Knowlegde discovery and data mining

Veel technologieën om data vast te leggen.
Veel technologieën om data op te slaan.

Data mining helpt om kennis te ontdekken.

Knowledge discovery

De wetenschap om informatie te halen uit grote data sets is knowledge discovery.

Gebruikt in marketing, fraude detectie, medische sector, risico analyse, beurzen...

Niet alle vragen zijn te beantwoorden met 1 SQL query.

Knowledge discovery in databases (KDD) = data mining

Data mining

Data mining is één van de stappen in het knowlegde discovery proces.
Data mining is een iteratief en interactief proces.

1. Goal indentification
   Begrijpen wat de klant wil
2. Creation of target data set
3. Cleansing/preprocessing
   Zeer belangrijk, maar saai
   Garbage in, garbage out
4. Data transformation
   Training data: om het model te bouwen (grootste deel van de data)
   Validatie data: om het model te evalueren (overfitting tegen te gaan)
   Test data: om het model te re-evalueren (deze data niet gebruiken in de vorige 2 stappen, anders wordt de fout van het model te laag ingeschat)
5. Choice of data mining task
   Voorspellen of beschrijven
6. Choice of data mining algoritm
7. Data mining
   Zoeken achter patronen
   Problemen:
   Is er overbodige data?
   Hebben we meer data nodig of een ander algoritme? Indien verkeerd kan men terug keren naar stap 6.
8. Interpretation and evaluation of the results
   Wordt door een mens geïnterpreteerd
9. Presentation of the information
10. Action
    Deze stap wordt typisch door andere mensen uitgevoerd dan de vorige stappen.

Voorbeelden:

- Classificatie: voorspellen tot welke categorie een item behoort
  - Spamfilter
  - Voorspellen tot welke familie bloem een bloem behoort
- Regressie:
  - Verkoopsprijs van een huis bepalen
  - Prijs van een aandeel voorspellen

Supervised

- Clustering
  - Verschillende groepen proberen te ontdekken (sociale netwerken, type klanten)

Unsupervised

- Summarisatie:
  - Voor elk veld gemiddelde en standaar deviatie bijhouden.
  - Word cloud
- Dependency modelling
  - kijken naar afhankelijkheden
  - proberen om afhankelijkheden te vinden tussen variabelen

Decision tree

Een beslissings boom neem een vector van attribuut waardes als input en retourneert een beslissing.

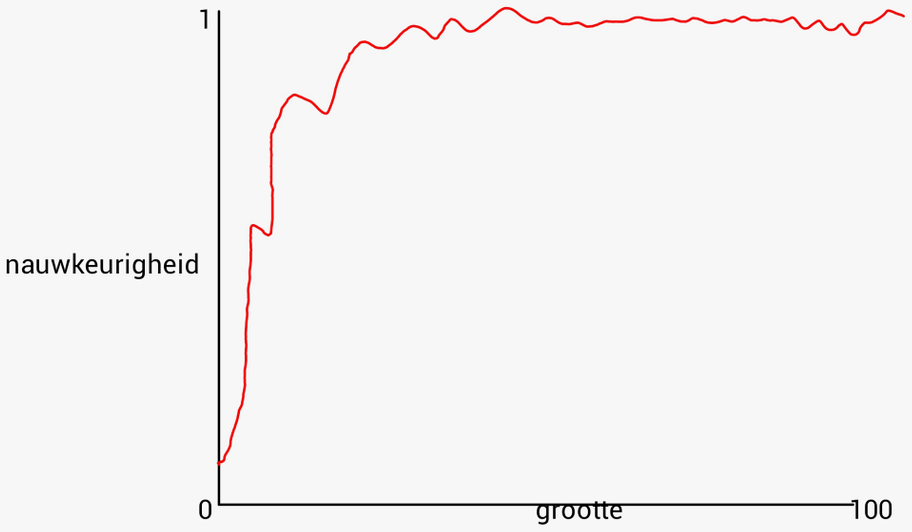
Ockhams razor: kiezen voor het eenvoudigste model dat de wereld verklaard.
- Test eerst op het meest belangrijke attribuut
  - De belangrijkheid van een attribuut wordt vaak berekend aan de hand van entropie.

Ockhams razor recursief toepassen om zo snel tot een beslissing te komen.

Hoe meer voorbeelden hoe beter de boom.

Learning algoritme accuracy

Learning curve (happy curve)



Als de grote van de training set groeit, dan verbetert het algoritme op de test sets.

Overfitting en generalisatie

Overfitting is het gekend groot probleem bij machine learning/ datamining
- Het leren focust te veel op de test data waardoor er de test data van buiten gekend is.
  - Goed scoren op test data, maar geen nieuwe output genereren bij nieuw voorbeeld.