Unsupervised Machine Learning

Joris de Ruiter, derdejaars AI student aan de VU
(jdruiter@few.vu.nl)

Maarten Keijzer is onderzoeker op het gebied van *Machine Learing*. Zijn onderzoek is met name gericht op het toepassen van genetisch programmeren in diverse vakgebieden, zoals hydrologie, hydraulica, bioinformatica en datamining. Hij is momenteel werkzaam bij Chordiant, waar hij intelligente analytische software onderzoekt en ontwerpt. Tevens is hij dit jaar hoofdredacteur van de grootste conferentie op het gebied van Evolutionaire Technieken: GECCO 2006. Zijn wetenschappelijke publicaties zijn vooral gericht op gebieden als genetisch programmeren, evolutionaire computatie, neurale netwerken en andere statistische programmeertechnieken.

Maarten Keijzer

Een mens wordt geboren met beperkte voorkennis. De mens leert spreken, schrijven en rekenen, en de mens leert hoe de wereld om hem of haar heen in elkaar zit. Het feit dat wij onszelf kunnen programmeren stelt ons niet alleen in staat ons goed aan te passen aan sterk wisselende omgevingen, maar zorgt er ook voor dat we complex gedrag kunnen vertonen, zonder dat dit vooraf in ons geprogrammeerd is. Door gebruik te maken van *Machine Learning*-technieken (ML-technieken) komt dit doel ook voor computers in zicht, bijvoorbeeld in adaptieve spamfilters en lopende robots.

'Machine Learning bestaat uit het vinden van lerende algoritmen, en uit het vinden van bewijzen over generaliseren.' Daarmee is het zowel praktisch, als puur wiskundig,

Ook heeft ML veel te maken met statistiek, aangezien beiden te maken hebben met het analyseren van data. Keijzer noemt zijn vakgebied dan ook graag "geavanceerde statistiek": 'Neem bijvoorbeeld het vinden van een gemiddelde. De statistische methode is het nemen van een steekproef waarover een gemiddelde wordt berekend dat vervolgens wordt gegeneraliseerd naar de gehele populatie. Bij ML zou men in dit geval een trainingset maken door een steekproef te nemen, en maakt men tevens een model, wat vervolgens met de eerder verkregen set wordt getraind. Op die manier wordt het model geleidelijk aan beter in het voorspellen van het werkelijke gemiddelde. Een goed model is daarbij niet zozeer perfect op de trainingset, maar kan wel goed generaliseren over de gehele set of populatie.'

"Een goed model is niet zozeer perfect op de trainingset, maar kan wel goed generaliseren over de gehele set of populatie"

Een algemeen principe in ML is het leren door middel van inductieve technieken, wat neerkomt op het generaliseren van het model voorbij de trainingset. Mensen doen dit volgens Keijzer continu: 'Stel je voor dat alle zwanen die we observeren wit zijn (de trainingset), dan nemen we hierdoor aan dat alle zwanen altijd wit zijn (de afgeleide theorie). Dit stelt ons in staat theorieën op te stellen en zo de wereld te begrijpen en te voorspellen. Echter, het is geenszins zeker dat alle zwanen altijd wit zijn. Er hoeft maar één zwarte zwaan te zijn, en onze theorie gaat onderuit. In dat geval zou onze trainingset niet representatief genoeg zijn geweest, iets waarvoor men altijd moet uitkijken bij Machine Learning.'

ML-modellen zijn onder te verdelen in statische en adaptieve modellen. Bij de statische modellen wordt een model éénmaal getraind, waarna het telkens kan worden toegepast. Een adaptief model begint met weinig data en wordt beter naarmate er meer data bij komt. Een voorbeeld van de laatste methode zijn adaptieve *spamfilters*, waarbij het filter zich aanpast aan het soort spam wat langskomt. Ze zijn te vinden in de betere mailprogramma's.

Een ander onderscheid binnen ML is te maken tussen *supervised* en *unsupervised learning*. Dit onderscheid uit zich in het soort algoritmes dat ervoor wordt gebruikt.

Bij supervised learning wordt het model getraind met vooraf gegeven in- en uitvoerparen, een soort antwoordmodel. Gegereneerde uitvoer wordt vergeleken met dit antwoordmodel, waarna eventueel het model wordt aangepast. Vervolgens moet het model kunnen generaliseren over invoer die het niet eerder heeft gezien, waardoor het de uitvoer kan voorspellen. Een mooi voorbeeld is wederom het spamfilter, waarbij de gebruiker van het mailprogramma aangeeft of iets spam is of niet. Op die manier creëert de gebruiker zijn eigen trainingset. De spam komt automatisch binnen (invoer), waarna de gebruiker aangeeft wat er met de spam moet gebeuren (uitvoer). Als de gebruiker dit enigszins consistent doet en het ML-algoritme goed kan generaliseren, dan zal het voortaan bij nieuwe mail zelf kunnen aangeven of het spam is of niet.

Unsupervised learning is het leren zonder hulp van buitenaf. Er zijn geen vooraf vastgestelde in- en uitvoerparen. Het model moet zelf uitzoeken hoe het orde aanbrengt in de data. Een voorbeeld hiervan is clusteren, waarbij objecten worden geclas-

De Connectie
nummer 1, jaargang 2, April 2006

"Software moet vooraf een betrouwbare inschatting kunnen maken of een klant wel in staat is om zijn krediet of hypotheek terug te betalen."

sificeerd en gegroepeerd al naargelang zij op elkaar lijken.

Er zijn inmiddels een hoop algoritmes bedacht waarmee ML kan werken. Enkele voorbeelden zijn: *decision trees*, *Artificial Neural Networks* (ANN's), *het Perceptron*, *Genetic Programming* (GP) en evolutionaire algoritmes. De zakenwereld maakt daar dankbaar gebruik van en toepassingen van ML-technieken zijn dan ook ruim voor handen.

Het bedrijf waar Maarten Keijzer voor werkt is gespecialiseerd in marketing, management en customer experience. De klanten van het bedrijf zijn over het algemeen bedrijven met grote databases vol met consumenteninformatie, zoals naam, leeftijd, sekse, etc. Keijzer's opdracht is dataminingsoftware te ontwerpen, die uit die databases de zinvolle informatie kan halen. Hierbij maakt hij onder andere gebruik van statistische methodes en genetisch programmeren.

Het soort software dat hij ontwerpt kan bijvoorbeeld worden gebruikt binnen de financiële sector bij het verstrekken van kredieten. De software moet dan vooraf een betrouwbare inschatting kunnen maken of een klant wel in staat is om zijn krediet of hypotheek terug te betalen.

Een ander domein waar deze software van pas kan komen is de klantenservice. Hier moet onderscheid gemaakt worden tussen inkomend en uitgaand verkeer. Het inkomend verkeer bestaat bijvoorbeeld uit klanten die een callcenter opbellen met een klacht of een vraag. De klant moet dan zo snel mogelijk en zo goed mogelijk geholpen worden. ML-technieken kunnen hierbij helpen door bijvoorbeeld te leren naar welke afdeling de klant het beste kan worden doorverbonden, of door bijvoorbeeld te bepalen wat het meest geschikte keuzemenu is om de klanten voor te schotelen.

Bij uitgaand verkeer probeert een klantenservice mensen te bereiken, bijvoorbeeld via de telefoon (inderdaad, die vervelende telefoontjes rond etenstijd), post of via sms. Om geld en moeite te besparen dient uitgezocht te worden wie bereikt moet worden, hoe dat moet gebeuren en met wat voor een benadering. Keijzer: 'Traditionele marketeers werken met het opdelen van de consumentenmarkt in segmenten, zoals bijvoorbeeld "jong en wild", of "jong en rijk" Wij gebruiken ML-technieken om dit te doen. Zo geeft een bepaald model bijvoorbeeld de drie beste keuzes, waarna de klantenservice daaruit kan kiezen.'

Behalve binnen bedrijven met grote databases, wordt ML ook gebruikt bij digitale signaalverwerking. Voorbeelden hiervan zijn mobieltjes en *conference call software* waarin adaptieve filters worden gebruikt voor ruisonderdrukking.

De toepassingsgebieden voor ML-technieken zijn zeer wijdverspreid: Zoekmachines, medische diagnose, het detecteren van creditcardfraude, marktanalyse, spraak- en handschriftherkenning en bewegingen van robots. Een uitzonderlijk geval is de ontwikkeling van een neuraal netwerk, ALVINN genaamd, dat geheel autonoom en botsvrij met 70 km/u een auto over een snelweg kan rijden.

Hoewel er misschien geen ML-technieken aan te pas komen is de *pagerank*-techniek van Google als geheel te zien als zelflerende software. Deze techniek meet de populariteit van een internetpagina door onder andere te kijken naar de hoeveelheid bezoekers en het aantal pagina's wat ernaar verwijst.

De game-industrie is als toekomstig domein interessant, bijvoorbeeld voor het bepalen van gedrag van computertegenstanders (*bots*). Vooral bij *first person shooters* (FPS) dienen bots niet zozeer het beste te zijn, maar ze moeten wel leuk zijn om tegen te spelen. Dit vereist dat de bot zich baseert op waarnemingen naar menselijke maat. Hij moet dus bijvoorbeeld niet door muren kunnen kijken en dekking zoeken als er op hem geschoten wordt. Ook moet hij niet altijd raak schieten.

Op de vraag of we in de toekomst nog een echt *killer applicatie* kunnen verwachten, zei dr. Keijzer dat hij graag een geïntegreerd systeem zou zien, waarin gegevens door een heel bedrijf kunnen worden gedeeld. Keijzer verklaart: 'Momenteel zijn gegevens vaak beperkt toegankelijk, waardoor er allerlei misverstanden ontstaan. Een geïntegreerd ML-systeem zou dit kunnen verhelpen.'

Ook zou hij graag massale datamining op het web zien ontstaan, waarbij een persoonlijke agent interactief informatie voor je opzoekt. Een adaptieve robot met een combinatie van verscheidene ML-technieken, die desondanks toch nog in toom is te houden door mensen, lijkt hem ook wel wat.

Er zijn in de wereld een hoop problemen, en een hoop softwaretechnieken om ze op te lossen. ML-technieken zijn vooral handig bij open-einde-problemen, toepassingen waarbij adaptief gedrag gewenst is of toepassingen waarbij coderen niet mogelijk is, bijvoorbeeld bij handschrift-, spraak- en beeldherkenning. Deze domeinen zijn zo complex en ambigue, dat normaal coderen niet werkt. Onze hersenen lossen dit op met een enorm neuraal netwerk, Machine Learning zou dit met de artificiële variant hiervan moeten kunnen (ANN's).

Enthousiast geworden van de vele applicaties vroeg ik me af hoe het werk zelf is. Wat Keijzer erg bevalt is het feit dat je niet zelf alles hoeft te programmeren, zoals bijvoorbeeld bij kennisgebaseerde systemen, maar dat de computer zichzelf programmeert. Tevens is er altijd een open einde. Het staat nooit van tevoren vast wat uit je programma zal komen. Een carrière in Machine Learning is echter niet voor iedereen weggelegd: statistiek en wiskunde zijn belangrijke factoren in ML, en niet iedereen vindt dat even leuk. ø