

Advanced Programming & Maths

Tuur Vanhoutte

19 februari 2021

Inhoudsopgave

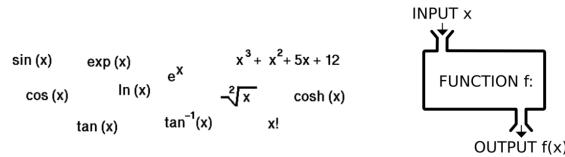
1 Basisfuncties in de wiskunde	1
1.1 Functies	1
1.2 Veelterm en veeltermfuncties	2
1.3 Bijzondere veeltermfuncties	2
1.3.1 Constante functie	3
1.3.2 Lineaire functie	3
1.3.3 Tweedegraadsfunctie	3
1.3.4 Derdegraadsfunctie	5
1.3.5 Exponentiële functie	5
2 Exponentiële verbanden in data	6
2.1 Lineaire groei	6
2.2 Exponentiële groei	7
2.3 Van groeipercentage naar groeifactor	7
2.3.1 Percentage naar factor	7
2.3.2 Factor naar percentage	8
2.4 Voorbeeld	8
2.5 Belangrijke maten voor exponentiële toename	9
2.5.1 Oefening: Combinatie van groeifactoren?	9
3 Belangrijke functies met betrekking tot machine learning	10
3.1 Logistische groei	10
3.1.1 Voorbeeld	10
3.1.2 De groei	10
3.1.3 Functievoorschrift	10
3.1.4 Voorbeeld	11
3.1.5 Algemene wiskundige notatie van een logistische functie	11
3.2 Regression analysis	12
3.2.1 Lineair regressiemodel	12
3.2.2 Logistisch regressiemodel	13
3.2.3 Lineair vs logistisch regressiemodel	14
3.2.4 Meerdere inputfactoren	14
3.3 Softmax functie	14
3.3.1 Kansen	15
3.3.2 Model	15
3.3.3 Wiskundig	16
3.4 Logistic regression cost function	16
3.4.1 Success meten	16
4 Pandas library	17
4.1 Inleiding	17
4.1.1 Welke data verwerken?	18
4.2 Pandas.core	18
4.3 Series	18
4.4 DataFrame	18
4.4.1 Select data from DataFrame	19
4.4.2 Veelgebruikte commandos bij dataframes	20
4.5 Loc vs iloc	21
4.5.1 iloc	21
4.5.2 loc	21

4.6	Plotten met pandas	22
4.6.1	Dataframe plotten	22
4.6.2	Series plotten	23
4.7	Demo: Iris Dataset	23
4.8	Complexe bewerkingen	24
5	Normaaldistributie	25
5.1	Doelstelling	25
5.2	Inleiding	25
5.3	Basisbegrippen	25
5.3.1	Kenmerken	25
5.3.2	Verdeling	27
5.3.3	Gevolg: Kansberekening	27
5.3.4	Wiskundig	27
5.4	Standaard normaalverdeling	28
5.4.1	Standardizeren van een normaalverdeling	28
5.5	Kansberekening	29
5.5.1	Via tabellen	30
5.6	Z-score	30
5.6.1	Wiskundig	30
5.6.2	Voorbeeld	30
5.6.3	Nut	31
5.7	Scheefheid	31
5.7.1	Berekening	31
5.8	Kurtosis	32
5.8.1	Wiskundig	32
5.9	Anomaly detection	32
5.9.1	Detectiemethode 1	33
5.9.2	Detectiemethode 2	33
5.9.3	Detectiemethode 3	34

1 Basisfuncties in de wiskunde

1.1 Functies

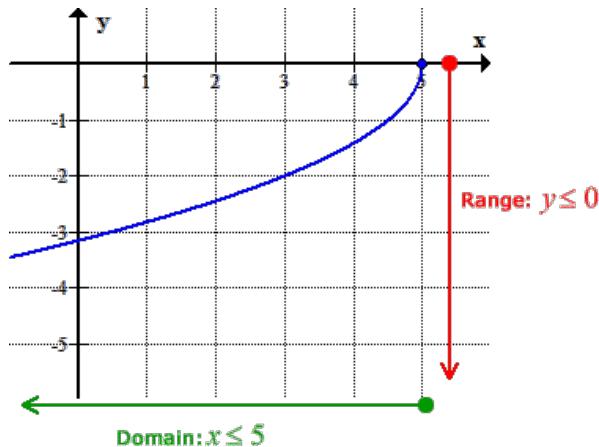
Definitie 1.1 (Reële functie) Een reële functie is een relatie in \mathbb{R} waarbij elke waarde x hoogstens één beeldwaarde $f(x)$ heeft



Figuur 1: Voorbeelden reële functies

Definitie 1.2 Voor elke functie geldt: er bestaat een ...

- (i) ... domein van de functie (domain)
- (ii) ... beeld van de functie (range)
- (iii) ... functievoorschrift van de functie

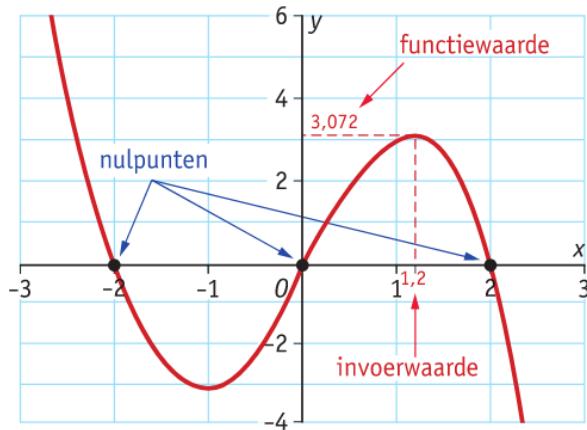


Figuur 2: Domein, bereik, functievoorschrift

$$f : \text{domein} \rightarrow \text{bereik} : x \rightarrow y = f(x)$$

$$f : \mathbb{R} \rightarrow \mathbb{R} : x \rightarrow y = x^3 - 4x$$

Definitie 1.3 Elke functie kan nulpunten hebben.



Figuur 3: $y = -x^3 + 4x$

Verloop van een functie wordt via een tekenschema verduidelijkt:

x		-2		0		2	
$f(x)$	+	0	-	0	+	0	-

Figuur 4: Tekenschema

1.2 Veelterm en veeltermfuncties

Definitie 1.4 (Veelterm)

$$A(x) = a_n x^n + a_{n-1} x^{n-1} + a_{n-2} x^{n-2} + \dots + a_2 x^2 + a_1 x + a_0 \quad (a_n, a_{n-1}, \dots, a_2, a_1, a_0 \in \mathbb{R}) \quad (1)$$

Definitie 1.5 (Veeltermfunctie)

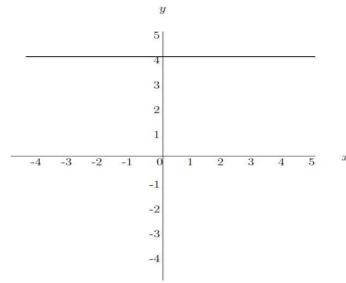
$$f(x) = a_n x^n + a_{n-1} x^{n-1} + a_{n-2} x^{n-2} + \dots + a_2 x^2 + a_1 x + a_0 \quad (2)$$

Graad van veelterm = n als $a_n \neq 0$

1.3 Bijzondere veeltermfuncties

- Constante functie: $f(x) = 4$
- Lineaire functie: $f(x) = 4$
- Tweedegraadsfunctie: $f(x) = 3x^2 + 2x + 1$
- Derdegraadsfunctie: $f(x) = 5x^3 - 3x^2 + 2x - 1$
- Exponentiële functie: $f(x) = 2^x$
- Logaritmische functie: $(fx) = \log_2(x)$

1.3.1 Constante functie



Figuur 5: $y = 4$

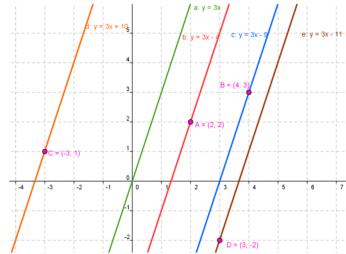
1.3.2 Lineaire functie

Definitie 1.6 (Lineaire functie)

$$f(x) = ax + b \quad (3)$$

Voorbeeld: $f(x) = 3x + 6$

- Betekenis van a : de richtingscoëfficiënt (rico)
- Betekenis van b : het snijpunt met de y -as
- Nulpunt: $f(x) = 0$
 $\Leftrightarrow 3x + 6 = 0$
 $\Leftrightarrow 3x = -6$
 $\Leftrightarrow x = -2$



Figuur 6: Meerdere evenwijdige lineaire functies

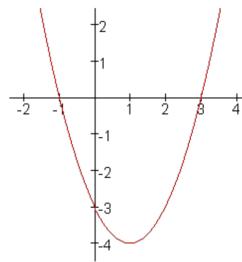
Evenwijdige rechten als: als $a_1 = a_2$

Loodrechte rechten als: als $a_1 \cdot a_2 = -1$

1.3.3 Tweedegraadsfunctie

Definitie 1.7

$$f(x) = ax^2 + bx + c, \quad (a \neq 0) \quad (4)$$



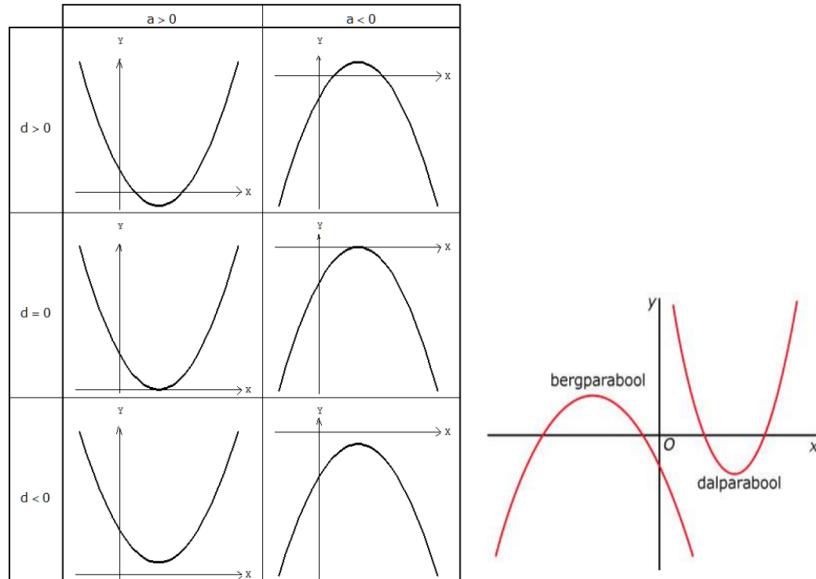
Figuur 7: $f(x) = x^2 - 2x - 3$

- Betekenis van a : positief \Rightarrow dalparabool, negatief \Rightarrow bergparabool
- Nulpunten: via de discriminant berekenen:

Definitie 1.8 (Discriminant) Bij een tweedegraadsvergelijking is de discriminant:

$$D = b^2 - 4ac \quad (5)$$

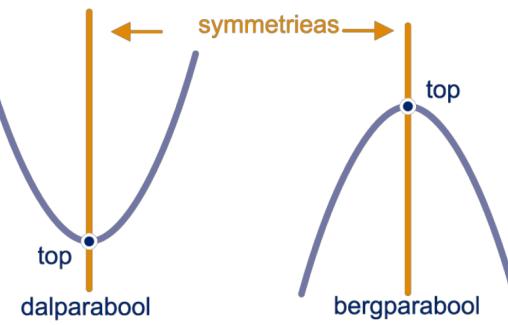
- Geval 1: $D > 0 \Rightarrow$ de functie heeft 2 nulpunten
- Geval 2: $D = 0 \Rightarrow$ de functie heeft 1 nulpunt
- Geval 3: $D < 0 \Rightarrow$ de functie heeft géén nulpunten



Figuur 8: De discriminant toont de nulpunten

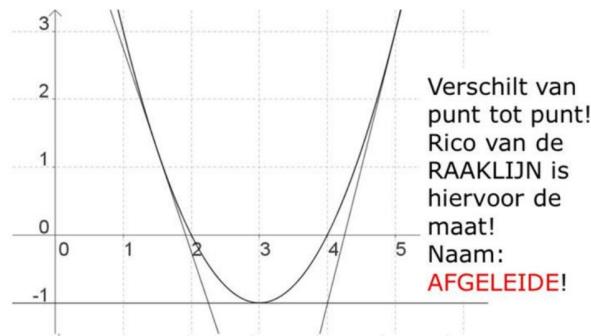
Nulpunten berekenen:

$$x_{1,2} = \frac{-b \pm \sqrt{D}}{2a} \quad (6)$$



Figuur 9: Symmetrieas: $x = \frac{-b}{2a}$

Voorbeeld:



Figuur 10: $y = x^2 - 6x + 8$

1.3.4 Derdegraadsfunctie

Definitie 1.9 (Derdegraadsfunctie)

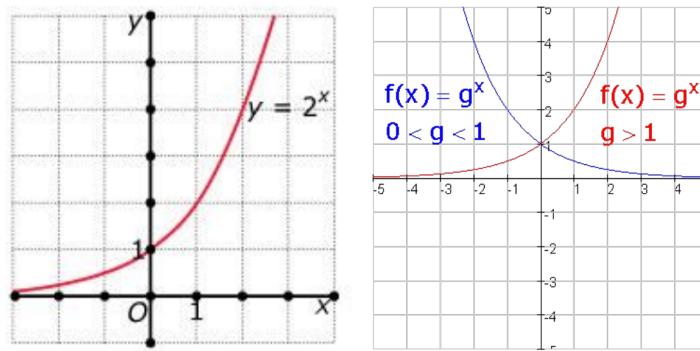
$$f(x) = ax^3 + bx^2 + cx + d (a \neq 0) \quad (7)$$

1.3.5 Exponentiële functie

Definitie 1.10 (Exponentiële functie)

$$f(x) = a^{g(x)} \quad (8)$$

Met grondtal $a \in \mathbb{R}_0^+ \setminus \{1\}$



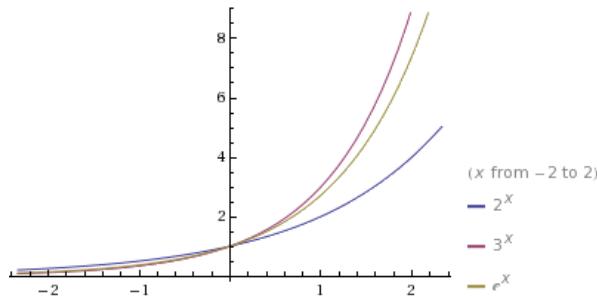
Figuur 11

- Betekenis van a: groefactor
- Wanneer stijgend?
- Wanneer dalend?
- Nulpunten:
- Vaststelling beeld functie

Definitie 1.11 (Constante van Euler)

$$e \approx 2.718281828 \dots \quad (9)$$

$f(x) = e^x$ is een bijzondere exponentiële functie



Figuur 12: Verschil tussen 2^x , 3^x en e^x

2 Exponentiële verbanden in data

2.1 Lineaire groei

Kenmerkend:

- Per tijdseenheid wordt hetzelfde getal **opgeteld**
- Grafiek is een rechte
- Algemene formule (N = aantal, t = tijd, b : beginhoeveelheid):

$$N = a \cdot t + b \quad (10)$$

t	0	1	2	3	4	5
N	750	780	810	840	870	900
	+30	+30	+30	+30	+30	+30

Figuur 13: Lineaire groei

2.2 Exponentiële groei

Kenmerkend:

- Per tijdseenheid wordt de hoeveelheid met hetzelfde getal **vermenigvuldigd**
- Grafiek is een exponentiële functie
- **Algemene formule:**

$$N = b \cdot g^t \quad (11)$$

t	0	1	2	3	4
N	1280	1600	2000	2500	3125
	x1,25	x1,25	x1,25	x1,25	x1,25

Figuur 14: Exponentiële groei

LENGTE FIETSPADEN IN NEDERLAND					
jaar	1998	2002	2006	2010	2014
aantal km	17 600	21 500	26 200	32 000	39 000

Figuur 15: Voorbeeld exponentiële groei met groeifactor ≈ 1.22

2.3 Van groeipercentage naar groeifactor

De toename/afname wordt vaak ook procentueel uitgedrukt

- Een jaarlijkse toename van 14.6%
- Een jaarlijkse afname van 14.6%

Definitie 2.1 (Groeifactor) *De groeifactor is de factor die per tijdseenheid wordt vermenigvuldigd met de vorige waarde.*

2.3.1 Percentage naar factor

$$g = \frac{p + 100}{100}\% \quad (12)$$

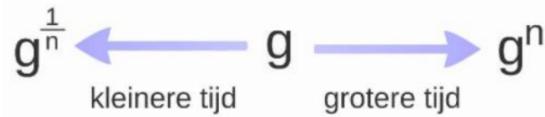
$$\begin{aligned} 100\% + 14,6\% &= 114,6\% = 1,146 \\ \times 1,146 \end{aligned} \qquad \begin{aligned} 100\% - 14,6\% &= 85,4\% = 0,854 \\ \times 0,854 \end{aligned}$$

Figuur 16: Van groeipercentage naar groeifactor

2.3.2 Factor naar percentage

$$0,765 = 76,5\% - 100\% = -23,5\%$$

Figuur 17: Van groeifactor naar groeipercentage



Voorbeeld

De groeifactor per uur is gelijk aan 3.

De groeifactor per 2 uur is gelijk aan $3^2 = 9$

De groeifactor per 20 minuten is gelijk aan $3^{\frac{1}{3}} = 1,44$.

Figuur 18: Let op: hier gebeuren vaak fouten bij het omrekenen

2.4 Voorbeeld

Een hoeveelheid groeit exponentieel. Na 5u is $N = 82$ en na 12u is $N = 246$.

Stel de formule van N op.

Oplossing

$$N = b \cdot g^t \quad (13)$$

Stap 1: groeifactor berekenen per tijdseenheid:

$$\left. \begin{array}{l} \text{Na 5u} \rightarrow N = 82 \\ \text{Na 12u} \rightarrow N = 246 \end{array} \right\} \Delta = 7u \rightarrow 164$$

Groeifactor voor 7 uren: $\frac{246}{82} = 3$

Groeifactor voor 1 uur: $3^{1/7} \approx 1.170$

Stap 2: 1 punt nemen waarvan we N weten:

Gekozen punt: (5, 82)

$$82 = b \cdot (1.170)^5$$

$$\Leftrightarrow b = \frac{82}{1.170}^5 \approx 37$$

$$\Leftrightarrow N = 37 \cdot 1.170^t$$

2.5 Belangrijke maten voor exponentiële toename

Definitie 2.2 (Verdubbelingstijd) *De verdubbelingstijd is de nodige tijd tot de hoeveelheid verdubbeld is.*

De verdubbelingstijd t kan je berekenen met:

$$g^t = 2 \quad (14)$$

Oefening

De populatie neemt toe met 8.3% per jaar. Bereken de verdubbelingstijd:

$$\begin{aligned} g^t &= 2 \\ \Leftrightarrow (1.083)^t &= 2 \\ \Leftrightarrow \log(1.083^t) &= \log(2) \\ \Leftrightarrow t \cdot \log(1.083) &= \log(2) \\ \Leftrightarrow t &= \frac{\log(2)}{\log(1.083)} \\ \Leftrightarrow t &= 8.69 \text{ jaar} \end{aligned}$$

Definitie 2.3 (Halveringstijd) *De halveringstijd is de nodige tijd tot de hoeveelheid gehalveerd is.*

De halveringstijd t kan je berekenen met:

$$g^t = 1/2 \quad (15)$$

2.5.1 Oefening: Combinatie van groeifactoren?

Een hoeveelheid neemt eerst 5 jaar lang met vast percentage (*) toe, om daarna nog 3 jaar met 10% per jaar toe te nemen. Na 8 jaar is de totale hoeveelheid verdubbeld.

(*) Bereken het jaarlijkse groeipercentage in de eerste 5 jaren.

Oplossing

We weten:

- Eerste 5 jaar: toename met vast percentage
- Volgende 3 jaar: toename met 10% (= factor van 1.1)
- Na 8 jaar: hoeveelheid verdubbeld (= factor van 2)

$$g^5 \cdot 1.1^3 = 2$$

We moeten g vinden:

$$\begin{aligned} \Leftrightarrow g^5 &= \frac{2}{1.1^3} \\ \Leftrightarrow g &= \sqrt[5]{\frac{2}{1.1^3}} \end{aligned}$$

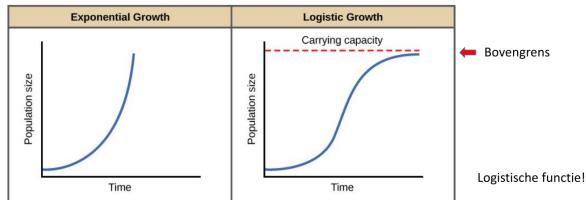
3 Belangrijke functies met betrekking tot machine learning

3.1 Logistische groei

3.1.1 Voorbeeld

Startsituatie: een bos (bv 10km²) waarin een konijnenepidemie uitbreekt. Boswachter houdt de populatie van de konijnen bij. Wat stelt hij vast?

De groei van de populatie verloopt volgens een typisch patroon (niet exponentieel):



Figuur 19: De rode lijn is de bovengrens

3.1.2 De groei

= de mate van toename

- Hangt af van hoeveel er al zijn tegenover hoeveel er nog bij kunnen
- Heel sterke verandering bij start, op het einde heel kleine verandering
- Hangt dus ook af van de tijd

Definitie 3.1 (De logistische groei) *De logistische groei is de mate van toename, afhankelijk van hoeveel er nog bij kan en hoeveel er al is*

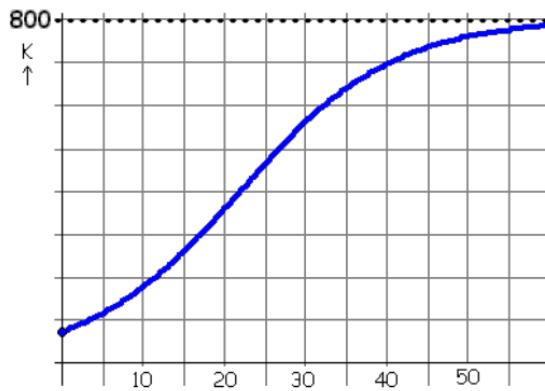
$$\frac{\text{Hoeveel er nog bij kan}}{\text{Hoeveel er al is}} = B \cdot g^t \quad (16)$$

- t = de tijd,
- B en g = constanten

3.1.3 Functievoorschrift

$$y = \frac{G}{1 + B \cdot g^t} \quad (17)$$

- t = de tijd
- B en constanten
- G = bovengrens



Figuur 20: Grafiek logistische groei met $G = 800$

3.1.4 Voorbeeld

Het aantal vissen in een meer is gegeven door:

$$N = \frac{2500}{1 + 5.5 \cdot 0.74^t}$$

waarbij N = aantal vissen, t = tijd

Beredeneer: Wanneer bereiken we het 'verzadigingsniveau'

Als t heel groot is:

- Dan wordt $0.74^t \approx 0$
- Dan wordt $5.5 \cdot 0.74^t \approx 0$
- Dan wordt $N \approx 2500$
- \Rightarrow Het meer is 'verzadigd'

3.1.5 Algemene wiskundige notatie van een logistische functie

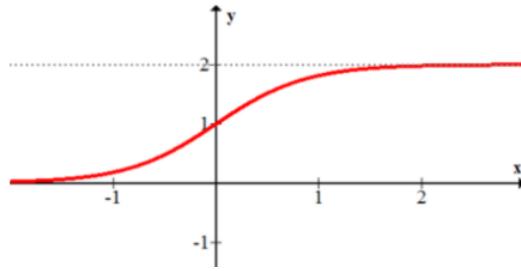
Definitie 3.2 (Logistische functie) De wiskundige notatie voor een logistische functie is:

$$f(x) = \frac{c}{1 + a \cdot b^x} \tag{18}$$

met a, b, c constanten waarbij de constante c de belangrijkste is:

c drukt uit wat de maximumwaarde kan zijn

$$f(x) = \frac{2}{1 + 0.1^x}$$



Figuur 21

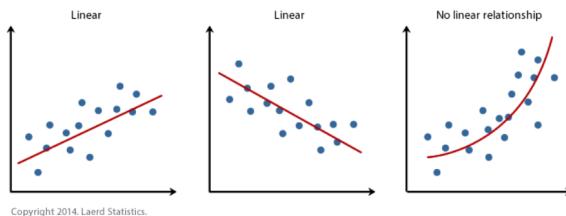
3.2 Regression analysis

Regressieanalyse:

- Is er een (voorspellend) verband tussen 2 variabelen
- Heeft de ene variabele een invloed op de andere variabele



Figuur 22: Regressieanalyse

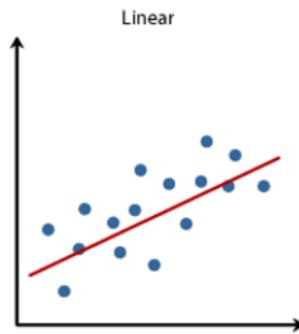


Figuur 23: Lineaire vs niet-lineaire samenhang

3.2.1 Lineair regressiemodel

Enkelvoudige vorm:

- 1 inputwaarde x
- via lineaire functie $h_\theta(x) = \theta_0 + \theta_1 x$



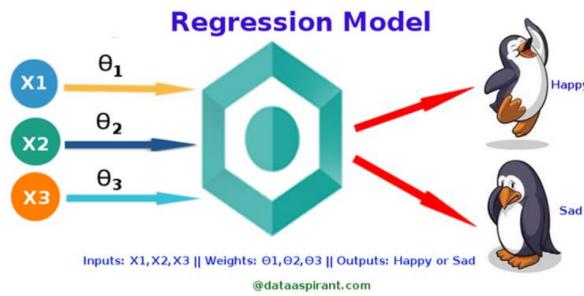
Figuur 24

- Aan de hand van de opgestelde functie doe je een voorspelling
- **Doel:** een zo goed mogelijke lineaire functie opstellen
- ⇒ zoektocht naar de beste θ_0 en θ_1

3.2.2 Logistisch regressiemodel

Logistische regressie = **Classificatie-algoritme**

Zoeken naar een model dat uitkomst (2 mogelijkheden) voorspelt mbv inputwaarden. Elke inputwaarde heeft een zeker belang (gewicht).



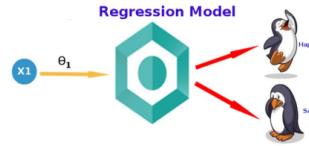
Figuur 25: 3 inputs met elk een bepaald gewicht, die een uitkomst zoekt (2 mogelijkheden)

Vereenvoudiging:

- 1 inputwaarde x
- Logistische functie $p = \frac{1}{1+e^{-(b_0+b_1x)}}$

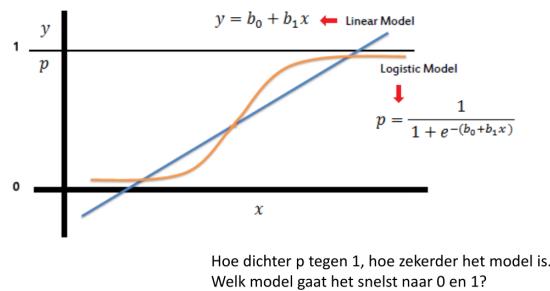
Uitkomst:

- de persoon slaagt als $h_\theta(x) \geq 0.5$
- de persoon slaagt niet als $h_\theta(x) < 0.5$



Figuur 26: 1 inputwaarde x , met twee uitkomsten

3.2.3 Lineair vs logistisch regressiemodel



Figuur 27: Hoe dichter p tegen 1, hoe zekerder het model is

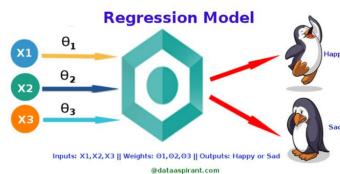
Welk model gaat het snelst naar 0 en 1?

- Het logistische model
- Daarom is het logistische model beter voor classificatie: je splitst de groep op in 2

3.2.4 Meerdere inputfactoren

Zelfde redenering:

- Meerdere inputwaarden x_1, x_2, \dots
- Gebruik $\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots$



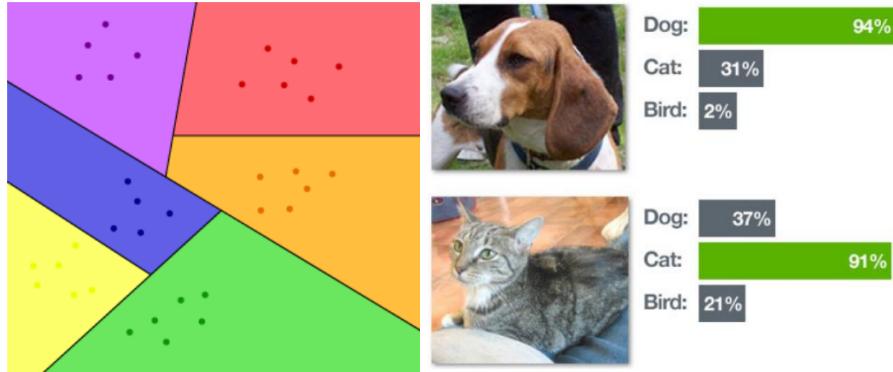
Figuur 28: Regressiemodel met meerdere inputfactoren

3.3 Softmax functie

Doelstelling:

- Model dat in staat is om data te gaan categoriseren
- Hoe?

- ⇒ Met behulp van verschillende inputvariabelen en bijhorende parameters



Figuur 29: Categoriseren met de softmax functie

3.3.1 Kansen

Kans dat de toestand tot groep A behoort:

- $\theta_{A,0} + \theta_{A,1}x_1 + \theta_{A,2}x_2$
- Voorbeeld: $0.01 + 0.1x_1 + 0.1x_2$

Kans dat de toestand tot groep B behoort:

- $\theta_{B,0} + \theta_{B,1}x_1 + \theta_{B,2}x_2$
- Voorbeeld: $0.1 + 0.2x_1 + 0.2x_2$

Kans dat de toestand tot groep C behoort:

- $\theta_{C,0} + \theta_{C,1}x_1 + \theta_{C,2}x_2$
- Voorbeeld: $0.1 + 0.3x_1 + 0.3x_2$

3.3.2 Model

Het softmax-model berekent de mate van zekerheid dat een toestand tot een bepaalde categorie behoort.

vb: volgende quotiënt drukt uit hoe zeker hij is dat (z_1, z_2) tot categorie A behoort:

$$\frac{e^{\theta_{A,0} + \theta_{A,1}z_1 + \theta_{A,2}z_2}}{e^{\theta_{A,0} + \theta_{A,1}z_1 + \theta_{A,2}z_2} + e^{\theta_{B,0} + \theta_{B,1}z_1 + \theta_{B,2}z_2} + e^{\theta_{C,0} + \theta_{C,1}z_1 + \theta_{C,2}z_2}}$$

(analoog voor categorie B en C: vervang de teller)

$$\frac{e^{0.01 + 0.1 \cdot 0.1 + 0.1 \cdot 0.5}}{e^{0.01 + 0.1 \cdot 0.1 + 0.1 \cdot 0.5} + e^{0.1 + 0.2 \cdot 0.1 + 0.2 \cdot 0.5} + e^{0.1 + 0.3 \cdot 0.1 + 0.3 \cdot 0.5}} = 0.2945$$

Figuur 30: Betekenis: het model is 29% zeker dat $(0.1, 0.5)$ tot categorie A behoort. Bereken zelf als oefening voor B en C

3.3.3 Wiskundig

Het gebruikte model wordt via volgende wiskundige formule algemeen beschreven:

$$\frac{e^{x_k}}{\sum_{i=1}^n e^{x_i}} \quad (19)$$

waarbij:

- $x_k = \theta_{k,0} + \theta_{k,1}x_1 + \theta_{k,2}x_2 + \cdots + \theta_{k,m}x_m$
- n = aantal groepen
- m = het aantal meetcriteria

3.4 Logistic regression cost function

Het model:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^\top x}} \quad (20)$$

waarbij:

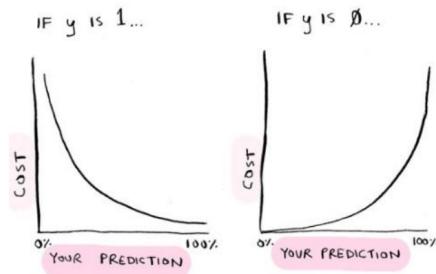
- $\theta^\top x = \theta_0 + \theta_1x_1 + \theta_2x_2$
- h_θ drukt uit wat de kans is dat voor opgegeven x_1 en x_2 de waarneming tot 1 groep behoort
- x_1 en x_2 zijn de inputwaardes
- θ_1 en θ_2 zijn gewichten (hoe belangrijk is de input)
- **Doel:** vinden van de beste gewichten zodat de voorspelling == de werkelijkheid

3.4.1 Success meten

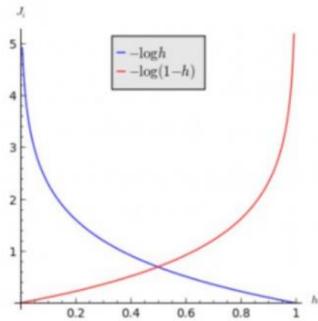
Stel: je maakt een logistisch regressiemodel die bepaalt of een object een groene appel of een tennisbal is.

- Bepalen van de kostenfunctie $J(\theta)$ met als doel deze zo laag mogelijk te brengen
- kost = afwijking tegenover de werkelijke situatie
- werkelijkheid kan 2 situaties zijn:
 - Indien de werkelijkheid een groene appel is $\Rightarrow y = 1$
 - Indien de werkelijkheid géén groene appel is $\Rightarrow y = 0$

Hoe ziet zo'n kostfunctie er dan uit?



Figuur 31: Als $y = 1$ en $y = 0$



Figuur 32

$$\begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Figuur 33

Hoe brengen we 2 mogelijke situaties in 1 functie samen?

(TODO: slide 24 - 32)

4 Pandas library

4.1 Inleiding

- Doelstelling:
 - Nut van de pandas library kunnen situeren
 - Data-analyse: basisbewerkingen
- Pandas = ‘Python Data Analysis Library’
- Pandas bouwt op de NumPy library
- Officiële website: <https://pandas.pydata.org/>
- Goede start: <http://pandas.pydata.org/pandas-docs/stable/10min.html>

4.1.1 Welke data verwerken?

- csv-files
- txt-files
- Excel-files
- Databases

4.2 Pandas.core

Beschikbare datastructuren:

- Series (1D)
- DataFrame (2D)
- Panel (3D)

4.3 Series

Bestemd voor 1-dimensionale data:

'a one-dimension labeled array capable of holding any data'

- Subklasse van numpy-ndarray
- Data: elk soort datatype
- Geordende index
- Duplicaten mag (maar niet optimaal)

	index	values
A	→	5
B	→	6
C	→	12
D	→	-5
E	→	6.7

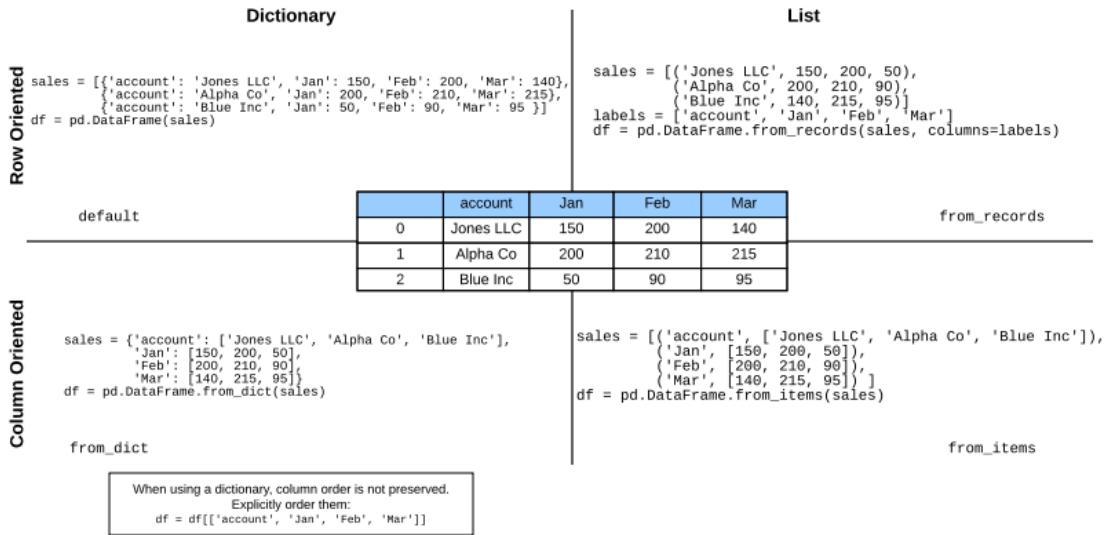
Figuur 34: Elk element heeft een index

4.4 DataFrame

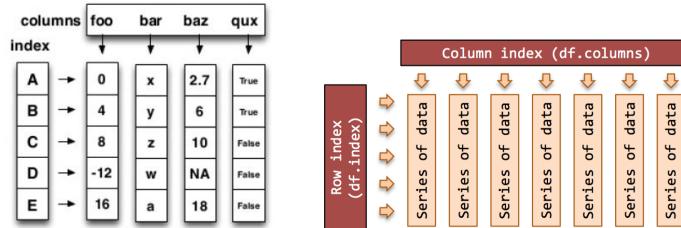
Bestemd voor meer-dimensionale data

- Subklasse van numpy-ndarray
- Elke kolom kan ander datatype hebben
- Rij en kolom index
- Grootte wijzigbaar (invoegen/verwijderen van rijen en kolommen)

Creating Pandas DataFrames from Python Lists and Dictionaries



Figuur 35: DataFrames maken uit Python lists en dictionaries



Figuur 36: Elk element heeft een rij en kolom

4.4.1 Select data from DataFrame

Via operator [] selecteer je een kolom:

	Feb	Jan	Mar	account
0	200	150	140	Jones LLC
1	210	200	215	Alpha Co
2	90	50	95	Blue Inc

Figuur 37: Voorbeeld DataFrame

```
# selecteer de kolom Jan uit dataframe
df['Jan']
```

```
# analoog: elke kolom is dus een attribuut van dataframe
df.Jan
```

```
# returnwaarde: Series-object
```

```

0    150
1    200
2     50
Name: Jan, dtype: int64

```

Via operator [] selecteer je een kolom & krijg je een dataframe terug:

```

>> df[['Jan']]
# returns:
   Jan
0  150
1  200
2   50

>> df[['Jan', 'Mar']]
# returns:
   Jan    Mar
0  150   140
1  200   215
2   50    95

```

Via de operator [] en met een conditie:

	Feb	Jan	Mar	account
0	200	150	140	Jones LLC
1	210	200	215	Alpha Co
2	90	50	95	Blue Inc

Figuur 38: Voorbeeld DataFrame

```

>> df[df.Jan > 60]
# returns:
   Feb    Jan    Mar    account
0  200   150   140  Jones LLC
1  210   200   215  Alpha Co

>> df[np.logical_and(df.Jan > 100, df.Feb <= 200)]
# returns:
   Feb    Jan    Mar    account
0  200   150   140  Jones LLC

>> df[df.account.str.startswith('Alpha')]
# test deze eens zelf uit als oefening :(

```

4.4.2 Veelgebruikte commandos bij dataframes

df.shape	# geeft de dimensie als een tuple terug
df.info()	# oplijsting van de aanwezige kolommen
df.head([aantal])	# eerste vijf/aantal rijen
df.tail([aantal])	# laatste vijf/aantal rijen
df.index	# geef de index-kolom weer
df.columns	# geef de kolomnamen weer
df.describe()	# geef snel overzicht van statistische data
df.T	# transponeer data (rij -> kol, kol -> rij)

```
df.sort_index()      # sorteert op basis van index
df.sort_values()    # sorteren op één of meerdere kolommen
```

4.5 Loc vs iloc

4.5.1 iloc

= Integer-location based indexing / selection by position

- Nut: selecteren van rijen en kolommen via rij/kolomnummer
- Syntax: data.iloc[<row>, <column>]
- Returnwaarde:
 - Indien 1 **rij** ⇒ series-object
 - Indien meerdere **rijen**: ⇒ dataframe-object
 - 1 of meerdere **kolommen**: ⇒ dataframe-object

iloc-voorbeelden:

```
# Rows:
data.iloc[0] # first row of data frame
data.iloc[1] # second row of data frame
data.iloc[-1] # last row of data frame

# Columns:
data.iloc[:,0] # first column of data frame
data.iloc[:,1] # second column of data frame
data.iloc[:, -1] # last column of data frame

data.iloc[0:5] # first five rows of dataframe

# first two columns of data frame with all rows
data.iloc[:, 0:2]

# 1st, 4th, 7th, 25th row + 1st 6th 7th columns.
data.iloc[[0,3,6,24], [0,5,6]]

# first 5 rows and 5th, 6th, 7th columns of data frame
data.iloc[0:5, 5:8]
```

4.5.2 loc

= label based indexing / selection

- Nut: selecteren van rijen en kolommen via label / via conditionele look-up
- Syntax: data.loc[<row>, <column>]
- Returnwaarde:
 - Indien 1 **rij/kol** ⇒ series-object
 - Indien meerdere **rijen**: ⇒ dataframe-object
 - 1 of meerdere **kolommen**: ⇒ dataframe-object

loc-voorbeelden:

	cars_per_cap	country	drives_right
US	809	United States	True
AUS	731	Australia	False
JAP	588	Japan	False
IN	18	India	False
RU	200	Russia	True
MOR	70	Morocco	True
EG	45	Egypt	True

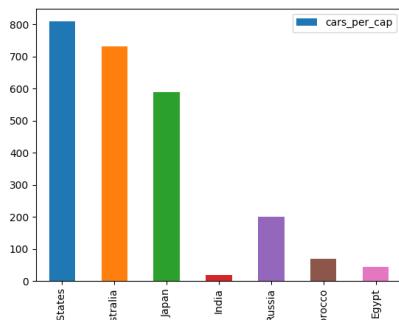
Figuur 39: Voorbeeld dataframe

```
reviews.loc[:2, "score"] # return type =  
  
reviews.loc[:2, ["score", "title"]] # return type =  
  
# select column "score" where value of index <= 5  
reviews.loc[:5, "score"]  
  
# select columns "country" and "cars_per_cap" where rowindex is "US" or "RU"  
cars.loc[ ["US","RU"] , ["country","cars_per_cap"]]  
  
# select columns "country" and "cars_per_cap" where rowindex is from "US" to "RU"  
cars.loc[ "US " : "RU " , ["country","cars_per_cap"]]  
  
# selectie rijen hoeft niet altijd op basis van row-index te zijn  
# select columns "country" and "drives_right", voor de landen 'Japan' en 'India'  
cars.loc[ cars.country.isin( ['Japan', 'India'] ) , ['country','drives_right']]
```

4.6 Plotten met pandas

4.6.1 Dataframe plotten

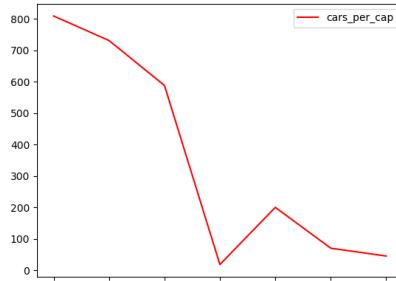
```
# print(cars[['country', 'cars_per_cap']])  
# werkwijze 1:  
cars[['country', 'cars_per_cap']].plot(kind='bar', legend=True)  
# werkwijze 2:  
cars.plot(x='country', y='cars_per_cap', kind='bar', legend=True)  
  
plt.show()
```



Figuur 40: Resultaat

4.6.2 Series plotten

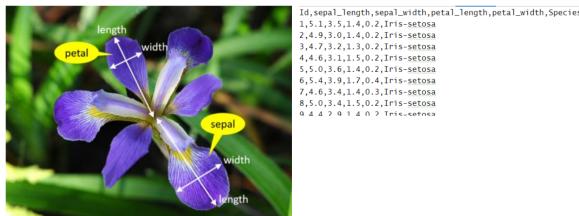
```
# werkwijsje 1:  
plt.plot(cars['cars_per_cap'])  
# werkwijsje 2:  
plt.plot(cars['cars_per_cap'].plot(color='r', legend=True))  
  
plt.show()
```



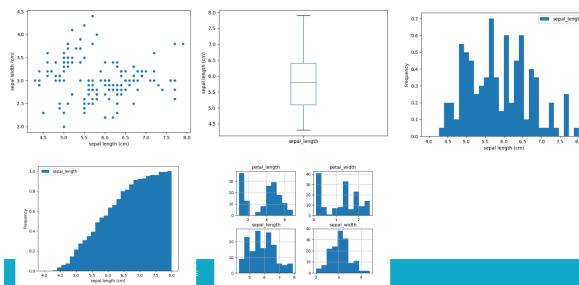
Figuur 41: Resultaat

4.7 Demo: Iris Dataset

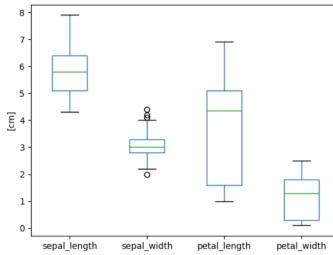
(Zie DemoPandas.zip op Leho voor de code)



Figuur 42: Een iris bestaat uit petals & sepals, met elk hun breedte en lengte

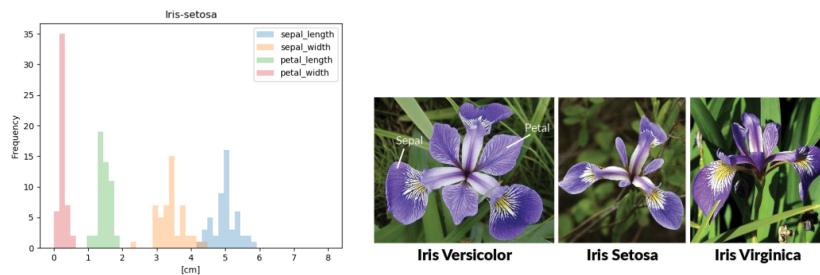


Figuur 43: Plotten van de beschikbare data (demo5.py)



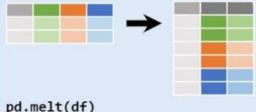
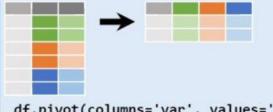
Figuur 44: Vergelijken van de lengtes en breedtes adhv box-plots (demo6.py)

```
# filteren van data
result_check = iris['Species'] == 'Iris-setosa'
# print(type(result_check)) == TODO
filtered_setosa = iris.loc[result_check, :]
# of in 1 lijn:
filtered_setosa = iris.loc[iris['Species'] == 'Iris-setosa', :]
```



Figuur 45: Frequentiediagram voor de Iris-setosa soort (demo7.py)

4.8 Complexe bewerkingen

		<code>df.sort_values('mpg')</code> Order rows by values of a column (low to high).
<code>pd.melt(df)</code> Gather columns into rows.	<code>df.pivot(columns='var', values='val')</code> Spread rows into columns.	<code>df.sort_values('mpg', ascending=False)</code> Order rows by values of a column (high to low).
		<code>df.rename(columns = {'y': 'year'})</code> Rename the columns of a DataFrame
<code>pd.concat([df1, df2])</code> Append rows of DataFrames	<code>pd.concat([df1, df2], axis=1)</code> Append columns of DataFrames	<code>df.sort_index()</code> Sort the index of a DataFrame
		<code>df.reset_index()</code> Reset index of DataFrame to row numbers, moving index to columns.
		<code>df.drop(['Length', 'Height'], axis=1)</code> Drop columns from DataFrame

Figuur 46: Reshaping data: change the layout of a data set

(komen we later nog op terug)

5 Normaalverdeling

5.1 Doelstelling

- Het herkennen van de eigenschappen van de normaalverdeling
- Verband standaard normaalverdeling en Z-waarden
- Z-index, Z-score tabel

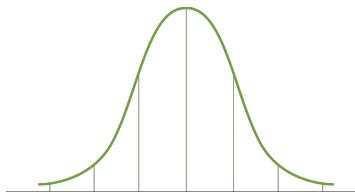
5.2 Inleiding

Wanneer komt een normaalverdeling voor?

- Grote aantallen onafhankelijke waarnemingen uit een willekeurige populatie
 - De lengte van personen
 - Productieprocessen die een bepaalde tijdsduur hebben
 - ...

5.3 Basisbegrippen

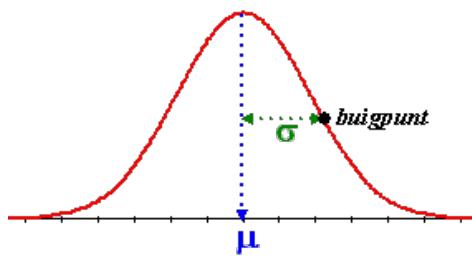
Definitie 5.1 (Normaalverdeling) *Een intervalwaarde dat afhankelijk is van oneindig aantal onafhankelijke factoren (die los van elkaar in werken) zal in de populatie een normaalverdeling vertonen (Gauss-curve)*



Figuur 47: Gauss-curve

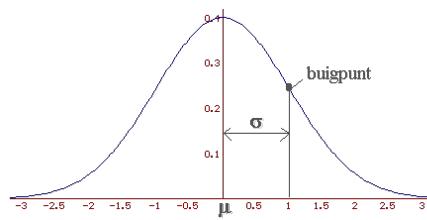
5.3.1 Kenmerken

- Klokvormig verloop
- 1 maximum = gemiddelde = mediaan (μ)
- Uitslagen vooral geconcentreerd rond gemiddelde
- Frequentie daalt naarmate scores afwijken
- Twee buigpunten
- Symmetrie



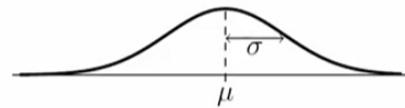
Figuur 48: Kenmerken Gauss-curve

- Standaarddeviatie (standaardafwijking) σ weerspiegelt mate van spreiding
- Wordt gemeten in de buigpunten



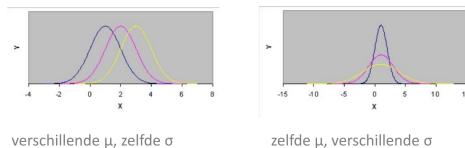
Figuur 49

- Functie is volledig te beschrijven met het gemiddelde μ en de Standaarddeviatie σ



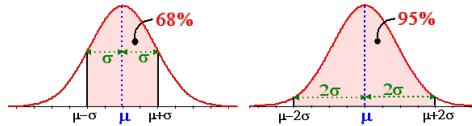
$$X \sim N(\mu, \sigma^2)$$

Figuur 50: Notatiewijze

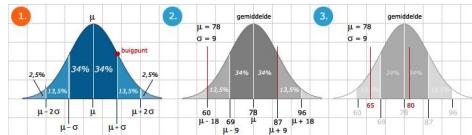


Figuur 51: Impact μ en σ

5.3.2 Verdeling



Figuur 52



Figuur 53

- Tussen beide buigpunten ($\mu - \sigma, \mu + \sigma$): $\approx 68\%$ van de populatie
- Tussen ($\mu - 2\sigma, \mu + 2\sigma$): $\approx 95\%$ van de populatie
- Tussen ($\mu - 3\sigma, \mu + 3\sigma$): $\approx 99\%$ van de populatie

5.3.3 Gevolg: Kansberekening

- Oppervlakte onder de grafiek = 100%
- Hiermee kunnen we nu uitdrukken wat de kans van een specifiek bereik kan zijn. Bv:
 - Met $\mu = 78, \sigma = 9 \Rightarrow$ Kans dat $x < 65$?
 - \Rightarrow bereken oppervlakte links van 65

5.3.4 Wiskundig

$$f(x) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \cdot e^{-\frac{(x-\mu)^2}{2 \cdot \sigma^2}} \quad (21)$$

Standaardafwijking berekenen:

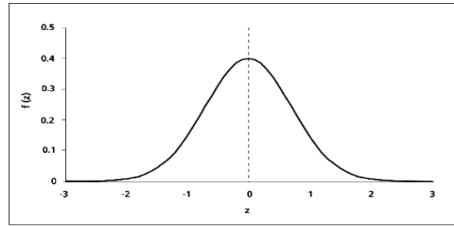
$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n - 1}} \quad (22)$$

- σ = Standaarddeviatie van getallenreeks x
- x_i = de waarde van getal i in de getallenreeks
- μ = het gemiddelde van de getallenreeks
- n = het aantal getallen in de proef

5.4 Standaard normaalverdeling

Definitie 5.2 (Standaard normaalverdeling) De standaardnormaalverdeling is een normaalverdeling waarbij:

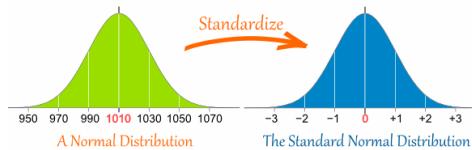
- $\mu = 0$
- $\sigma = 1$



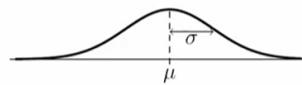
Figuur 54: Oppervlakte onder grafiek = 1

5.4.1 Standardizeren van een normaalverdeling

Doel: een normaalverdeling omzetten in een standaardnormaalverdeling waarbij $\mu = 0$ en $\sigma = 1$



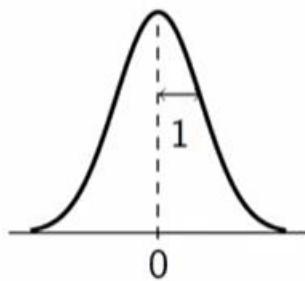
Figuur 55: Standardizeren



$$X \sim N(\mu, \sigma^2)$$

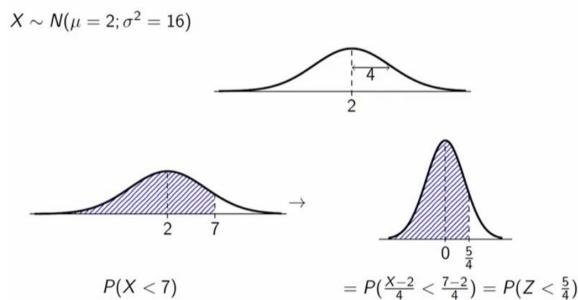
Figuur 56: Uitgangspunt: willekeurige normaalverdeling

- Verschuiven naar 0: $X - \mu$
- samendrukken/uitrekken tot $\sigma = 1$: $\frac{X-\mu}{\sigma}$



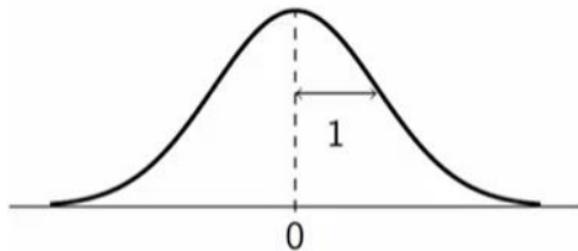
$$Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$$

Figuur 57: Gestandardiseerd: alle eigenschappen blijven bewaard, enkel de hoogte blijft hetzelfde



Figuur 58: Standardizeren: voorbeeld

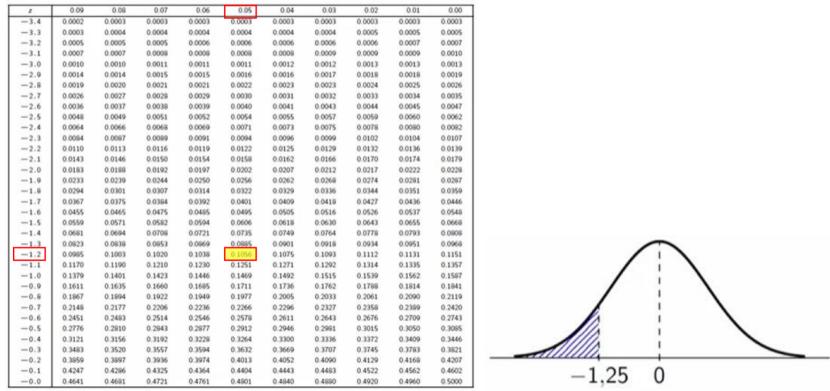
5.5 Kansberekening



Figuur 59: $Z \sim N(0, 1) \rightarrow \mu = 0$ en standaardafwijking $\sigma = 1$

- Grafiek beschrijft hoe waarschijnlijk uitkomsten zijn
- Uitkomsten dicht bij 0 waarschijnlijker dan in de staarten
- Kans = oppervlakte onder de grafiek, totale kans = 100% = 1

5.5.1 Via tabellen



Figuur 60: Kansberekening via tabellen: rijen = 1 getal na de komma, kolommen = 2de getal na de komma. $P(Z < 1.25) = 0.1056$

- Symmetrie: $P(Z < -1.25) = P(Z > 1.25) = 0.1056$
- Complement: $P(Z < -1.25) = 1 - P(Z < 1.25)$

5.6 Z-score

Definitie 5.3 (Z-score) Een Z-score geeft aan hoeveel standaardafwijkingen een observatie van het gemiddelde verwijderd is.

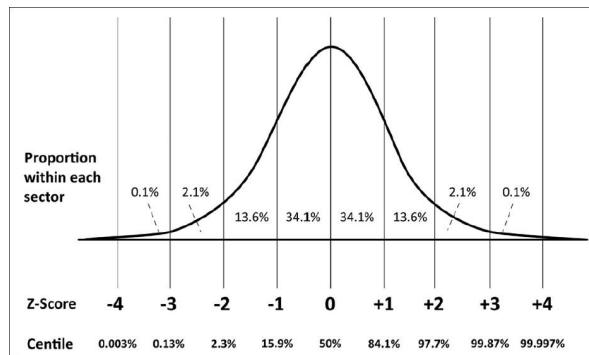
Je krijgt dus je plek ten opzichte van het gemiddelde, uitgedrukt in een standaard maat.

Voordeel: je weet direct hoe goed iemand scoort ten opzichte van de rest.

5.6.1 Wiskundig

$$z = \frac{x - \mu}{\sigma} \quad (23)$$

5.6.2 Voorbeeld



Figuur 61: Hoeveel standaarddeviations (σ) zit een score van het gemiddelde μ

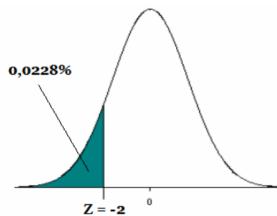
Voorbeeld: 20 studenten leggen een examen af.

- Gemiddelde $\mu = 7/10$
- Standaardafwijking $\sigma = 0.5$ (68% heeft score tussen [6.5, 7.5])
- Iemand heeft 6/10 $\Rightarrow z\text{-score} = -2 \Rightarrow 2 \cdot 0.5$ onder gemiddelde

5.6.3 Nut

Z-score is eveneens een uitdrukking van de kans:

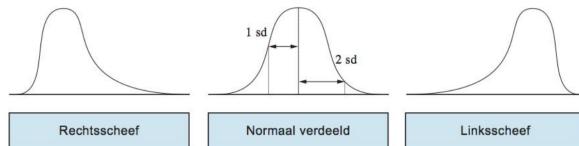
- Hoe (on)gebruikelijk is een score
- Bij vorig voorbeeld: 2.275% heeft een score $\leq 6/10$



Figuur 62

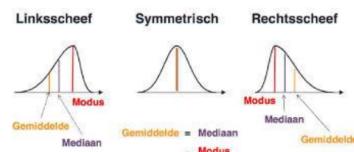
5.7 Scheefheid

Veel frequentieverdelingen hebben niet de vorm van de normaalverdeling. Ze zijn 'scheef'



Figuur 63: Scheefheid

- Linksscheef = gemiddelde is lager dan de mediaan
- Rechtsscheef = gemiddelde is hoger dan de mediaan



Figuur 64

5.7.1 Berekening

De maat voor scheefheid kan berekend worden:

- Linksscheef: negatieve waarde

- Rechtsscheef: positieve waarde

$$Scheefheid = \left(\frac{\left(\frac{\sum(x-\bar{x})^3}{n} \right)}{\left(\frac{\sum(x-\bar{x})^2}{n} \right)^{3/2}} \right) \quad (24)$$

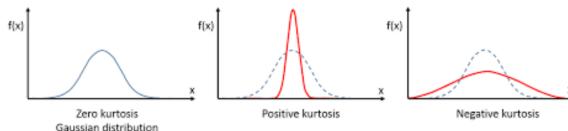
Interpretatie:

- Waarde tussen [-0.5, 0.5]: zeer goede normaalverdeling
- Waarde tussen [-1, 1]: redelijk goede normaalverdeling
- Waarde daarbuiten: geen goede normaalverdeling

5.8 Kurtosis

Definitie 5.4 (Kurtosis) *Kurtosis is de maat van de gepiektheid.*

Je gaat na of de verdeling een scherpe top heeft of een nogal vlakke top.



Figuur 65

5.8.1 Wiskundig

TODO (25)

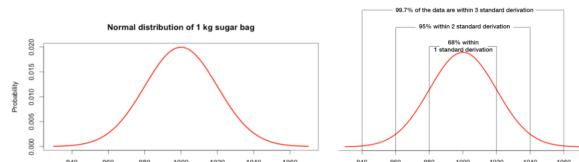
Interpretatie:

- 0 = normaalverdeling
- Minimum is -3: volledig plat
- Maximum is + inf

5.9 Anomaly detection

Voorbeeld: suikerfabriek produceert 1kg suikerzakken

- In werkelijkheid: nooit exact 1kg



Figuur 66: Verdeling 1kg suikerzak

Definitie 5.5 (Anomalie) Een anomalie in een normaalverdeling is een onregelmatigheid of afwijking
Soms verdacht, soms niet

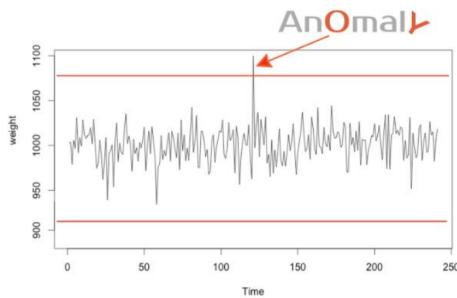
"Anomaly detection (also outlier detection) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset."

5.9.1 Detectiemethode 1

= Minimum- en maximumgrens vastleggen

Bv:

- wanneer een waarde kleiner is dan $\mu - 4\sigma$ (920 gram)
- wanneer een waarde groter is dan $\mu + 4\sigma$ (1080 gram)



Figuur 67

5.9.2 Detectiemethode 2

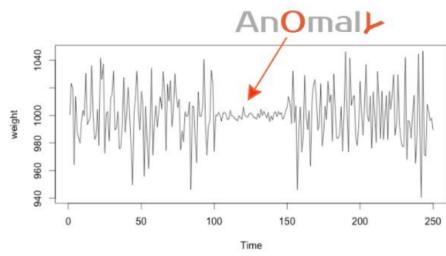
= detecteren van een zeldzame verdeling die afwijkt van de normaalverdeling

Bv:

- 4 opeenvolgende waarden die vallen in de buitengrenzen $\mu - 3\sigma$ en $\mu + 3\sigma$
- Kans voor 1 waarde: 0.3%
- Kans voor 4 opeenvolgende waardes: $(0.3\%)^4 = (0.003)^4 = 0.00000000081$

Bv:

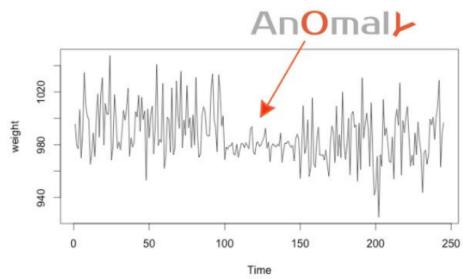
- 50 opeenvolgende waarden die vallen in de buitengrenzen $\mu - \sigma$ en $\mu + \sigma$
- Kans voor 1 waarde: 68%
- Kans 50 opeenvolgende waardes: $(0.68)^{50} = 4.221 \cdot 10^{-9}$



Figuur 68

5.9.3 Detectiemethode 3

Volledige verschuiving van de normaalverdeling



Figuur 69: Oplossing: gebruik van window