Samenvatting

Betere voorspellingen uit verkeerde en ondergespecificeerde modellen

Statistiek en machine learning behandelen de vraag wat we te weten kunnen komen over een onbekend proces, door te kijken naar de *data* (gegevens) die door het proces zijn voortgebracht. Zo wil een onderzoeker bijvoorbeeld iets te weten komen over de werking van een nieuw medicijn, op basis van gegevens over patiënten die dit medicijn eerder gebruikten. Met zulke kennis kunnen we dan *voorspellen* hoe het proces zich in de toekomst zal gedragen.

Voor de analyse van data worden vaak statistische *modellen* gebruikt. Een model is een verzameling *hypothesen*: mogelijke beschrijvingen van het onbekende proces. In veel realistische situaties is ieder beschikbaar model echter:

- *verkeerd* — geen enkele beschrijving in het model klopt precies met de werkelijkheid; of
- *ondergespecificeerd* — de beschrijvingen zijn niet volledig.

Toch willen we van data leren en goede voorspellingen doen, óók als er geen beter model voorhanden is. In dit proefschrift presenteren we methoden waarmee dit gedaan kan worden.

In hoofdstuk 2 kijken we naar *Akaike's informatiecriterium (AIC)*, in 1973 geïntroduceerd door de Japanse statisticus Hirotugu Akaike. Dit is een methode voor modelselectie, waarbij met een eenvoudige formule wordt bepaald welk van een aantal modellen naar verwachting het best in staat zal zijn om nieuwe data te voorspellen. Deze methode wordt in de praktijk vaak toegepast op een bepaald type ondergespecificeerde modellen: de datapunten bestaan uit twee delen, x en y, maar de modellen beschrijven alleen hoe y zich gedraagt als x al bekend is, en niet waar x vandaan komt. Toegepast op zulke modellen, blijkt AIC echter niet precies te doen wat we zouden willen. De formule geeft dan namelijk niet een inschatting van de voorspelfout op een níéuw datapunt, maar op een datapunt dat slechts gedeeltelijk nieuw is: een oude x met een nieuwe y. In hoofdstuk 2 laten we zien dat het ook mogelijk is om de voorspelfout voor een volledig nieuw datapunt in te schatten, namelijk met de nieuwe methode XAIC. Ook zien we dat door XAIC te gebruiken in plaats van AIC, de kwaliteit van voorspellingen in veel gevallen aanzienlijk beter wordt.

Een andere manier om een voorspeltaak aan te pakken, is *Bayesiaanse statistiek*, vernoemd naar dominee Thomas Bayes, een Engelse wiskundige die

leefde van 1702 tot 1761. In deze tak van statistiek wordt het begrip kans niet alleen gebruikt voor toevallige gebeurtenissen (zoals de kans om zes te gooien met een dobbelsteen), maar ook om onze onzekerheid uit te drukken (zoals de kans dat een bepaald model correct is). Als we onze onzekerheid over een onbekend proces in een kansverdeling hebben uitgedrukt en daarna nieuwe data uit dat proces observeren, dan volgt uit de wetten van de kansrekening een nieuwe kansverdeling, die uitdrukt hoe onze onzekerheid over het onbekende proces is veranderd nu we de data gezien hebben. Op basis van deze kansverdeling kunnen we vervolgens voorspellingen doen.

Wat gebeurt er echter als geen van de modellen correct is? In veel gevallen blijkt deze methode dan nog steeds goede voorspellingen op te leveren, hoewel ook bekend is dat het in zulke gevallen mis kan gaan. In hoofdstukken 3 tot en met 5 zien we een voorbeeld van een situatie waar de modellen allemaal verkeerd zijn (hoewel het verschil op het eerste gezicht niet problematisch lijkt), en waar Bayesiaanse methoden tot heel slechte voorspellingen leiden.

Deze problemen treden niet op als we een lagere *leersnelheid* kiezen. Een lagere leersnelheid betekent dat we onze kansverdeling minder sterk aanpassen wanneer we nieuwe data zien. Het nadeel hiervan is, dat we minder efficiënt gebruikmaken van de data. Daarom willen we de grootste leersnelheid gebruiken die nog veilig is voor onze data. De juiste snelheid is echter moeilijk te bepalen: in onze experimenten blijkt, dat veel voor de hand liggende methoden hier niet in slagen. We zien echter dat onze nieuwe methode *SafeBayes* er wel in slaagt een goede leersnelheid te kiezen, en daardoor betrouwbare voorspellingen blijft doen.

De laatste drie hoofdstukken van dit proefschrift behandelen een andere voorspeltaak, namelijk een generalisatie van het *Monty Hall-probleem* (in het Nederlands ook bekend als het *driedeurenprobleem*). In het spelprogramma *Let's make a deal*, met presentator Monty Hall, krijgt de deelnemer de keus uit drie gesloten deuren. Achter één van deze deuren staat een dure auto, achter de twee andere staan (relatief) waardeloze geiten. Nadat de deelnemer een keuze gemaakt heeft, wordt hij onderbroken door Monty, die één van de andere twee deuren opent en een geit laat zien. Verrassend genoeg kan de deelnemer zijn kans op de auto vergroten door nu van deur te veranderen!

Een algemener probleem is het volgende. Een uitkomst wordt willekeurig getrokken; we kennen de kansverdeling waarmee dit gebeurt, maar krijgen de uitkomst nog niet te zien. Vervolgens ontvangen we nieuwe informatie, waardoor sommige mogelijke uitkomsten afvallen. We weten dat deze informatie klopt, maar we weten niet door wat voor mechanisme de informatie is gekozen (zoals de deelnemer in het Monty Hall-probleem niet weet op wat voor manier de presentator kiest welke deur hij openmaakt).

De Bayesiaanse manier om kansverdelingen bij te werken, vertelt ons niet hoe we met zulke informatie om moeten gaan. Daarom pakken we dit probleem op een andere manier aan. We willen goede voorspellingen over de ware uitkomst kunnen doen, wát het mechanisme ook is dat ons de informatie gaf. We willen het dus zelfs goed doen als het mechanisme ontworpen is om het ons zo moeilijk mogelijk te maken. Met andere woorden: we kijken naar de

worst case (het slechtste geval).

Om verschillende voorspelstrategieën te vergelijken, hebben we getallen nodig die beschrijven hoe 'ver' een voorspelling ernaast zat. Zo'n toekenning van getallen aan voorspellingen heet een *verliesfunctie*. Omdat er oneindig veel verschillende manieren zijn om die getallen toe te kennen, moeten we een keuze maken.

In hoofdstuk 6 bekijken we hoe worst-case optimale strategieën voor allerlei verliesfuncties eruit zien. In het algemeen blijkt dat een voorspelstrategie die volgens de ene verliesfunctie optimaal is, dit volgens andere verliesfuncties niet hoeft te zijn. Het belangrijkste theoretische resultaat in hoofdstuk 7 is, dat er ook situaties zijn waarin de optimale strategie níét op deze manier afhankelijk is van de keuze van een verliesfunctie. Dit geldt in twee gevallen: als door iedere boodschap die het mechanisme ons zou kunnen geven, precies twee van de uitkomsten niet afvallen (de boodschappen vormen dan een *graaf*); of als de boodschappen op een specifieke manier met elkaar samenhangen (ze vormen een *matroïde*). Het Monty Hall-probleem heeft deze eigenschappen, en daarom is de optimale strategie daar niet afhankelijk van de verliesfunctie.

De resultaten van hoofdstuk 6 beschrijven hoe worst-case optimale strategieën voor allerlei verliesfuncties eruit zien, maar niet hoe je ze kunt vinden. Dit probleem wordt gedeeltelijk opgelost in de rest van hoofdstuk 7 en in hoofdstuk 8. In het bijzonder worden in dit laatste hoofdstuk efficiënte *algoritmen* voor grafen en matroïden gegeven: precieze stappenplannen die geheel automatisch uitgevoerd kunnen worden en na een eindig aantal stappen een worst-case optimale strategie hebben berekend.