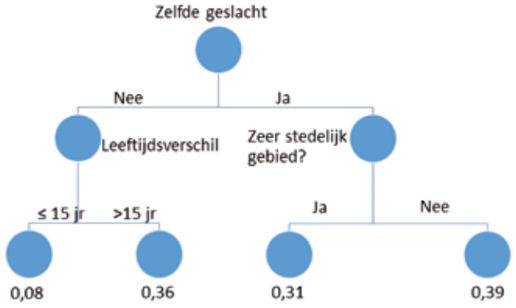


Figuur 1. Een fictief voorbeeld van een beslisboom

eindknoop van een boom, een zogenaamd blad, hoort bij een specifieke, homogene groep. Bijvoorbeeld alle duo's op één adres van hetzelfde geslacht met minder dan 15 jaar leeftijdsverschil. Voor ieder blad wordt er ofwel een waarde van de doelvariabele geschat (hier: de kans op twee huishoudens op één adres), of een correctie ten opzichte van een eerdere schatting.

Als resultaat van Gradient Boosting krijgen we voor ieder adres een geschatte kans voor één versus twee huishoudens. Deze kansen kunnen we gebruiken voor het afleiden van imputaties. Dit is gedaan door het trekken van random getallen tussen nul en één. Stel dat we voor een bepaald adres afleiden dat er 60% kans is op één huishouden en 40% op twee huishoudens. We trekken dan vervolgens uit een uniforme verdeling op het interval [0,1]. Als de uitkomst kleiner of gelijk is aan 0,6 dan imputeren we 'één huishouden', is de uitkomst groter dan imputeren we 'twee huishoudens'. Deze stochastische methode om imputaties af te leiden wijkt af van de gangbare methode bij machine learning die kansen afrondt. Als er een kans van 0,6 is geschat op één huishouden, dan wordt die afgerond naar 1 en wordt er dus 'één huishouden' geïmputeerd. Hoewel deze benadering op individueel niveau de kleinste schattingsfout oplevert, heeft deze als effect dat de verdeling van de doelvariabele sterk af kan wijken van die van de geobserveerde waarden. Stel bijvoorbeeld dat voor alle adressen 60% kans wordt geschat op 'één huishouden'. Afronden zou dan betekenen dat voor alle adressen 'één huishouden' wordt geïmputeerd. De categorie 'één huishouden' wordt dan dus geobserveerd in 60% van de adressen, maar komt voor in 100% van de imputaties. Voor veel toepassingen bij statistische bureaus is dit zeer ongewenst, aangezien een juiste verdeling op geaggregeerd niveau belangrijker is dan een nauwkeurige schatting op individueel niveau.

Resultaten

Gradient Boosting blijkt betere schattingen te geven dan de huidige regressiemethode. Door toepassing van de methode op woningen met een bekende huishoudsamenstelling (zogenaamde cross-validatie) is een inschatting te maken van de precisie van de imputaties. Gradient Boosting classificeert 77% van de adressen correct, terwijl dit percentage voor de huidige regressiemethode op 74% ligt. Daarnaast is er ook gekeken naar de zogenaamde AUC (Area Under the Receiver Operating Characteristics (ROC) Curve). Eén van de interpretaties hiervan is hoe waarschijnlijk het is dat een grotere kans voor klasse 1 wordt voorspeld voor iemand uit klasse 1 vergeleken met iemand uit klasse 0. De score ligt tussen 0,5 en 1. Een score van 0,5 betekent dat een model willekeurig gokt en 1 staat voor een perfecte discriminatie tussen de twee groepen. De AUC/ROC voor het regressiemodel en Gradient Boosting bedragen respectievelijk 0,87 en 0,90.

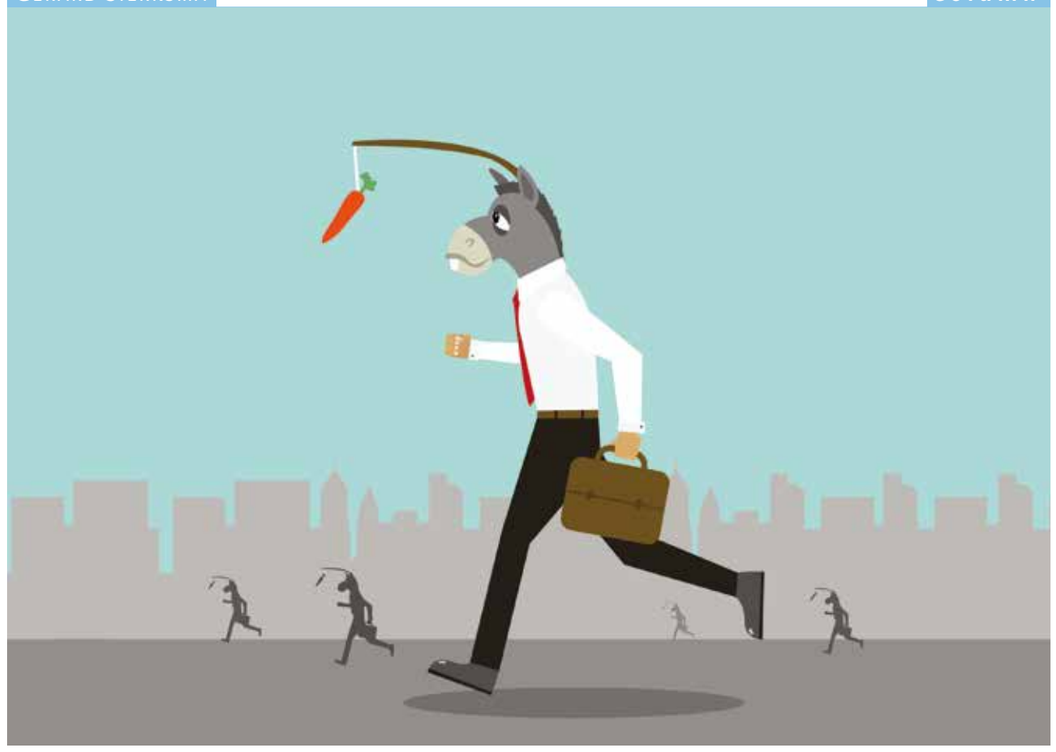
Gradient Boosting versus regressie

Zoals hierboven beschreven geeft Gradient Boosting nauwkeurigere schattingen dan de huidige regressiemethode. Gradient Boosting heeft echter ook andere voordelen. Zo is deze methode makkelijker toepasbaar. Bij regressie moet van te voren worden bepaald wat de relatie is tussen de doelvariabele en de verklarende variabelen. Dit kan bijvoorbeeld een lineair verband zijn, maar ook een exponentieel verband. Bij Gradient Boosting is het niet nodig om dit van te voren te vast te leggen; dit wordt door de methode bepaald. Ook kan Gradient Boosting eenvoudiger dan regressie overweg met ontbrekende waarden in verklarende variabelen. Een nadeel van Gradient Boosting is dat de uitkomsten lastiger zijn te duiden. Het is niet erg eenvoudig om achteraf te achterhalen hoe specifieke schattingen tot stand zijn gekomen. Bij regressie is dit makkelijker. Vanwege de bovenstaande voordelen is geadviseerd om Gradient Boosting te implementeren. Momenteel wordt de methode verder uitgewerkt en worden er voorbereidingen getroffen om de methode in het productieproces op te nemen.

Jacco Daalmans heeft econometrie gestudeerd aan Tilburg University. Hij werkt als methodoloog voor het Centraal Bureau voor de Statistiek en is in 2019 gepromoveerd op toepassingen van macro-integratie in de officiële statistiek.
E-mail: j.daalmans@cbs.nl

Gerard Sierksma

column



Over P≠NP en een Eeuwige Student

In het decembernummer van het tijdschrift *NewScientist* staat een mooi verhaal met de titel 'P=NP?'. Een vraag als titel dus. Die vraag betreft een van de zeven beroemde millenniumproblemen met een miljoen dollar voor elke eerste oplosser. Het antwoord laat al ruim 50 jaar op zich wachten en Tamara Florijn, de auteur, twijfelt of het er ooit van komt. Ze eindigt nogal pessimistisch met zinsneden als: '(...) dat het nog wel honderd jaar kan duren voordat (...)' en 'Misschien zullen we wel nooit weten of (...)'. Je moet kennelijk wel een beetje gek zijn om er tijd en energie in te steken. Ik heb zo'n 'gek' gekend, een eeuwige student.

Iedereen heeft wel een beeld van een eeuwige student, maar wat P≠NP of P=NP betekent is minder bekend, zeker ook doordat de uitleg ervan een behoorlijke dosis wiskundige voorkennis vereist. Dat de P en NP niets van doen hebben met parkeren of zo lijkt me duidelijk, maar met wat dan wel? Om te beginnen, de P staat voor 'polynomiaal'. Je zou denken dat NP dan voor 'niet-polynomiaal' staat, maar dat is dan weer niet het geval. Schiet niet op dus, zelfs als ik toevoeg dat NP staat voor 'niet-deterministisch polynomiaal'. Dan maar kort-door-de-bocht uitgelegd, wat Tamara ook doet.

De letter P staat voor de klasse van wiskundige problemen die (met een algoritme) zijn op te lossen in polynomiale tijd, wat kort-door-de-bocht wil zeggen 'snel op te lossen'. En wat is 'snel' dan wel? Ik kom daar zo op terug. Eerst even naar NP. Dat is de klasse van problemen waarvan 'snel kan worden gecheckt' of een gevonden 'oplossing' echt wel oplossing is. Oké, nu 'snel'. Interessant in deze context is dat er een helder onderscheid is tussen 'snel' en 'niet snel'. Een probleem heet 'snel', sommigen zeggen zelfs 'makkelijk', oplosbaar als voor dat probleem een algoritme bestaat waarmee, voor alle mogelijke input data, een oplossing wordt geproduceerd binnen een beperkte tijdspanne, ook wel *real-time* genoemd. Een voorbeeldje. Als om vier uur 's ochtends de vrachtwagens moeten beginnen met het rijden van de routes en alle bezoekadressen zijn bekend om, zeg, drie uur in de morgen, dan heeft het algoritme maximaal een

uur ter beschikking om een optimale planning te berekenen. En dat dag-in-dag-uit, met elke dag andere input data en elke dag dat ene uurtje. Als dat lukt heb je in je computer een snel algoritme ter beschikking en heet het (routerings)probleem snel oplosbaar. Daarentegen een probleem dat 'niet snel' oplosbaar is, kent deze luxe niet en moet het doen met algoritmen die in het slechtste geval uren-dagen-maanden, ja zelfs eeuwen moeten rekenen om een gewenste (vaak een optimale) oplossing te berekenen. Het 'handelsreizigersprobleem' is hiervan het prototypevoorbeeld. In de praktijk heb je aan dergelijke algoritmen dus helemaal niets en wordt gewerkt met algoritmen die slechts suboptimaal de wensen (van de planner) vervullen. Hierbij moet worden aangetekend dat het aantal kandidaatoplossingen exponentieel toeneemt met de omvang van de input data: één locatie erbij betekent onmiddellijk een verdubbeling van het aantal routes (de zoekruimte) waaruit de beste berekend moet worden. Exponentiële groei van de zoekruimte derhalve.

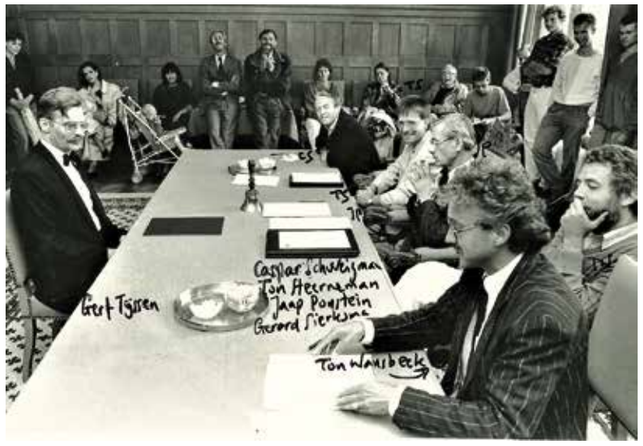
De vraag 'P=NP?' betekent dan: Als je snel kunt controleren of een oplossing van een probleem klopt, kun je dan het probleem zelf ook snel oplossen? Voor alle problemen, waarvoor dus een snel checking-algoritme bestaat, zou dan ook het probleem zelf snel zijn op te lossen, hoe bizar de input data er ook uit ziet. En geloof het of niet, maar als op de vraag P=NP? het antwoord 'ja' is, dan kunnen computers vervolgens veel werk van wiskundigen overnemen en zelfs pincodes van banken kraken. Geen wonder dat we hier te maken hebben met een 1-miljoen dollarprobleem. Nog korter door de bocht met een metafoor: Controleren of een aangeboden 'voorwerp' inderdaad de gezochte speld in de gigagrote hooiberg is moge simpel zijn, maar dat daarmee ook het vinden van die speld zelf simpel is ligt gelukkig niet echt voor de hand. Geen paniek over krakende codes dus, voorlopig.

De nu 82-jarige informaticus Stephen Cook stelde ruim 50 jaar geleden als eerste de vraag 'Is P gelijk aan NP?' Tegenwoordig spreken we van het 'PvsNP-probleem' – met 'versus' dus tussen P en NP –, omdat het gilde der complexiteitswiskundigen inmiddels is gepolariseerd: De paniekzaaiers (pincodes worden gekraakt) die geloven in P=NP, met daartegenover de 'ongelovigen', zij die hun ziel en zaligheid hebben verkocht aan P≠NP. De gelovigen vormen een kwijnende minderheid en in lijn met Tamara moeten we ook vrezen voor het uitsterven van het P≠NP-ras. Ik reken mezelf tot de ongelovigen: ooit wordt bewezen dat P niet gelijk is aan NP en worden geen pincodes gekraakt. En misschien is het nog beter om Tamara te zien als de voorloper van een derde denominatie, waar ik ook wel iets in zie: de PvsNP-agnosten, die beweren dat de vraag zelf eigenlijk helemaal niet heeft bestaan en ook niet zal bestaan. Zeker is dat de kwestie niet bestond voor Cooks eerste formulering ervan. En ooit zou de vraag compleet kunnen verdwijnen zonder antwoord. Misschien, heel misschien gebeurt dat zodra de kwantumcomputer zijn intrede heeft gedaan en NP-problemen, zoals het handelsreizigersprobleem, wel snel maar niet polynomiaal worden opgelost. Het zou dus weleens heel lang kunnen gaan duren. Tja, en dan is de 1-miljoen dollar niks anders geweest dan een in de eeuwigheid verdwenen bijna-grijpbare wortel voor de neus van de ezel.

Koffer is bestemd voor de wortel: de PvsNP 1-miljoen dollar

Gert Tijssen zal midden vijftig zijn geweest. Het was in de zomer van 2005. De televisie stond nog aan toen zijn zuster hem vond. Op de glazen tafel naast hem lag de as van z'n laatste sigaret. De fles leeg, het glas vol. Gert is vanuit Groningen bijgezet in het familiegraf in zijn geboorteplaats Apeldoorn. Een zoektocht op het internet levert mij niks op: geen sterfdag, geen bio, niks. *Zapped into eternity*? Een dikke tien jaar heb ik zijn obsessieve P≠NP?-gezwoeg meegemaakt. Hij was zeker niet gek, wel markant en uiterst scherpzinnig. *Never a dull moment* in de tien jaren met Gert. Maar wat heeft het opgeleverd?

In 1992 studeerde Tijssen cum laude af als de laatste 'eeuwige student'. Daarna heb ik het gewaagd om hem een promotieplaats aan te bieden als vervolg op zijn succesvolle afstudeerscriptie over *cutting stock* problemen.



Gert Tijssen (links) bij zijn afstuderen; rechts aan tafel van boven naar beneden: Caspar Schweigman, Ton Steerneman, Jaap Ponstein, Ton Wansbeek en Gerard Sierksma.)

Hij heeft die aanbieding graag geaccepteerd. In de perioden waarin zijn manische toppen overgingen in depressieve dalen was Gert een uiterst aimabele man die echter meer en meer begon te vergen van luisterende oren. Na een paar maanden als PhD-student veranderde plots zijn gedrag: geen lange monologen meer bij de koffieautomaat en er kwam een soort grauwe stilte om hem heen te hangen. Ineens was hij verdwenen met, in een laatste e-mail, de mededeling dat hij voor een half jaar naar Griekenland was vertrokken, naar zijn zuster in Athene.

Zo'n zes maanden later, op een goede maandagochtend, trof ik hem aan bij de koffieautomaat, monter en zijn oude maatpak wat strakker om de inhoud. Het oude vertrouwde tussen hem en mij was niet verdwenen en zonder omhaal meldde hij te zijn gestopt met snijproblemen. En met een grijns: 'die zitten allemaal in P en zijn dus opgelost.' Maar wat dan wel? Weer die grijns: 'Het wordt vanaf nu de simplexmethode met een nieuwe pivottruc die LP in P brengt' en voegde eraan toe 'LP gaat straks ook met simplex in P.' Ik ben akkoord gegaan, onder de voorwaarde dat er binnen een half jaar een toptijdschriftrijp *paper* van hem zou zijn. *Believe it or not*, in juli 1995 lag zo'n artikel op mijn bureau. Het duurde daarna minstens drie jaar tot de publicatie in *Mathematical Programming*. Die lange aanlooptijd kwam doordat we in eerste instantie het *paper* hebben aangeboden aan het toptijdschrift *Mathematics of Operations Research* en de toenmalige *editor-in-chief* het terugstuurde met de mededeling '*this is folklore*'. Kort gezegd, ging het onder meer over de stelling dat voor LP-modellen met eindige oplossingen geldt: *The dimension of the optimal primal face is equal to the degeneracy degree of the corresponding optimal dual face.* Door Tijssen fraai bewezen met Balinski-Tucker simplex tableaus. Dat klinkt toch niet als folklore, zou ik zeggen. Maar dit terzijde.

Tijssen is gepromoveerd op 24 januari van het jaar 2000. In datzelfde jaar was het dat het PvsNP-probleem verheven werd tot een van de zeven millenniumproblemen met voor elk probleem de hoofdprijs van 1-miljoen dollar, uit te keren door het *Clay Mathematics Institute* aan de allereerste oplosser. Slechts een van de zeven problemen is inmiddels opgelost; 6 miljoen ligt dus nog op de plank. Tien jaren heeft Gert de 1-miljoenwortel voor z'n neus zien bungelen.

In de nadagen van zijn leven veranderde Tijssen in de rijzige grijzende man, die veel ouderen onder ons zich zullen herinneren. Gerts heilig geloof in het vinden van een bewijs voor P≠NP leek zijn regelmatig terugkerende depressies te verhevigen. Zeker driemaal heeft hij me bij de koffieautomaat 'zijn bewijs' van P≠NP verteld. Euforisch. De driemaal die ik me herinner had hij de nacht ervoor niet geslapen, maar wel te diep in het glas met P≠NP gekeken. Een vierde 'bewijs' is er niet meer van gekomen: de fles was leeg, het glas nog vol. De 1-miljoen dollar zit nog in het vat en is voor eeuwig waardevast geïndexeerd. Voor eeuwig?

P.S. Een meer dan opmerkzame lezer van een eerdere versie van deze column vroeg zich af waarom ik niet voor de titel 'In Memoriam Gerhard Antonie Tijssen' heb gekozen. Die uitleg over P en NP kan iedereen toch vinden op Wikipedia? Ze heeft hier een punt. Voor een in memoriam vind ik het in het geval van Gert Tijssen te laat. Tja en dat lange verhaal waarin ik kort-door-de-bocht... En of ik NP-agnost wordt of zoiets? Ik hou me daar eerlijk gezegd niet mee bezig. Wel raad ik elke jonge wetenschapper aan, die een van de zes overgebleven 1-miljoenen wil verdienen, eerst een ontdekking te doen die minimaal in *Math of OR* verschijnt.

Literatuur

Tijssen, G. A., & Sierksma, G. (1998), Balinski-Tucker Simplex Tableaus: Dimensions, degeneracy degrees, and interior points of optimal faces. *Mathematical Programming, 81*, 349–372.

Tijssen, G. A. (2000), *Theoretical and practical aspects of linear optimization*. PhD Thesis, SOM Research School, University of Groningen.

Tijssen, G. A., & Sierksma, G. (2006), Simplex adjacency graphs in linear optimization. *Algorithmic Operations Research, 1,* 46–51.

Gerard Sierksma is emeritus hoogleraar Operations Research aan de Rijksuniversiteit Groningen.
E-mail: g.sierksma@rug.nl