**Special Interest Group Machine Learning**

Eerste meeting – 2019-06-24

Aanwezig:

* Rob Verseijden
* Rob van Ginkel
* Frits Keereweer
* Mischa de Blaauw
* Manuel Osses

Afwezig:

* Peter van Loghem

**Verwachtingen**

De verwachtingen van iedereen worden gepeild. De algemene conclusie is dat men graag meer wil weten van Machine Learning en dit vakgebied binnen de Artificial Intelligence ook gezien wordt als het best aansluitend op eventuele inzet binnen het vakgebied van Magion. Men wil er graag van leren en zou het liefst zien dat we hier ook daadwerkelijk iets mee kunnen doen binnen Magion.

**Presentatie**

Pascal laat de presentatie Machine Learning zien, deze zal meegestuurd worden met dit verslag. Door de animaties erin is deze helaas alleen goed zichtbaar binnen PowerPoint en niet als PDF.

**Talen en tools**

Als tools worden aangeraden Matlab of Octave, R (via R-Studio) en Python:

Octave is gratis beschikbaar en wordt net als Matlab ook gebruikt in de academische wereld, het is beschikbaar voor de meest gangbare OSen en te vinden op:  
<https://www.gnu.org/software/octave/>

De taal R is een beetje de gulden middenweg: deze heeft wat weg van Matlab, maar is net zoals Python voorzien van vele libaries waaruit allerlei leuke dingen zijn te halen. R is vooral handig voor Feature Engineering en heeft daarvoor heel veel handige functies in het standaardpakket.

Python is een wat bekendere taal is ook makkelijk te leren. Het biedt veel vrijheden en door de enorme hoeveelheid libraries is ook hier veel mee te bereiken.

Octave wordt in de Coursera tutorial gebruikt (zie verderop).

Python en R komen in heel veel tutorials voor, zoals Udacity, Eduonix etc…

Om Python en R te gebruiken is het gebruik van Anaconda aan te raden. Hier wordt Python (kies voor versie 3.7) en R-Studio standaard bij meegeleverd. Doordat je diverse environments kunt aanmaken heb je ook altijd beschikking tot de juiste (versie-)mix van libraries. Voor Python gebruikje Jupyter Notebook, waar je je code mengt met aantekeningen en uitkomsten zoals grafieken. Anaconda is beschikbaar voor de meest gangbare OSen en is hier te vinden:  
<https://www.anaconda.com/distribution/>

Note: voor R-studio zijn de libraries vaak alleen voor Linux, dus het kan handig zijn daar een versie van achter de hand te hebben (misschien op een Raspberry Pi 4?).

Verder heeft Frits ervaring met Power BI en zal hier ook een en ander van laten zien tijdens de volgende skills & trends. Power BI kan tegenwoordig zowel tegen R als tegen Python aanpraten.

Power BI is niet gratis, maar binnen Magion zou er wel wat te regelen moeten zijn voor die 8 a 9 EURO per maand.

**Cursussen**

Wat links om de grondbeginselen van statistieken weer even op te frissen:

<https://www.udacity.com/course/intro-to-descriptive-statistics--ud827>

<https://www.udacity.com/course/intro-to-inferential-statistics--ud201>

Als je nog niet heel erg bekend bent met Python, is het aan te raden ook deze cursus te volgen:

<https://www.udacity.com/course/intro-to-computer-science--cs101>

Dan de Machine Learning zelf:

- <https://www.udacity.com/course/intro-to-machine-learning--ud120>

Deze cursus is redelijk beknopt en geeft met behulp van Python een goed overzicht van ML

- <https://www.coursera.org/learn/machine-learning/home/welcome>

Deze cursus is vrij intens, maar geeft je daardoor wel goed inzicht hoe en waarom je met het tunen van parameters veel meer uit je systeem kunt halen. De cursus is gebaseerd op Matlab/Octave.

- <http://www-bcf.usc.edu/~gareth/ISL/>

Een boek, online te downloaden. Ook hier weer een goed overzicht in Machine Learning, deze keer echter met R.

Deze heb ik nog niet bekeken, maar zou zomaar goed kunnen zijn (Samenwerking tussen Google en Kaggle):  
<https://developers.google.com/machine-learning/crash-course/>

**Datasets**

Om een start te kunnen maken en te kunnen oefenen zullen we 1 of 2 datasets kiezen die aangeboden worden op het internet, waar we ook naar voorbeelden kunnen kijken van andere inzendingen. Er zijn ook enkele (2) datasets beschikbaar op het internet met betrekking tot ons vakgebied en ook binnen Magion hebben we mogelijk de beschikking over nog meer data, maar we willen graag eerst onze resultaten met anderen vergelijken.

We zouden naar een dataset kunnen zoeken waar al redelijk veel mensen aan gewerkt hebben. Het mooiste zou zijn iets met time-series, omdat we daar in praktijk natuurlijk ook mee te maken hebben. Dan zouden we een voorbeeld van regressie (een continue waarde voorspellen) en een van categorisatie (bijvoorbeeld een trip voorspellen).

Sites waar veel datasets te vinden zijn:  
<https://www.kaggle.com/competitions>  
[https://www.topcoder.com/challenges?filter[tracks][data\_science]=true](https://www.topcoder.com/challenges?filter%5btracks%5d%5bdata_science%5d=true)  
<https://www.crowdanalytix.com/>  
<http://tunedit.org/challenges>  
<https://www.innocentive.com/ar/challenge/browse>

Enkele datasets die richitng ons werkgebied komen (maar helaas met geen concurrentie/voorbeelden):  
<https://www.kaggle.com/c/vsb-power-line-fault-detection/data>  
<https://archive.ics.uci.edu/ml/datasets/Condition+monitoring+of+hydraulic+systems>  
<http://archive.ics.uci.edu/ml/datasets/Gas+sensor+array+temperature+modulation>

Antwoorden achteraf op vragen tijdens de meeting:

Waar komt F-score vandaan? De F is blijkbaar geen afkorting, zie:  
<https://en.wikipedia.org/wiki/F1_score>

Waarom is er ook nog een validation set, naast train en test?  
Dit hangt af van de validatiemethode, zie bijgevoegde PDF.

