Data Mining

= op een geautomatiseerde manier bruikbare patronen en relaties ontdekken in grote hoeveelheden gegevens (= grote datasets)

Toepassingen in:

- Medische wereld: kans op ziekte bepalen
- Economische wereld:
    - Wie zal ingaan op een aanbod?
    - Gemiddelde uitgave van die mensen?
    - Fraude, b.v. bij verzekeringen
- Informatica:
    - Classificatie spam – non spam mails
    - Welke datapakketjes vormen aanval?
- ...

Op het grensvlak van:

- statistiek,
    - lineaire en logistische regressie,...
- machine learning, artificial intelligence
    - neural networks, decision trees,...
- data base management

Gevolg: meervoudige terminologie!

- Variabele, kenmerk, attribuut, “field”
  → kolom in een dataset
- Observatie, “record”
  → rij in een dataset

- Output variabele, target variabele, afhankelijke variabele
- Input variabelen, predictor variabelen

Verschillen statistiek ↔ data mining

Statistiek:

- “kleine” hoeveelheid data (= steekproef)
- zelfde steekproef voor voorspelling en voor betrouwbaarheid van dat resultaat
- hypothesetoetsen, betrouwbaarheidsint.

  → moeilijk

  → veel beperkingen

Data mining

- “grote” hoeveelheid data
- een grote steekproef om model te “fitten”
- tweede steekproef voor performantie model

  → veel inzichtelijker

  → minder beperkingen
- **Gevaar! Overfitting !!** (zie verder)

⇒ Data mininig is statistics at scale, speed and simplicity!

Belangrijkste ‘domeinen’ data mining:

Classificiation

- Klasse voorspellen

- Op basis van data met gekende klasse
  → model ‘fitten’
- Toepassen op data met ongekende klasse
- Voorbeelden:
  - Spam – non spam
  - Aanval netwerk – geen aanval

Prediction

- Analoog als classificiation
- Geen klasse, maar waarde **continue** variabele voorspellen
- Voorbeelden:
  - Gemiddelde uitgave klant
  - Kijkcijfers voorspellen

Affinity Analysis – Association Rules

- Welke variabelen worden geassocieerd met elkaar?
- Voorbeelden:
  - Opstelling producten warenhuis
  - Amazon (Market Basket)

Clustering Analysis

- Groeperen van ‘gelijkaardige’ data
- Voorbeelden:
  - Marktsegmentatie-onderzoek
  - http://www.music-map.com

(Data Exploration – Data Visualization)

(Data Dimension Reduction)

Supervised ↔ Unsupervised learning

Supervised

- Classification en Prediction
- “Training set” waarbij waarde target variabele gekend is
  → gebruiken om model te “trainen”
  → model leert van data in “training set”
- Model daarna ‘tunen’
- Model gebruiken voor voorspelling target variabele bij nieuwe data
- Voorbeeld: lineaire regressie

Unsupervised

- Geen target variabele om te voorspellen
- Model kan niet leren van een “training set” met gekende target variabele
- Voorbeelden: clustering analysis

Stappen in het data mining proces

1. Doel van de analyse duidelijk omschrijven
2. Dataset voor de analyse ‘samenstellen’
    - Random sampling uit grote database
    - Gegevens uit databases samenbrengen
    - Interne en externe data
3. Exploreren, ‘opkuisen’, voorbereiden data
    - Ontbrekende data (missing values)?
    - Outliers?
    - Verbanden tussen variabelen?
    - …
4. Variabelen verwijderen, bijmaken, ‘transformeren’
5. Data verdelen in training, validatie, test set
6. Welke data mining taak?
    - Association, Prediction, Clustering?
7. Welk data mining techniek?
    - Regressie, Naïve Bayes, Neural net?

1. Data mining algoritmes implementeren en uitvoeren
2. Beste algoritme/techniek kiezen
3. Model in productie brengen / omzetten in beslissingsregels

In Sas: **SEMMA**-methodologie:

Sample
Explore
Modify

Model
Assess

Random Sampling

- data mining: veel variabelen en records
- beperkingen rekencapaciteit, software

→ sampling, kleinere dataset

Oversampelen van zeldzame gebeurtenissen

- vaak in classificatie: veel ‘0’, weinig ‘1’

→ eventueel random sample met weinig ‘1’

→ weinig informatie over records klasse ‘1’

→ moeilijk goed model te trainen voor classifcatie ‘0’ en ‘1’

- oplossing: klasse ‘1’ oversampelen
- kan belangrijk zijn:

→ een ‘1’ missen kan duur zijn! (cfr. aanval computernetwerk)

→ ‘0’ foutief als ‘1’ klasseren is minder erg!

Data voorbereiden

Continue variabele

→ eventueel discrete variabele

Discrete variabele met n klassen

→ n-1 binaire dummy variabelen

Outliers, uitschieters

- kan wijzen op fout gemeten/ingegeven
- kan groot effect hebben op het model

- eventueel verklaring vinden voor outlier
  → inbreng domeinexpert belangrijk!

Ontbrekende data (missing values)

- indien weinig records met ontbrekende data
  → records verwijderen
- **maar** stel 30 variabalen, voor elke record, voor elke variabele 5% kans missing value
  → kans missing value in gegeven record?
- eventueel vervangen door b.v. gemiddelde
  (= 'imputed value')
  → nadeel: geen nieuwe informatie voor die variabele
  → voordeel: informatie andere variabelen gaat niet verloren!

Data normaliseren

- records onderling vergelijken kan alleen op dezelfde schaal
- anders kan 1 variabele domineren
  → normaliseren
  → (waarde – gemiddelde)/standaardafw.
  → schaal: "aantal standaardafwijkingen van het gemiddelde"
- afhankelijk van techniek of dit nodig is

Hoeveel records nodig voor training model?

- v.b. Delmater en Hancock voor classificatie
  aantal records = minstens 6*M*N

M = aantal klassen
N = aantal variabelen
→ liefst aantal variabelen minimaliseren

Overfitting

- training data gebruiken om model te fitten
- probleem: training data = signaal + ruis
- gevaar: ruis modelleren i.p.v. signaal
  = **overfitting**
- altijd wel model dat past op training data
- maar: model moet niet alleen training data modelleren!
- gevolg: model niet toepasbaar op toekomstige data
- belangrijk: waar stop je met fitten?
- voorbeeld: lineaire regressie en regressie van hogere orde

- probleem kan ook ontstaan bij te veel 'predictor' variabelen
  → model beter te fitten bij meer var.
  → maar: eventueel variabelen in model opgenomen die niet belangrijk zijn bij het toepassen van model op toekomstige data

Training Set – Validation Set – Test Set

“Partitioning”

Typisch voor data mining (grote datasets!)

- Training set
  - gebruiken om model te trainen
  - model aanpassen/fitten aan de data
  - gevaar: overfitting

- Validation set
  - performantie gevonden model onderzoeken
  - gevonden model verfijnen/tunen
  - gevaar: opnieuw overfitting

- Test Set
  - niet altijd aanwezig
  - performantie beste modellen per techniek met elkaar vergelijken

Partitioning at random of volgens een variabele

Typische verhoudingen:

- training/validation (60/40)
- training/validaton/test (50/30/20)