The Art of Intelligence

A PRACTICAL INTRODUCTION TO MACHINE LEARNING (FOR ORACLE PROFESSIONALS)

Machine Learning is een hype. Iedereen heeft het erover- en elke organisatie moet dat hoognodig gaan doen. Althans die indruk zou gemakkelijk kunnen ontstaan. In dit artikel kijken we iets nuchterder naar Machine Learning en het enthousiasme daaromheen. Wat is het, waar zou je het voor kunnen gebruiken, hoe pak je dat dan aan en doet Oracle ook iets op dit gebied – dat zijn ruwweg de onderwerpen.

Lucas Jellema is Oracle ACE Director en werkzaam als Chief Technology Officer bij AMIS.

Wat en Waarom?

Machine Learning (ML) is een deelgebied van Artificial Intelligence, dat zich bezighoudt met zelflerende computersystemen. Computersystemen die antwoorden kunnen geven die niet volgen uit regeltjes die programmeurs hebben ingebouwd, maar die op basis van ervaringen uit het verleden voorspellingen kunnen doen over de toekomst. Of meer concreet: op basis van historische data hebben deze systemen een model opgesteld van een verband tussen input van bepaalde (predictor) attributen en de (target) output waarin we geïnteresseerd zijn. Worden er nieuwe data aangeboden aan dit model, dan krijgen we een inschatting van de target waarde.

Bekende toepassingen van ML zijn fraudedetectie (vaststellen of een geval in de categorie grote kans op fraude wordt geclassificeerd), e-mailspamfiltering, persoonlijke aanbevelingen, spraakherkenning en OCR (optical character recognition), predictive maintenance (op basis van beschikbare kentallen van het huidige gedrag van een component kan het model - dat is afgeleid van historische waarden van component-kentallen en het moment waarop ze defect raakten – inschatten of de component in de gevarenzone komt), chatbot, autocorrectie, schaken en Go spelen.

Machine learning is niet nieuw. De basisconcepten stammen van decennia geleden en de eerste Science Fiction over Artificial Intelligence is al ruimschoots achterhaald. Vanwaar dan nu de hype-achtige belangstelling? Daar zijn verschillende oorzaken voor aan te wijzen:

ML is veel bereikbaarder geworden vanwege betaalbare (veelal open source) en begrijpelijke en eenvoudiger toe te passen technologie (zelfs door de citizen data scientist)

De hardware resources die nodig zijn voor het toepassen van data mining en machine learning op grote (Big Data) gegevensets zijn haalbaar voor een veel groter publiek, dankzij technologieën als Hadoop en Spark en de prijzenslag rond IaaS.

De beschikbaarheid van data om machine learning op toe te passen is enorm toegenomen. Bronnen als Internet of Things en Social Media zorgen in veel organisaties voor onontgonnen datastromen. Daarnaast worden door veel organisaties in het overheidsdomein en de wereld van wetenschap en ook het bedrijfsleven, datasets gepubliceerd die kunnen worden benut voor data verkenningen en het opbouwen van ML modellen.

Verwachtingen en eisen van burgers en consumenten naast en druk van concurrenten zetten veel organisaties in beweging. Een smartphone moet spraak herkennen en foto's kunnen rubriceren. Negatieve sentimenten onder klanten op Twitter, Instagram of Facebook moeten onderkend worden. Cross-selling met op maat gemaakte suggesties en aanbiedingen is de voornaamste winstgenerator voor webshops. Analyse van terroristische dreiging, ontwikkelingen in het verkeer, rond virussen en het klimaat is noodzakelijk om tijdig maatregelen te kunnen overwegen.

Kortom: bijna iedereen kan het en vrijwel iedereen heeft een dringende reden om machine learning in te zetten.

Hoe en Waarmee?

Het inzetten van machine learning verloopt anders dan de gebruikelijke ontwikkeling van software: voordat modellen worden ingehaakt in web shops, klantservice systemen en IoT-applicaties ten behoeve van predictive maintenance wordt de data science workflow doorlopen. Hierin onderkennen we de volgende fasen:

- **Bepaal het doel van het project** – of tenminste de scope; welke vragen zijn het meest prangend, welke vermoedens willen we bevestigd krijgen, wat zouden we in de toekomst graag willen kunnen?
- **Verzamel data** –uit allerlei bronnen en in allerlei vormen.
- **Prepareer data-** voor verkenning en learning – door opschonen, filteren, omvormen, verrijken, aggregeren en combineren (de termen wrangle wordt voor een deel van deze activiteiten gebruikt).
- **Verken data** –op zoek naar trends, verbanden, mogelijk interessant patronen; bepaal welke attributen betekenisvol zijn – en welke niets lijken toe te voegen; bepaal datatypes, waardebereiken en categorieën, tijd-afhankelijkheden. Deze verkenning kan – evenals eventueel andere stappen in het proces – in eerste instantie heel goed op een sample van de data worden uitgevoerd – een subset die nog in het geheugen past. Als een goed beeld is ontstaan van een gewenste richting zou de volledige dataset verwerkt kunnen worden, bijvoorbeeld op een gedistribueerd storage en job execution platform zoals Hadoop.
- **Modelleer data**- in deze fase wordt geprobeerd op basis van de beschikbare data en de doelstellingen te komen tot een model dat waardevolle inzichten geeft in al beschikbare data (bijvoorbeeld correlaties tussen attributen en segmentering in clusters) of zinvolle voorspellingen kan doen op basis van nieuwe data. Hiertoe wordt een ogenschijnlijk passend ML algoritme geselecteerd en geconfigureerd. Vervolgens wordt een deel van de beschikbare data door dit model verwerkt: training van het model. Als dit resulteert in een model met enige betrouwbaarheid, wordt voor de rest van de data – waarvan we al weten wat de uitkomst zou moeten zijn - de voorspelling gedaan van de target-waarde. De betrouwbaarheid van het model wordt vervolgens bepaald door de voorspellingen te vergelijken met de vooraf bekende waarden. Afhankelijk van deze evaluatie kunnen we besluiten het model bij te stellen – met andere configuratie instellingen, aanvullende data of zelfs een compleet ander algoritme.
- **Presenteer bevindingen**, doe aanbevelingen en beslis over vervolgstappen.
- **Toepassen**- als een betrouwbaar model is gevonden waarmee relevante voorspellingen kunnen worden gedaan uit beschikbare data met acceptabele verwerkingstijd, dan zou dat model – of de uitkomsten van dat model – kunnen worden opgenomen in het run-time applicatielandschap.

Op dit moment start het reguliere software ontwikkelproces. Het is goed mogelijk dat voor het afleiden van het model een enorm Big Data Lake is ingezet, terwijl voor het toepassen van het model slechts heel bescheiden resources nodig zijn. Ook is het niet ongebruikelijk dat in deze fase de technologie van de datas cientist wordt vervangen door de tools van de software ontwikkelaar – en bijvoorbeeld het machine learning model wordt geconverteerd van Python naar Java on Spark.

Machine learning is gebaseerd op wiskunde – met name statistiek. Voor de stappen van het bovenstaande proces kunnen allerlei tools worden ingezet waarin deze wiskunde is ingekapseld, van Excel tot IBM Watson. Hieronder een - niet uitputtend - overzicht van populaire tools voor data science:

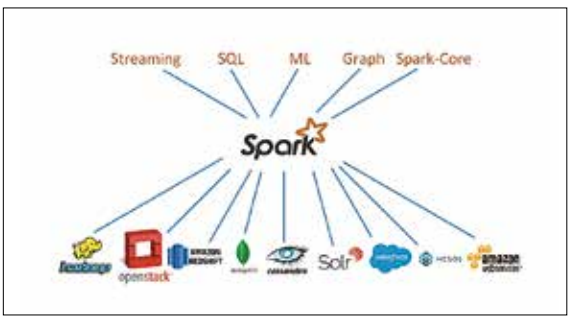
Veel (citizen) data scientists maken gebruik van Python – een toegankelijke programmeertaal met zeer krachtige libraries voor data wrangling, statistiek en machine learning en integratie met andere technologieën en platformen zoals R (een taal en omgeving voor statistische computerbewerking), Apache Spark, Fortran, C en SQL.

Een gebruikelijke, leuke en laagdrempelige manier om met Python en data science aan de slag te gaan is met behulp van Jupyter Notebooks. Een notebook is als een laboratorium-journaal waarin je verslag doet van al je hypotheses, experimenten en bevinden. In een Jupyter

»

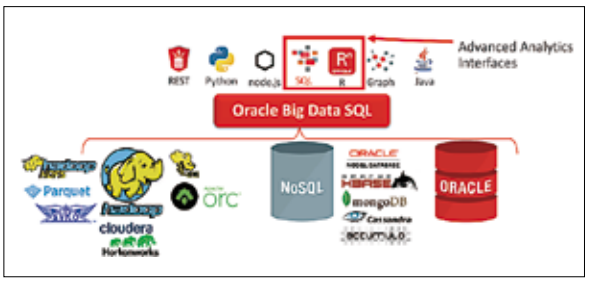
Notebook – een browser based Wiki-achtige pagina – kun je code opnemen en uitvoeren en stap voor stap je data verzamelen, bewerken, verkennen, visualiseren, en uiteindelijk modelleren. Een voorbeeld van een heel eenvoudig notebook op basis van Python en een eenvoudige CSV file is te vinden op: http://tinyurl.com/simple-linear.

Als een hele grote dataset moet worden gepreppareerd en ten behoeve van machine learning moet worden verwerkt, maken we graag gebruik van een gedistribueerd platform waar data over vele nodes is opgeslagen en waar de bewerkingen op die nodes kan plaatsvinden, zodat we netwerkverkeer beperken en optimaal van parallellisatie gebruik kunnen maken. Dat platform is in veel gevallen Hadoop en om niet de low-level map-reduce operaties van Hadoop te hoeven toepassen is met Apache Spark een framework beschikbaar om veel productiever gedistribueerde taken te ontwikkelen en te laten uitvoeren. Apache Spark kan onder meer SQL-statements uitvoeren tegen gedistribueerde Big Data-sets en heeft een library voor machine learning waarin enkele tientallen algoritmes beschikbaar zijn die de data op het Hadoop cluster kunnen verwerken.



Oracle en The Art of Intelligence

De stap van data naar informatie en inzicht werd door Oracle vanaf tweede helft jaren ’90 gezet. Met onder meer Oracle Text, Analytische SQL functies, de Darwin Data Mining engine, en later Enterprise R, honderden statistische functies in standard edition en ook ondersteuning voor spatial en semantische queries is de Oracle Database een platform waarmee geavanceerde analyse en ook machine learning kan worden gedaan. De echte ML algoritmes zijn onderdeel van de Advanced Analytics Database Option die bovenop de Enterprise Edition moet worden aangeschaft.
Met Big Data SQL heeft Oracle de rol van de database als platform voor advanced analytics nog verder versterkt: via de bekende database interfaces (SQL, PL/SQL en Enterprise R) kunnen queries, statistische bewerkingen en ook data mining en machine learning worden uitgevoerd tegen data in de database zelf, in NoSQL databases en op een Hadoop gedistribueerd systeem. Data scientists en tools die tegen Oracle Database kunnen praten kunnen daarmee op transparante wijze ook data in aanvullende databronnen ontsluiten.



In de Oracle Public Cloud zijn data science en machine learning op allerlei wijzen ondersteund. Machine learning algoritmes wordt toegepast in diverse SaaS producten – zoals Log Analytics in Oracle Management Cloud. De Big Data Discovery CS is voor verkenning van data (gebaseerd op Endeca) en Big Data Preparation CS is voor data wrangling, uitmondend in Big Data Cloud Service waar een Hadoop & Spark cluster het zware werk doet. Oracle heeft een Machine Learning CS aangekondigd tijdens Oracle OpenWorld 2016. Deze service zal een notebook –achtige interface bieden van waaruit Oracle specifieke en open source technologie- en ontsloten en geïntegreerd kunnen worden. Meer nieuws over deze service rond OOW2017.

Conclusie

Machine Learning is geen zwarte magie, al voelt dat misschien wel eens zo. Krachtige wiskundige modellen, het resultaat van decennia aan wetenschappelijk onderzoek, kunnen allerlei samenhangen in grote datasets aan het licht brengen, en die toepassen voor classificeren, beslissen en voorspellen naar aanleiding van nieuwe gegevens. Machine learning is dankzij snelle ontwikkelingen in software en hardware binnen handbereik gekomen van ons allemaal. Het zetten van eerste stappen met machine learning is dankzij deze technologie, gratis VM’s en free-tier cloud services en ook met behulp van gratis cursussen van hoge kwaliteit bij bijvoorbeeld Udacity en edX erg eenvoudig geworden. ■