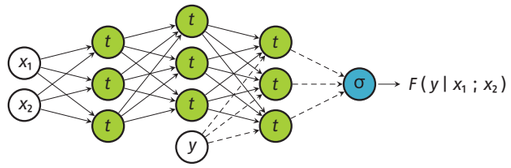
DOOR ANNE VAN DER SCHEER

Een kansrijk model voor Business Analytics

Het schatten van 'niet-parametrische' modellen met Machine Learning (ML) methoden kan aanvoelen als een concessie aan stochastiek. Een bijzonder neuraal netwerk maakt het echter mogelijk om een generieke verdelingsfunctie te schatten op basis van het maximum likelihood principe. Bovendien kan een informatiecriterium worden gebruikt om een optimaal model te selecteren. Hiermee ontstaat een interessant alternatief of benchmark voor het Generalized Linear Model (GLM).

NEURAAL NETWERK

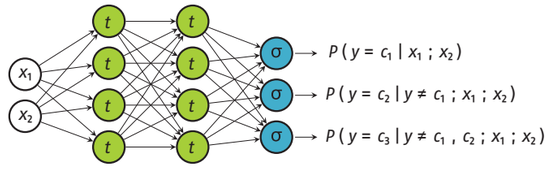
Binnen ML zijn neurale netwerken een bekende techniek om complexe functies te benaderen. Het basiselement is de neuron, die een lineaire combinatie van ingangswaarden plus een *bias* omzet naar een uitgangswaarde via een activatiefunctie, bijvoorbeeld de logistische (sigmoid) functie. In een netwerk worden neuronen gegroepeerd in lagen, waarbij opeenvolgende lagen volledig zijn verbonden. Door voor elke neuron de bias en de gewichten op de ingangen te variëren, kan een enorm bereik aan netwerkfuncties worden verkregen.
Als we een continue afhankelijke variabele y willen schatten, dan is het gebruikelijk om de covariaten x_i als input mee te geven en de schatting van y de output van het netwerk te laten zijn. Met een gekozen metriek wordt de fout in de schatting van y geminimaliseerd, waardoor punt-schattingen ontstaan zonder stochastiek. De crux in onze benadering is om ook variabele y als input in het netwerk op te nemen en de voor-waardelijke cumulatieve verdelingsfunctie $F(y|x_i)$ de output van het netwerk te laten zijn. In plaats van een fout in y te minimaliseren kiezen we er nu voor om de likelihood van de observaties te maximaliseren [1].



Figuur 1: voorbeeld netwerk continue verdeling

In figuur 1 wordt dit inzichtelijk gemaakt, met daarbij de volgende toelichting:
- De laatste laag bevat één neuron met een logistische activatiefunctie, die waarden tussen 0 en 1 aanneemt. Voorgaande lagen bevatten uitsluitend neuronen met de tanh activatiefunctie. De tanh functie is een lineaire transformatie van de logistische functie en is vanwege symmetrie aantrekkelijk;
- De eerste lagen bevatten alleen relaties tussen de covariaten, daarna wordt variabele y ingevoegd;
- De gestreepte verbindingen (na invoeging van variabele y) hebben verplicht een niet-negatief gewicht, waardoor F niet-dalend is in variabele y voor alle waarden van de covariaten;
- De geschatte functie F heeft geen linkerlimiet 0 en rechterlimiet 1. Bij maximaliseren van de likelihood ontstaat dit echter bij benadering automatisch;
- Door de gekozen activatiefuncties is de uiteindelijke netwerkfunctie differentieerbaar naar variabele y en kunnen waarden van de dichtheidsfunctie f worden gevonden.

De variant voor categorische afhankelijke variabelen wordt in figuur 2 getoond. De σ-neuronen in de laatste laag zijn hier als sequentiële kansen gedefinieerd, waardoor ze afzonderlijk van elkaar elke waarde tussen 0 en 1 kunnen aannemen. Het aantal σ-neuronen is altijd één minder dan het aantal categorieën.



Figuur 2: voorbeeld netwerk categorische verdeling

MAXIMUM LIKELIHOOD EN OPTIMALISATIE

Met de gradiënt methode en een geschikt algoritme, bijvoorbeeld Adam [2], worden alle gewichten en biases zodanig ingesteld dat de likelihood van de observaties wordt gemaximaliseerd. Voor het continue model moet daartoe eerst naar variabele y worden gedifferentieerd om de dichtheidsfunctie f te krijgen, de gradiënt ontstaat na een tweede keer differentiëren [3].

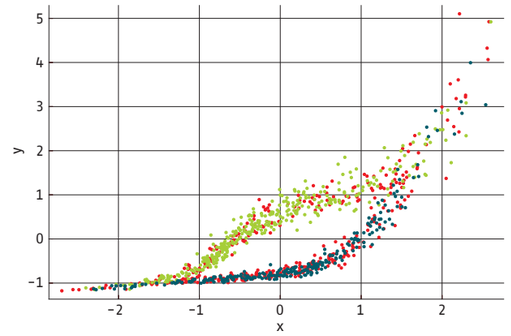
Bij een niet-triviale netwerkgrootte bestaat in het algemeen geen wiskundig maximum van de likelihood. Onrealistisch veronderstelde smalle 'spikes' kunnen de likelihood namelijk opstuwen tot on-begrensde hoogte. Met het instellen van een begrenzing van toe-gestane parameter-waarden is een theoretisch maximum wel gegarandeerd. Er is echter geen garantie dat het algoritme leidt tot het globale maximum. Door het starten met een eenvoudig netwerk en dat geleidelijk uit te breiden met extra (lagen) neuronen, kunnen we de ontwikkeling van de geschatte maximum likelihood en daarmee de kwaliteit van de optimalisatie volgen. Via een informatiecriterium zoals het Akaike Information Criterion (AIC) wordt vervolgens een optimaal model gekozen, waarmee underfitting en overfitting wordt voorkomen. Het hierin te gebruiken aantal model parameters is de som van het aantal gewichten en biases.

IMPLEMENTATIE

Het formularium inclusief een nieuw gradiënt algoritme is gerealiseerd in een eigen script in de programmeertaal Julia [4]. In vergelijking met de populaire talen R en Python is Julia zeer snel, hetgeen bruikbaar is bij de vereiste intensieve berekeningen. Uiteraard draagt ook de keuze van een maatwerk script positief bij aan de performance.

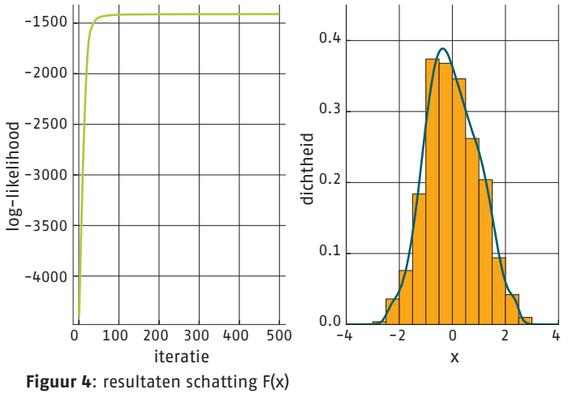
VOORBEELD

Als voorbeeld nemen we een kunstmatig geconstrueerde dataset met drie variabelen: x en y zijn continue variabelen en z is een categorische kleur (rood, groen, blauw). De set bestaat uit 1000 onafhankelijke observaties, zie figuur 3.



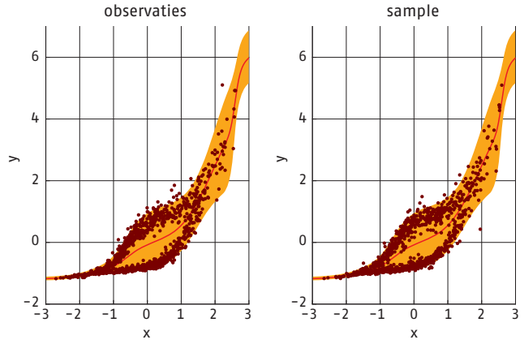
Figuur 3: scatter plot van de voorbeeld observaties

We kiezen ervoor om alle variabelen stochastisch te modelleren. Daartoe schatten we achtereenvolgens F(x), F(y|x) en F(z|x,y) met neurale netwerken. Een optimaal model voor F(x) ontstaat bij één tussenlaag met twee tanh-neuronen, met in totaal zeven parameters. Figuur 4 toont de ontwikkeling van de log-likelihood en de grafiek van de uiteindelijke dichtheid.



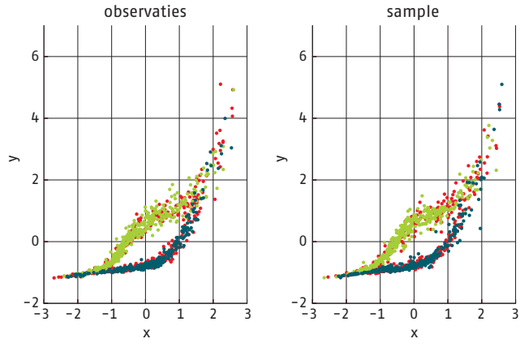
Figuur 4: resultaten schatting F(x)

Het optimale model voor F(y|x) is een netwerk met vier tussenlagen van elk drie tanh-neuronen. De variabele y wordt ingevoegd na de tweede tussenlaag. Dit netwerk kent 49 parameters. In figuur 5 worden links de geschatte verwachtingen en 95% betrouwbaarheidsintervallen getoond als functie van variabele x. Deze waarden zijn afgeleid door gebruik te maken van de inverse van F(y|x) via een binair-zoeken algoritme. Rechts zien we een gegenereerd sample van omvang 1000 uit het geschatte model, waaruit het goede schattingsresultaat blijkt.



Figuur 5: resultaten schatting F(y|x)

De categorische verdeling van model F(z|x,y) is geschat met twee tussenlagen met elk twee neuronen, in totaal 18 parameters. In figuur 6 wordt een model sample weergegeven naast de observaties. We zien hier een eindresultaat dat praktisch ondenkbaar is met GLM.



Figuur 6: resultaten schatting F(z|x,y)

CONCLUSIE

Het geschetste model geeft een ML-toepassing die raakvlakken heeft met GLM, maar vanwege de generieke formuleontwikkeling voordelen biedt in analyses met complexe verbanden. Of in situaties waar een korte ontwikkeltijd belangrijk is.

Bovendien hebben we hiermee voor data met slechts één variabele een techniek beschikbaar voor het schatten van een continue verdeling die niet past in een rijtje van bekende kandidaten.

Omdat het uiteindelijk resulterende neurale netwerk geen inzichtelijke formule levert, in tegenstelling tot GLM, is het noodzakelijk dat gebruikers simulaties uitvoeren in het geschatte model. Een voordeel hiervan is dat deze werkwijze een stochastische invalshoek bevordert en geschikt is om scenario's te construeren.

Tenslotte kan dit model waardevol zijn als ML benchmark, vanwege de brede toepasbaarheid en de stochastische basis. ■

Referenties
1. Chilinski, Pawel and Silva, Ricardo, Neural Likelihoods via Cumulative Distribution Functions, arXiv preprint arXiv:1811.00974 (2020).
2. Kingma, Diederik P and Ba, Jimmy Lei, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
3. Cardaliaguet, P and Euvrard, G, Approximation of a function and its derivative with a neural network, Neural Networks Vol. 5 (1992), 207–220.
4. Bezanson, Jeff and Edelman, Alan and Karpinski, Stefan and Shah, Viral B, Julia: A fresh approach to numerical computing, SIAM Rev. 59-1 (2017), 65-98.

Ir. A.W. van der Scheer AAG is zelfstandig actuaris bij Perunum Actuarieel Advies.