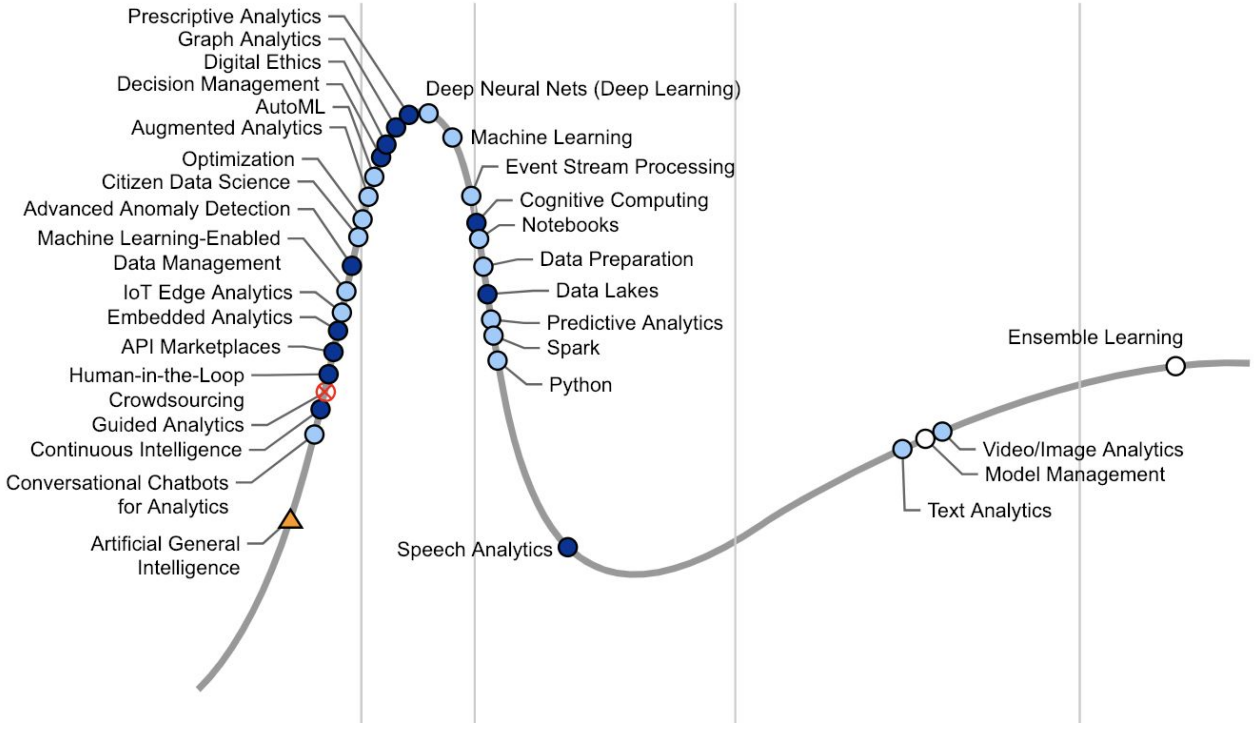
Lesbrief: Machine Learning



Gartner, Hype Cycle for Data Science and Machine Learning, 2018

Machine Learning is net over de top van de Hype Cycle, en zal naar verwachting binnen 2 tot 5 jaar volledig 'normaal' zijn. Het is dus goed om hier wat meer over te weten en te zien dat veel van de magie gewoon te verklaren is.

Worden de voorspellingen van Netflix echt beter door het werk van machines? (https://en.wikipedia.org/wiki/Netflix_Prize). Stuurt Facebook straks de hulpverleners bij je langs, voordat je zelf weet dat dat nodig is? (https://www.forbes.com/sites/bernardmarr/2017/03/02/facebook-uses-machine-learning-to-spot-suicidal-users/#ab6d9413b2b6)
Weet de computer wat voor weer het morgen wordt (https://www.ibm.com/developerworks/community/blogs/jfp/entry/Hindsight?lang=en) of kan hij zelfs vertellen dat je een korte broek aan mag? (http://www.magikeenkortebroekaan.nl/)

Praten we straks tegen computers (https://www.pullstring.com/blog/hybrid-ai-and-machine-learning-letting-computers-talk-back) of praten ze straks alleen nog met elkaar? (http://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html)

Ga op onderzoek naar Machine Learning. Onderzoek de mogelijkheden, maar ook de onmogelijkheden hiermee. En probeer zelf diverse technieken uit.

Uitzoeken

- Wat machine learning is / wat kan wel, wat kan niet
- Wat het verschil is tussen Supervised learning en Unsupervised learning
- Hoe k-means werkt
- Hoe logistic regression en decision trees werken
- Wat Over- en underfitting is
- Hoe je je resultaten valideert (verschil test en training, en hoe vaak je beiden mag gebruiken) en accuratie meet

Opdracht

Demonstreer:

- clustering (kmeans)
  Haal data op van
  https://programmeren9.cmgt.hr.nl:9000/{studentnummer}/clustering/training
  - Plot de data
  - Doe een 'gok' voor het aantal clusters, print de plot en teken waar je denkt dat de clusters en centroids komen
  - Voer kmeans uit voor verschillende k's (jouw gok, +1 en -1) en plot deze data
  - Plot ook de centroids

- classification (logistic regression en decision tree)
  Haal data op van https://programmeren9.cmgt.hr.nl:9000/{studentnummer}/classification/training
  - Plot de data
  - Train de data
  - Geef de trainingsdata nu aan je classifier om te zien hoe goed hij deze geleerd heeft. Wat is de accuratie van de twee algoritmes (op de trainingsdata)
    Gebruik hiervoor de sklearn functie accuracy_score()
  - Plot de resultaten voor beide classifiers
  - Welke van de twee zou je op basis hiervan kiezen

  - Valideer nu m.b.v. de testset en vergelijk de accuracy op test met de accuracy op training
    Haal data en post result op:
    https://programmeren9.cmgt.hr.nl:9000/{studentnummer}/classification/test
  - Wat is de accuratie van de twee algoritmes (op de test data)

  - Vergelijk de accuratie op de trainings- en testdata. Welke van de twee algoritmes zou je nu kiezen? Leg uit waarom.

NB. Voor Python is er startercode die het ophalen en posten van data voor je afhandelt. Wil je zelf (met een andere programmeertaal) aan de slag kijk dan naar de documentatie onderaan dit document.

Leerdoelen

- Ik kan het basisprincipe achter machine learning uitleggen en toepassen
- Ik ken het verschil tussen supervised en unsupervised learning, en kan uitleggen wanneer je welke van de twee toe kunt toepassen.
- Ik begrijp de werking van kmeans en logistic regression en decision trees en kan deze toepassen
- Ik weet welke problemen kunnen ontstaan bij het leren (over/under-fitting), kan maatregelen hiertegen nemen, en testen of ik dit goed gedaan heb

Toetsing

Om deze lesbrief te behalen moet je de volledig uitgewerkte opdracht kunnen demonstreren (zie boven) en moet je vragen kunnen beantwoorden over het uitzoek-werk (zie boven).

Tools

- sklearn
- tensorflow
- watson

Startcode

- https://github.com/HR-CMGT/PRG09-ML

Bronnen

Naast de handleidingen van bovengenoemde tools zijn er bronnen in overvloed over machine learning.

Een toegankelijke gratis cursus die ik adviseer is: Intro to Machine Learning (Udacity)
https://www.udacity.com/course/intro-to-machine-learning--ud120
Waarin Python gebruikt wordt met de sklearn toolkit.

Documentatie

clustering

https://programmeren9.cmgt.hr.nl:9000/{studentnummer}/clustering/training
{studentnummer} : Je eigen studentnummer (iedereen krijgt eigen data)

GET
Geeft lijst met 100 2D-datapunten

Default csv, stuur header Accept: application/json voor json-data

classificatie

https://programmeren9.cmgt.hr.nl:9000/{studentnummer}/classification/training
{studentnummer} : Je eigen studentnummer (iedereen krijgt eigen data)

GET
Geeft lijst met 1000 2D-datapunten en classificatie (0 of 1)
Default csv, stuur header Accept: application/json voor json-data

https://programmeren9.cmgt.hr.nl:9000/{studentnummer}/classification/test
{studentnummer} : Je eigen studentnummer (iedereen krijgt eigen data)

GET
Geeft lijst met 100 2D-datapunten zonder classificatie voor test (elke dag anders)
POST
Geeft een score aan jouw classificatie. Je stuurt hierheen een lijst met classificaties (0 of 1) voor de test-data van die dag in json-formaat.