Inhoudsopgave

[8.1 Inleiding 4](#_Toc91791303)

[8.2 Drie types van Machinaal Leren 4](#_Toc91791304)

[8.2.1 Gesuperviseerd Leren 4](#_Toc91791305)

[8.2.2 Ongesuperviseerd Leren 6](#_Toc91791306)

[8.2.3 Reinforcement Learning 6](#_Toc91791307)

[8.3 Clustering en het K-Gemiddelden Algoritme 7](#_Toc91791308)

[8.3.1 Het K-gemiddelden Algoritme als optimalisatie 9](#_Toc91791309)

[8.3.2 Bepalen van het aantal clusters 12](#_Toc91791310)

Hoofdstuk 8 Machinaal Leren

We starten met de definitie van machinaal leren om dan direct over te gaan op de drie types van machinaal leren:

* Gesuperviseerd leren
* Ongesuperviseerd leren
* Reinforcement learning

Voor de eerste 2 types worden nog een aantal deeltypes besproken:

* Regressie
* Classificatie
* Clustering

In de laatste sectie bestuderen we de implementatie van het K-gemiddelden algoritme. Een algoritme voor clustering.

# 8.1 Inleiding

Machinaal leren is een belangrijk deelveld van artificiële intelligentie en wordt gebuikt in zeer veel toepassingen die door miljoenen mensen dagelijks gebruikt worden.

* Voorbeelden:
  + - Spam filters voor email
    - Gezichtsdetectie camera smartphone
    - …

Wanneer men een brief verzend, dan zijn er machines in postsorteercentra die automatisch het handschrift van het adres lezen adhv karakterherkenninssoftware.

**Definitie:** machinaal leren is het deelveld van artificiële intelligentie dat computers de mogelijkheid geeft om te leren zonder hiervoor expliciet geprogrammeerd te zijn.

## 8.2 Drie types van Machinaal Leren

Algemeen worden er 3 grote types van machinaal leren beschouwd. We definiëren deze nu kort een voor een.

### 8.2.1 Gesuperviseerd Leren

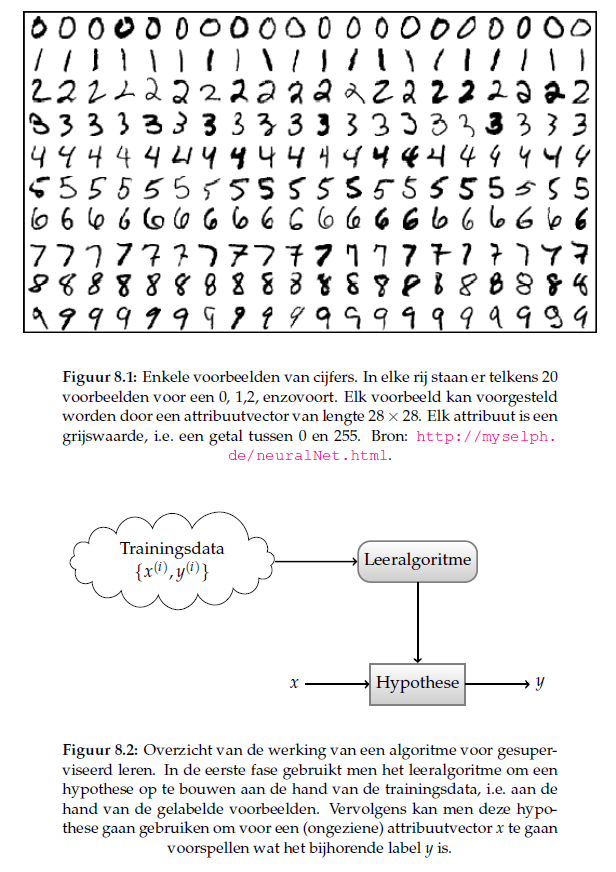
**Voorbeeld:** veronderstel dat we trachten te voorspellen wat de verkoopwaarde van een huis is op basis van de woonbare oppervlakte van dit huis. Om dit te realiseren beschikken we over een aantal voorbeelden van koppels. De taak bestaat erin om de verkoopwaarde te voorspellen wanneer een bepaalde bewoonbare oppervlakte werd gegeven.

**Voorbeeld:** bij spamdetectie bestaat de opdracht erin om op basis van een groot aantal e-mails, waarvan geweten is of ze spam zijn of niet, te voorspellen of een nieuwe email al dan niet spam is. Typisch gaat men eerst de tekst bewerken door woord normalisatie uit te voeren. Daarna alle stopwoorden die praktisch in alle e-mails voorkomen en dus praktisch geen enkele bijdrage leveren tot de inhoud van de email gaan verwijderen. ….

**Voorbeeld:** cijferherkenning: veronderstel dat men een groot aantal afbeeldingen heeft waarvan gegeven is of ze respectievelijk een 1,2,3,… 9 voorstellen. De taak van een cijferherkenningsalgoritme bestaat er dan in om voor een nieuwe afbeelding te beslissen welk cijfer het is.

**Definitie** de taak van gesuperviseerd leren bestaat erin om op basis van een gelabelde trainingsdataset een hypothese op te bouwen waarmee, voor elke invoer het label kan voorspeld worden. Wanner het label een getal is, dan spreekt men van een regressieprobleem, wanneer het label een va neen aantal voor gedefinieerde klassen is, dan spreekt men van een classificatieprobleem.

Het voorspellen van de verkoopwaarde is een regressieprobleem; spamdetectie en cijferherkenning zijn voorbeelden van classificatieproblemen. Omdat er bij spamdetectie slechts twee klassen zijn spreekt men ook van een binair classificatieprobleem.



Men ziet een overzicht van de werking van gesuperviseerd leren (fig8.2) we geven een groot aantal voorbeelden  aan het leeralgoritme waarbij y^(i) de “juiste” uitvoer is bij het attribuutvector x^(i). Het leer algoritme heeft als uitvoer een functie h die de hypothese genoemd wordt. Wanneer men aan deze functie een vector x meegeeft, dan antwoordt deze met het voorspelde label y dat bij x hoort.

### 8.2.2 Ongesuperviseerd Leren

(zie pdf voorbeelden)

**Definitie:** de taak van een algoritme voor ongesuperviseerd leren bestaat erin om structuur te ontdekken in een ongelabelde dataset. De meest voorkomende taak in ongesuperviseerd leren is clustering, het ontdekken van coherente groepen. Andere taken zijn anomaliedetectie en primaire componenten analyse.

**Voorbeeld** marktsegmentatie en nieuwsaggregatie zijn voorbeelden van clustering. Fraudedetectie is een voorbeeld van anomaliedetectie, terwijl dimensiereductie kan gerealiseerd worden mbv primaire componenten analyse.

### 8.2.3 Reinforcement Learning

Reinforcement learning is verschillend van gesuperviseerd en ongesuperviseerd leren in die zin dat er hier niet wordt gewerkt met datasets. Iphv leert de agent van een reeks “beloningssignalen”, die negatief zijn wanneer de agent een “slechte” handeling stelt en positief wanneer hij een goede handeling doet. Reinforcement learning bestaat uit de taak te leren welke acties leiden tot de hoogste totale beloning.

## 8.3 Clustering en het K-Gemiddelden Algoritme

Bij ongesuperviseerd leren is de invoer een ongelabelde dataset met elke we zien dat de dataset ongelabeld is door het ontbreken van de labels y. De taak van het algoritme bestaat erin om structuur te ontdekken in deze ongelabelde dataset.

De meest voorkomende vorm van het ontdekken van structuur is clustering, het vinden van coherente groepen.

De invoer bestaat uit een ongelabelde dataset en het aantal clusters K dat men wenst te ontdekken in de dataset. Elke cluster wordt voorgesteld adhv zijn zwaartepunt. Een element x^(i) van de dataset wordt toegewezen aan de cluster waarvan het zwaartepunt het dichtst bij x(i) ligt. De afstand wordt eenvoudig berekend adhv de Euclidische afstand:

Afbeelding met tekst

Automatisch gegenereerde beschrijving

Het K-gemiddelden algoritme is een iteratief algoritme. De K initiële zwaartepunten worden op een random manier bepaald. Daarna start de hoofdlus van het algoritme. In elke iteratie worden de volgende twee stappen na elkaar uitgevoerd:

* De toewijzingsstap: waarin voor elk punt x^(i) van de dataset wordt berekend tot welke cluster het behoort, er wordt dus bepaald welk zwaartepunt er het dichtst bij x^(i) ligt.
* De update zwaartepuntenstap: waarin op basis van de toekenningen in de vorige stappen het nieuwe zwaartepunt wordt berekent voor elk van de K clusters.

Deze stappen worden herhaalt totdat er in de toewijzingsstap geen wijzigingen meer gebeuren.

Afbeelding met tekst

Automatisch gegenereerde beschrijving

Veronderstel dat we beschikken over deze ongelabelde dataset:

Afbeelding met tafel

Automatisch gegenereerde beschrijving

We werken met K = 2 en we kiezen random de punten



als initiële zwaartepunten. Als we bv berekenen dan vinden we uiteraard dat  terwijl de afstand tot het tweede zwaartepunt gegeven wordt door . Dit betekent dat het eerste punt aan de eerste cluster wordt toegewezen, of c^(1) = 1. We doen dit nu voor elk van de 5 punten:

Afbeelding met tafel

Automatisch gegenereerde beschrijving

Vervolgens worden de nieuwe zwaartepunten u1 en u2 berekend: Afbeelding met tekst

Automatisch gegenereerde beschrijving

Nu wordt de afstand van de vijf punten in de dataset tot deze nieuwe zwaartepunten berekend. Elk punt wordt toegewezen aan de cluster bepaald door het dichtstbijzijnde zwaartepunt.

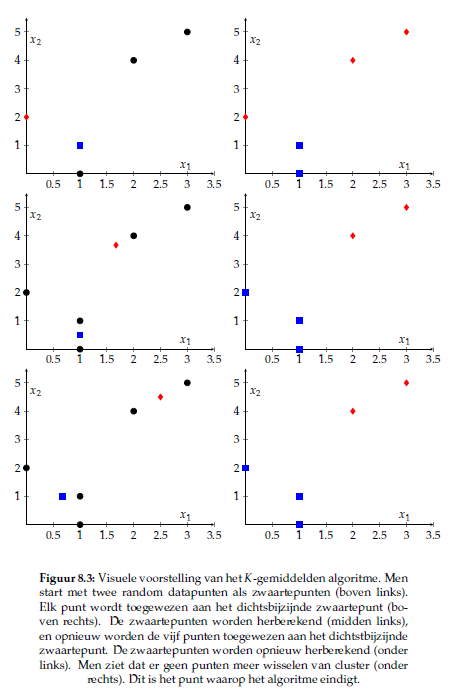
Afbeelding met tafel

Automatisch gegenereerde beschrijving

### 8.3.1 Het K-gemiddelden Algoritme als optimalisatie

In zekere zin is het algoritme ook een optimalisatieprobleem. De kennis ervan kan helpen om eventuele fouten in de implementatie van het K-gemiddelden algoritme op te sporen, maar het kan ook helpen voor het bepalen van het juiste aantal clusters.

Men kan aantonen dat het k-gemiddelden algoritme de gemiddelde kwadratische afwijking tot de zwaartepunten minimaliseert. Onthoud dat c^(i) de index van de huidige cluster van het punt x^(i) voorstelt.



We noemen de kostfunctie J en stellen 

Het K-gemiddelden algoritme zoekt zwaartepunten u^(k) en toewijzingen c^(i) zodanig dat de kostfunctie J minimaal wordt. J kan echter meer minimale lokale minima hebben.

**Voorbeeld:** lokale minima voor de kostfunctie mogelijk zijn beschouwen we volgende dataset:

Afbeelding met tafel

Automatisch gegenereerde beschrijving

De zwaartepunten zijn  . Het algoritme maakte de volgende assignaties:

Afbeelding met tafel

Automatisch gegenereerde beschrijving

De kostfunctie is dus:

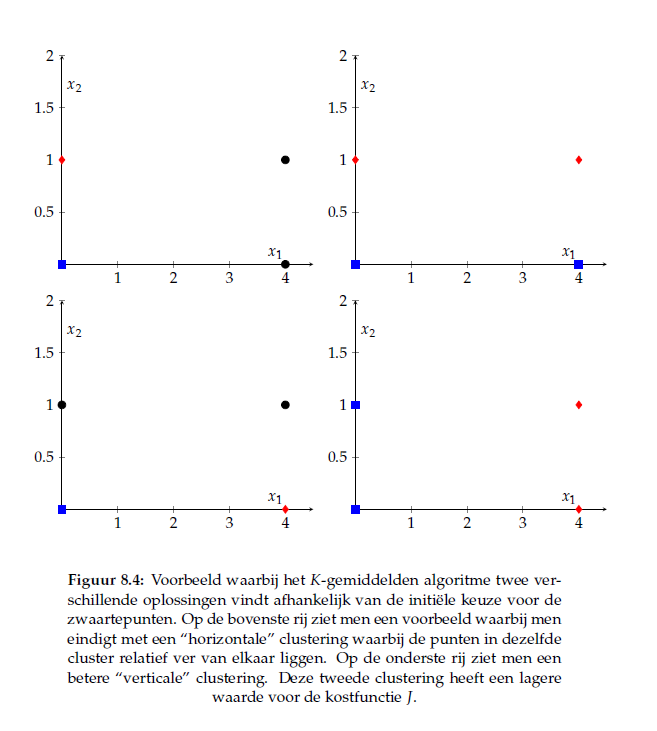
Afbeelding met tekst

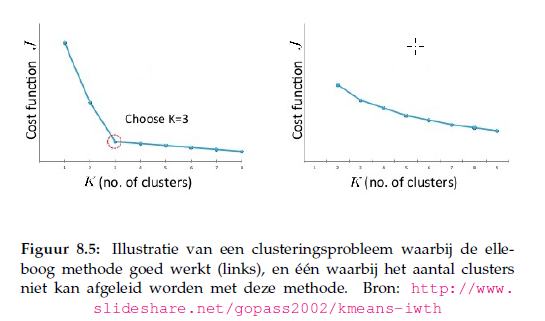
Automatisch gegenereerde beschrijving

Afbeelding met tekst

Automatisch gegenereerde beschrijving

### 8.3.2 Bepalen van het aantal clusters

In sommige gevallen dicteert het probleem het aantal clusters dat men moet ontdekken. 



Wanneer het aantal clusters K toeneemt, dan daalt de minimumwaarde van kostfunctie J.

* Zolang er echte, nuttige clusters worden ontdekt, zal de kostfunctie J snel dalen

**Elleboog methode:** bestaat erin om voor de waarden, K=1,K=2,K=3,… de waarde van de kostfunctie te plotten. Men verwacht in het begin een sterk dalende functie die plots afvlakt.