

Nicos Starreveld

Platform Wiskunde Nederland
Korteweg-de Vries Instituut voor Wiskunde
Universiteit van Amsterdam
n.j.starreveld@uva.nl

Interview Tim van Erven

De wiskundige fundering voor uitlegbare kunstmatige intelligentie

Tim van Erven, wiskundige aan de Universiteit van Amsterdam, heeft een VICI-beurs toegekend gekregen door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). Zijn voorstel gaat over “De wiskundige fundering voor uitlegbare kunstmatige intelligentie”. Met dit interview willen we meer leren over dit onderwerp en van Tim horen wat zijn visie is. Kunstmatige intelligentie (AI) wordt steeds beter, maar hoe deze tot beslissingen komt, blijft vaak ondoorgrondelijk. Deze beslissingen hangen af van vele onbegrijpelijke getallen in de software. Tim van Erven ontwikkelt in zijn onderzoek wiskundige methoden waarmee men beter kan begrijpen waarom AI-algoritmes een specifiek antwoord geven. Dit vakgebied wordt aangeduid als “explainable AI”.

AI is zo’n breed begrip geworden dat tegenwoordig alles wat met algoritmes te maken heeft, onder AI valt. Wat mij betreft probeer je binnen AI de computer iets slims te laten doen. Maar je ziet ook dat alle moderne doorbraken eigenlijk uit de machine learning komen, wat een concreter en afgebakend deelgebied is. Het grootste deel van machine learning-onderzoek

Tim, ten eerste gefeliciteerd met je VICI beurs. Om te beginnen: ik ben benieuwd naar hoe je interesse in AI is ontstaan?

Ik was altijd enthousiast over AI als onderwerp, daarom had ik de keuze gemaakt om een studie Kunstmatige Intelligentie te gaan doen. Tijdens mijn studie merkte ik dat ik concepten op een dieper niveau wilde begrijpen, en wiskunde is daar vaak geschikt voor. Tijdens mijn promotieonderzoek heb ik veel wiskunde moeten bijleren. Ik was ook een beetje jaloers op alle wiskundigen die maattheorie en functionaalanalyse bijvoorbeeld al hadden geleerd. Ik moest het allemaal gaan leren tijdens mijn promotietijd. Maar later besepte ik dat deze houding ook voordelen had, als ik iets uit de wiskunde nodig had dan moest ik die kennis gewoon gaan bijleren. Voor mij kon de wiskunde een dieper begripsniveau bieden van hoe AI werkt.

We horen het overal in het nieuws en in moderne technologieën, zeker na de lancering van ChatGPT. Wat betekent AI eigenlijk?



wordt gedreven door empirische evaluatie: iemand verzint een methode en test deze op een dataset. In de meer theoretische benadering verzin je een algoritme en daarna bewijs je garanties over hoe goed dat werkt. Vervolgens kun je bijvoorbeeld kijken hoe goed je het verband kan analyseren tussen wat je algoritme moet leren en hoeveel data je nodig hebt.

Sommige ontwikkelingen in de machine learning zijn historisch gezien vanuit de theorie gedreven: theoretici hebben een algoritme voorgesteld en vervolgens is het in de praktijk heel succesvol gebleken. Voorbeelden daarvan zijn support vector machines en boosting. Voor kleine datasets zijn dat nog steeds state-of-the-art methodes. Wat je tegenwoordig ziet, is dat neurale netwerken en deep learning voor plaatjes, audio en taal niet te verslaan zijn. De ontwikkelingen in deep learning zijn grotendeels gedreven door empirisch werk. Mensen gaan vooral verschillende algoritmes testen zonder te bewijzen hoe goed deze werken.

En kun je iets meer vertellen over wat uitlegbare AI betekent?

Het doel in machine learning was in het verleden altijd om zo goed mogelijk een voorspelling te maken. De taak die je moet leren zo accuraat mogelijk leren. Omdat alle aandacht lag op accurate voorspellingen maken werden AI-modellen beschouwd als een “black-box”. Het model genereert een output, bijvoorbeeld “er staat een mens in deze foto”, zonder bewijs te geven hoe het tot deze beslissing is gekomen.

In explainable (uitlegbare) AI is het doel anders: een taak goed uitvoeren is vaak niet voldoende. Zeker niet als een programma een effect heeft in de samenleving. Verantwoording is belangrijk. De term “uitlegbaar” betekent dat je model ook informatie geeft over waarom een specifieke beslissing is genomen. Denk aan een bank bijvoorbeeld, die met een AI-tool aanvragen voor leningen wil controleren. Dan willen wethouders en burgers dat zo’n tool niet alleen een “ja” of “nee” uittekent, maar ook dat er voldoende uitleg wordt

gegeven waarom iemand wel of niet de lening kan krijgen.

Er is op dit moment geen consensus over wat goede uitleg is. Wiskunde kan hier een bijdrage aan leveren. Wiskunde dwingt je om alles concreet en precies te maken! In de afgelopen jaren hebben we in mijn team goed gekeken naar hoe we “uitlegbaarheid” in wiskundige definities kunnen vatten. Laten we terugkijken naar het voorbeeld met de bankleningen. Hier krijg je weer de vraag: wat zijn belangrijke voorwaarden waaraan deze uitleg moet voldoen? Een bekende methode om uitleg concreet te maken heet “algorithmic recourse”. Naast een “ja” of “nee” antwoord moet je model ook voldoende toelichting geven opdat diegenen die de lening aanvragen weten wat ze moeten doen om de lening wel te krijgen. Een voorbeeld van uitleg zou kunnen zijn dat “als je je inkomen verhoogt tot boven de 40.000, dan kun je wel deze lening krijgen”. Je stelt dus de voorwaarde dat je AI-model actionable feedback moet geven.

Naast actionable feedback wil je vaak ook kwaliteitseisen stellen aan hoe je verwacht dat algorithmic recourse gaat werken. Bijvoorbeeld je wilt dat mensen die vergelijkbare aanvragen inleveren ook vergelijkbare antwoorden en uitleg krijgen. Dit noemen we een continuïteitsvoorwaarde. Een tweede belangrijke kwaliteitseis is dat je AI-model geen beslissingen neemt op basis van irrelevante criteria. Denk bijvoorbeeld aan MRI-scans, waar op sommige scans het logo van de fabrikant kan staan. Een AI-model zal soms een correlatie detecteren tussen het logo en wat er op de scan staat. Je wilt voorkomen dat het antwoord van een model afhangt van zulke irrelevante kenmerken.

In onze meest recente studie hebben we een wiskundige formulering gegeven van algorithmic recourse die in lijn is met bestaande methoden en technieken. Maar dit alleen is niet voldoende om mensen te enthousiasmeren voor je onderzoek. Je moet ook laten zien dat je wiskundige formulering iets kan bijdragen. Recentelijk hebben we wiskundig bewezen dat er

geen methode kan bestaan die zowel algorithmic recourse aanbiedt, alsook aan een continuïteits-kwaliteitseis voldoet! En mensen waren lang op zoek naar zo’n methode, maar dit kan helemaal niet. Je moet je doel veranderen. Met wiskunde kun je redeneren over alle mogelijke methodes, waarvan er oneindig veel zijn.

Kun je ook iets meer vertellen over wat je wilt bereiken met je VICI beurs?

In de twee publicaties die we tot nu toe hebben geschreven, zijn we vooral kritisch geweest. We hebben namelijk bewezen dat het niet mogelijk is om sommige eisen te combineren. Zoals hierboven genoemd, is het niet mogelijk om een methode te vinden die algorithmic recourse en continuïteit combineert. Met mijn VICI wil ik laten zien hoe het dan wel kan. Niet alleen kritisch zijn, maar ook constructief bijdragen aan het vakgebied. We willen een toolbox van wiskundige definities en methodes ontwikkelen, waarvan je weet dat als ze onder gegeven voorwaarden gebruikt worden, je dan ook kunt garanderen dat ze goed werken.

Je bent ook lid van het bestuur van AIM, en recent ook voorzitter geworden van de commissie onderzoek van Platform Wiskunde Nederland. Wat was de motivatie om ook bestuurlijke taken op je te nemen?

We hebben gemerkt dat er veel mensen over het hele land verspreid zijn die onderzoek doen dat belangrijk is voor AI. Denk aan statistiek, causaliteit, inverse problemen, netwerkentheorie, dynamische systemen, enzovoort. Met zowel AIM als met de commissie onderzoek van PWN willen we mensen met elkaar in contact brengen, een gemeenschap vormen en wiskundigen ondersteunen. Op de lange termijn willen we dat mensen beter weten hoe ze elkaar kunnen zoeken en misschien gezamenlijk beurzen kunnen gaan aanvragen. Bovendien wil ik ook iets teruggeven aan de wiskundegemeenschap. Ik geloof sterk dat als je bij een gemeenschap hoort en daarvan profiteert, dat je dan ook iets terug moet geven!

