**Data Representatie**

**Gegevens type**

1. **Categorisch**

Nominaal:

De data kunnen alleen worden gecategoriseerd (geen rangorde).

Ordinaal:

De data kunnen worden gecategoriseerd en gerangschikt.

1. **Numerieke**

Discrete data is countable while continuous — measurable.

* Discrete

Discrete data is the type of data that has clear spaces between values. You can count the data, but not measure it.

fi = absolute frequentie

ϕi = relatieve frequentie

cfi = cumulatieve absolute frequentie

cϕi = cumulatieve relatieve frequentie

The relative frequency is calculated by dividing the absolute frequency by the total number of values for the variable.

* Continu

Continuous data is data that falls in a constant sequence. It cannot be counted, but measured, like height or time

Methode:

- Zoek min en max data point

- Bereken het verschil (14.97)

- deel door 5 en 15 (uit 20) en kies klassenbreedte 'b' tussen uitkomsten:

- 0.998 <= b <= 2.994

Klassengrenzen:min en max grens. [15.89; 17.89[

Klassenbreedte: verschil tussen min en max

Klassenmidden: min + max / 2

Klassenfrequentie: the number of times the items corresponding to a class interval repeat in the series.

**Kengetallen**

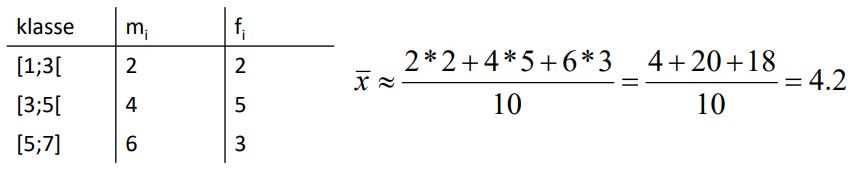
1. **Kengetallen voor locatie**

* Rekenkundig gemiddelde

Som/totaal aantal

Frequentietabellen: (hoeveelheid \* categorie) + (hoeveelheid \* categorie) +… / cf

Frequentietabel met klassenindeling:



Oplossing voor uitschieters: trimmed mean

* Gewogen rekenkundig gemiddelde

Def: de som van alle waarnemingen vermenigvuldigd met het juiste gewicht, gedeeld door de som van de gewichten.

* Mediaan
* De middelste waarde als n oneven is
* Het gemiddelde van de middelste twee waarden als n even is
* Kwartielen

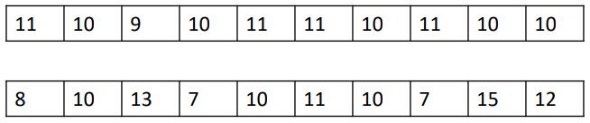
Een van de drie waarden die een dataset in vier gelijke delen opdeelt.

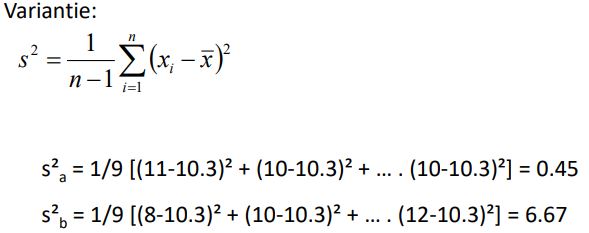
* Modus

De observatie met de hoogste frequentie

1. **Kengetallen voor spreiding**

* Variantie





X^- = gemiddelde

* Standaardafwijking

Sqrt(variantie)

* Spreidingsbreedte

Max - min

* Interkwartielafstand

Q3 – Q1

**Verbanden tussen variabelen**

Visuele representatie: Kruistabel

Geen verband => variabelen zijn onafhankelijk

Onafhankelijk als:

* De (relatieve) frequenties hebben in elke **kolom** dezelfde verhouding
* De (relatieve) frequenties hebben in elke **rij** dezelfde verhouding
* De relatieve frequentie in elke cel is gelijk aan het product van de bijbehorende rij/kolomtotalen

Er is een verband:

* Toename in een variabele 🡪 toename in andere variabele: **positief** correleren
* Toename in een variabele 🡪 Afname in andere variabele: **negatief** correleren

Verband tussen 1 numerieke en 1 categorische variabele: BoxPlot

Verband tussen 2 numerieke waarden: Scatter, bar, etc.

In ScatterPlots zou je of een lineair/non-lineair verband kunnen bespreken. Uitschieters en verstrengelende factoren dienen ook besproken te worden.

**Missing Values**

Reden?

Soorten:

* **MCAR**

Missing Completely At Random: De neiging waarom een data point niet aanwezig is, is totaal random. Er is geen verband tussen waarom de data niet aanwezig is, en andere variabelen.

* **MAR**

Missing at Random: De redenen waarom bepaalde data niet aanwezig is heeft wel te maken met een andere variabele. Voorbeeld: *“So for example if older people are more likely to skip survey question #13 than younger people, the missingness mechanism is based on age, a different variable.”*

**Outliers**

Oorzaak: Onjuiste waarde, of wel juiste waarde, maar dat is zeldzaam.

Visuele methode: Boxplot.

Missing Values/Outliers actie:

* Negeren
* Feature (kolom) verwijderen
* Observatie (rij) verwijderen
* Waarden imputeren
* Data transformeren (met binnen bijv.)

Imputeren: Het proces van data vervangen met gesubstitueerde waarden.

**Sampling/Steekproeven**

2 soorten:

* Selecte Steekproef

Niet iedereen maakt kans om terecht te komen in de steekproef: De resultaten gelden enkel voor de onderzochte groep.

* Aselecte Steekproef

Iedereen heeft even grote kans om terecht te komen in de steekproef. Completely at random.

**Dichtheidsfunctie**

“ideale” beschrijving van de volledige dataset.

Soorten dichtheidsfuncties:

* Symmetrische verdeling
* Rechts-scheve verdelingen en dichtheidsfuncties: Gemiddelde > Mediaan
* Links-scheve verdeling en dichtheidsfuncties: Gemiddeld < Mediaan

De dichtheidsfunctie van een normaal verdeelde stochastische variabele X heeft een typische klokvorm en wordt Gausscurve gemoed). De breedte en hoogte van de klok varieert afhankelijk van de waarde van de standaard afwijking.

**Machine Learning**

Soorten:

* Supervised Learning
* Inputvariabelen en outputvariabelen zijn aanwezig in de training data set
* Functie f die het verband geeft in y=f(…)
* Functie f voorspelt op basis van een nieuwe income de outcome y
* De functie f wordt continu bijgestuurd.

Types:

* Classificatie
* Regressie

Doel: om continue variabele te voorspellen

* Unsupervised Learning
* Enkel de inputvariabele is aanwezig
* Het algoritme brengt patronen, structuren, groepen, … naar voren die op voorhand niet opgenomen zijn in de dataset

Type:

* Clustering

Doel: items grouperen

* Reinforcement Learning
* Reward/Punishment

ML problemen herkennen:

* Rule based: je past een aantal statische regels toe op de huidige context( tijdstip, dag…)
* ML: Ook in de ML benadering pas je een aantal regels toe maar het verschil is dat deze regels automatisch geüpdatet worden op basis van nieuwe data.
* 3 stappen:
* Verzamel een grote dataset
* Gebruik een algoritme wat zelfstandig het verband vindt
* Update dit verband voortdurend met nieuwe data.

Het verschil tussen een klassieke aanpak vs. ML aanpak:

**Klassiek**  
Je bent een “alien” en kijkt naar de bevolking ergens op de wereld en je weet dat er 2 soorten mensen op de wereld zijn: zeg man en vrouw. Maar je kent de karakteristieken om te differentiëren tussen man en vrouw niet. Maar je kunt wel zien dat mensen die groter zijn dan bvb 1.8 tot de groep van de mannen behoren. De andere groep zijn dan de vrouwen. Je steekt data én een programma in een computer en je verkrijgt output; dit is een benadering van het probleem op basis van regels; op basis van programmatielogica.

**ML aanpak**

Die alien kidnapt jou en jij kan per persoon aanduiden of het een man of een vrouw is. Nadat die alien geobserveerd heeft hoe jij een 100- of 1000-tal mensen classificeert als man of vrouw (= dit noemen we de training data set); leert hij “intuïtief” hoe te differentiëren tussen geslachten.

Vuistregels ML:

* Het is moeilijk om het probleem te gieten in regels
* Je beschikt over een grote set historische data
* De patronen of relaties tussen de data zijn dynamisch

**ML problemen uitgewerkt:  
  
Classificatie**

Soorten:

* Binair classificatie
* Multi-class classificatie
* Multi-label classificatie
* Multi-output classificatie

6 verschillende classificatie algoritmes:

* Naive Bayes algoritme
* Support Vector Machine algoritme
* Decision Tree algoritme
* Logistic Regression algoritme
* Linear Discriminant Analysis
* Nearest Neighbor Algoritme

**Clustering**

Bestaande clustering algoritmes:

* K – mean clustering
* K – median clustering
* Hierarchical clustering
* Density-based clustering
* Distribution-based clustering

Presentatie:

Tot 95% nauwkeurigheid herkennen kan de mens, 5% erlangs

Mensen van andere etnische achtergrond zijn moeilijker te herkennen

Series: CSI/NCIS

Gezichtsdetectie: Gaan detecteren op een foto/video waar het gezicht zich bevindt

Gezichtsherkenning: Het gezicht gaan herkennen met een naam of dergelijke

Geschiedenis:

1960 begin eerste pogingen, Woodrow Bledsoe

1987: M.Kirby voor het eerst automatisch

2001: Paul Viola Michael Jones: features herkennen

2010: ImageNet Muffin/Chihuahua, Labradoodle Fried Chicken

2012: AlexNet: Neural Networks, Foutemarge van 16%

2015: Deepface, FaceNet, foutemarge < 5%

Toepassingen

China Social Credit Scores vs. Access apartments automatically or more safety at night?

China Publishes Draft Security Standard on Facial Recognition

Facebook – Tag

iPhone – Face Unlock

China – AliPay

Privacy:

Cameragebruik – Camerawetgeving, vooral o.t. bewakingscamera’s, alleen identificatie is toegelaten

Dataverwerking – Privacy wetgeving. Uitzonderingen:

1. Biometrische data voor identificatie is verboden, tenzij het te maken heeft met wetenschappelijk onderzoek
2. Pseudonymisatie is belangrijk
3. Als de persoon toestemt ervoor

Gezichtsherkenning in de praktijk – Servaas

Hoe kunnen we de voordelen van GH toepassen in het dagelijkse leven?

Vergelijking van getallen uit een neuraal netwerk. Eerst ref foto, dan nieuw foto

Meer ref fotos van dezelfde persoon toevoegen, om hogere accuracy te bereiken.