedrijf. De website heeft de vorm van een vragenlijst die in alle gevallen adviseert om het nieuwe medicijn van dit bedrijf te gebruiken. Pas toen later bleek dat het medicijn depressie verergert en een aantal jonge vrouwen tot zelfmoord doet aanzetten realiseert Sourour zich dat hij in meer of mindere mate heeft bijgedragen aan deze tragedie (Sourour, 2016).

> We [programmers] rule the world, the world does not know this yet, we don't quite know it yet (...) but we write the rules that go into the machines that execute everything that happens on this planet (Martin, 2016)

"Uncle" Bob (Martin) maakt hier een vrij dramatisch punt, maar dit punt raakt dichter aan de werkelijkheid dan een gemiddeld persoon zou willen toegeven. Er is nog maar weinig in de menselijke wereld dat niet geregeerd wordt door software. En met software bedoelen we niet alleen dat wat op je telefoon en je laptop draait, maar ook de software die je motor aanstuurt en je gaspedaal bedient, ook software die regelt of je wasmachine open mag, ook software die zorgt dat wetten ingevoerd kunnen worden, en als dat nog niet ingrijpend genoeg was, nu ook software die bepaalt hoe lang je in de gevangenis zal verblijven. Het is daarom niet houdbaar om programmeurs alleen op te leiden in het schrijven van code en het bedenken van algoritmes als deze code en algoritmes directe invloed hebben op de levens van individuele mensen.

Machine-learning maakt dit probleem alleen maar ondoorzichtiger, onder andere door de bovengenoemde redenen. Het grote ethische probleem van machine-learning is dat het de schijn van objectiviteit wekt. Het idee dat een machine van data leert geeft mensen, ook programmeurs, het foutieve idee dat de uitkomst volledig onafhankelijk is omdat het tot stand is gekomen zonder inmenging van de mens. Maar, zoals we zien uit het voorbeeld van ProPublica, minder menselijke inmenging maakt niet per definitie een onafhankelijk resultaat. Wel maakt het mensen minder waakzaam en op die manier zelfs minder bestand tegen menselijke fouten. Zolang software en machine-learning het leven van mensen beïnvloedt en zolang software en data door mensen wordt gemaakt zit er ook in machine-learning een menselijke factor waarvoor we zullen moeten waken.

Referenties

Kamishima, Toshihiro e.a. (2012). „Fairness-aware classifier with prejudice remover regularizer". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, p. 35–50.

Ristanoski, Goce, Wei Liu en James Bailey (2013). „Discrimination Aware Classification for Imbalanced Datasets". In: *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*. CIKM '13. San Francisco, California, USA: ACM, p. 1529–1532. ISBN: 978-1-4503-2263-8. DOI: 10.1145/2505515.2507836. URL: http://doi.acm.org/10.1145/2505515.2507836.

Julia Angwin Jeff Larson, Surya Mattu en Lauren Kirchner (2016). *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. (Geraadpleegd op 18/11/2016).

Martin, Robert (2016). *"Uncle"Bob Martin - "The Future of Programming- YouTube*. https://www.youtube.com/watch?v=ecIWPzGEbFc&feature=youtu.be&t=1h9m49s. (Geraadpleegd op 23/11/2016).

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Penguin Books Limited. ISBN: 9780141985428. URL: https://books.google.nl/books?id=60n0DAAAQBAJ.

Sourour, Bill (2016). *The code I'm still ashamed of*. https://medium.freecodecamp.com/the-code-im-still-ashamed-of-e4c021dff55e. (Geraadpleegd op 23/11/2016).

Utrecht, Universiteit (2016). *Studieprogramma - Informatica - Bachelors - Universiteit Utrecht*. http://www.uu.nl/bachelors/informatica/studieprogramma. (Geraadpleegd op 23/11/2016).