Werkmiddag

VWO 5 Wiskunde A

Netflix-aanbevelingen:

Het k-means algoritme

Inleiding

Welkom op deze werkmiddag. Je gaat vandaag van 12:40 uur tot 16:00 uur aan de slag met het k-means algoritme. Dit onderwerp ligt op het raakvlak van wiskunde en informatica. Je gaat het k-means algoritme toepassen op een aanbevelingssysteem voor Netflix-films. Wat dit algoritme precies inhoudt en hoe je dit gaat toepassen, zal in de loop van de opdracht duidelijk worden.

De opdracht maak je in tweetallen. Het is niet toegestaan om met andere groepjes over de opdracht te overleggen. Als jullie ergens echt niet uitkomen, mag je uiteraard wel de docent om extra uitleg vragen.

Jullie eindproduct is een getypt verslag waarin de opgaven uit dit boekje zijn uitgewerkt. Het verslag leveren jullie vandaag tussen 15:45 en 16:00 in bij de docent.

Deze opdracht bestaat uit 7 hoofdstukken. Hoofdstuk 1 t/m 4 vormen het voorwerk. Besteed hier niet meer dan 45 minuten aan. In hoofdstuk 5 leer je het k-means algoritme. Dit is de kern van deze werkmiddag. Besteed hier voldoende tijd aan, zo'n 60 tot 90 minuten. Mochten jullie niet uit een bepaalde opgave komen uit hoofdstuk 1 t/m 4, maak je dan niet druk. Kom je niet uit een vraag in hoofdstuk 5, vraag dan om hulp. Probeer wel, zolang de tijd dit toelaat, alle opgaven en schrijf ook over elke opgave iets in jullie verslag (naast het antwoord kan dit bijv. zijn wat jullie geprobeerd hebben, waarom jullie er niet uitkomen, wat voor soort antwoord jullie verwachten dat er uitkomt, etc.).

We raden jullie aan om de opdrachten niet op te delen, dat gaat niet lukken! Samen erdoorheen en telkens overleggen en samen nadenken zal tot de beste antwoorden leiden.

Veel succes met deze opdracht!

1.Introductie

Netflix is een online streamingdienst waarbij consumenten films, series en documentaires kunnen bekijken. Er zijn ruim 100 miljoen gebruikers wereldwijd en het bedrijf maakt elk jaar miljarden dollar omzet. Een belangrijk aspect is dat Netflix in staat is om gebruikers films te adviseren op basis van hun voorkeuren en kijkgeschiedenis.

Vandaag kruip je in de rol van data-analist bij Netflix. Je gaat verschillen en overeenkomsten in de filmvoorkeuren van Netflix-gebruikers onderzoeken. Dit doe je aan de hand van de beoordeling die gebruikers aan een film geven. Kunnen deze beoordelingen gebruikt worden om een aanbevelingssysteem op te zetten?

In 2006 schreef Netflix een prijsvraag uit waarin zij het publiek vroeg om een zo goed mogelijke manier te vinden om de beoordelingen van gebruikers te voorspellen. Hiervoor was er een dataset beschikbaar met maar liefst 100.000 beoordelingen van films. Degene die meer dan 10% verbetering wist te bereiken ten opzichte van de voorspelling van Netflix kon maar liefst 1.000.000 dollar winnen.

Opdracht 1
Welke data denk je dat er gebruikt kan worden om een aanbevelingssysteem op te zetten? Noem ten minste drie variabelen. ■

Elke film in de database van Netflix heeft beoordelingen van gebruikers die de film bekeken hebben. Een film heeft een beoordeling van 0 tot en met 5, waarbij 5 de hoogste beoordeling is. Verder is er informatie beschikbaar over het genre van elke film.

Films in de dataset behoren tot één of meerdere van de volgende genres:

- Action
- Adventure
- Animation
- Comedy
- Children
- Crime
- Documentary
- …..

Opdracht 2
Ga naar https://www.vustat.eu/apps/stat/index.html.

Open het bestand (…).

Bekijk alle beoordelingen voor de film Star Track (2009).
a) Wat is de gemiddelde beoordeling van deze film?

In het bestand (…) staan voor de films die gelabeld zijn met het genre Romantiek en Sciencefiction de gemiddelde beoordelingen die de gebruikers geven aan films met dit genre.
b) Geef deze gemiddelde beoordelingen op een overzichtelijke manier weer in één diagram.

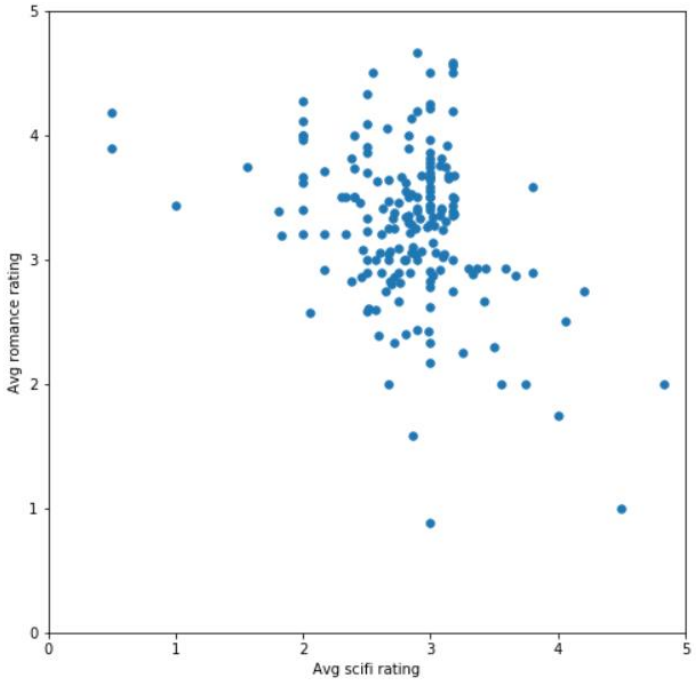
Er wordt het volgende beweerd: een gebruiker met een gemiddelde beoordeling van 5 voor Sciencefiction films is een liefhebber van dit soort films.
c) Geef een argument waarom dit niet juist hoeft te zijn. ■

2. Clustering

Het doel van het clusteren van data is het opdelen van de dataset in groepen bestaande uit soortgelijke items. Deze groepen worden clusters genoemd. In de context van Netflix willen we bijvoorbeeld groepen maken met gebruikers die een gelijke smaak hebben zodat we zo op basis van data uit deze groep films kunnen adviseren aan de gebruikers.

Ter illustratie van het begrip cluster bekijken we onderstaande puntenwolk bestaande uit 200 punten. Elk punt stelt een gebruiker voor en bestaat uit twee coördinaten: een horizontale coördinaat voor de gemiddelde romantiekbeoordeling en een verticale coördinaat voor de gemiddelde sci-fi beoordeling.

Voorbeeld: een gebruiker heeft films met het label romantiek beoordeeld met een gemiddelde van 3 en scifi met een gemiddelde van 3,5. Dit levert het punt $(3\ ;\ 3{,}5)$ op.



Opdracht 3
Stel we deze puntenwolk in twee delen willen splitsen, door bijvoorbeeld een deel van de punten rood te kleuren en een deel groen. Het doel is om het zo op te delen, dat de personen in een groep een enigszins overeenkomstige filmsmaak hebben. Elk punt uit de puntenwolk moet tot één van de twee delen behoren. Er zijn vele manier om deze punten wolk in twee delen op te delen.
a) Geef grafisch weer welke twee groepen jij zou maken en licht toe hoe je elk cluster zou kunnen karakteriseren.

Je kunt ook drie clusters maken.
b) Geef grafisch weer welke drie clusters jij zou maken en licht toe hoe je elk cluster zou kunnen omschrijven.

Wanneer het aantal clusters dat je maakt toeneemt, zal elk cluster gemiddeld uit minder punten bestaan.
c) Noem een voordeel en een nadeel van een groot aantal clusters.

Neem aan dat er geen lege cluster zijn.
d) Wat is het maximaal aantal clusters dat je zou kunnen maken? ■

3.Afstandsmaten

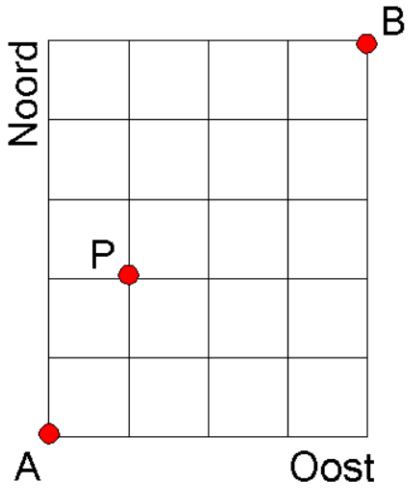
We vragen ons nu af hoeveel de ene film op de andere film lijkt. Om de mate van overeenkomst te bepalen bekijken we de afstand tussen twee punten. We gaan twee verschillende *afstandsmaten* bekijken*:* de Manhattan-afstand en de Euclidische afstand.

Manhattan afstand:
De naam van deze afstandsmaat verwijst naar de roostervormige opzet van de meeste straten op het eiland Manhattan in New York. Een route tussen twee punten is altijd de kortst mogelijk weg tussen twee punten, je loopt dus niet 'om'. Verder is schuin oversteken niet toegestaan, een route gaat altijd over de lijnen in het rooster. De Manhattan-afstand tussen twee punten is nu de lengte van de route (dat is hetzelfde als de optelling van de verticale en horizontale afstand).

Euclidische afstand:
De Euclidische afstand tussen twee punten is de lengte van de kortste verbindingsweg, waarbij nu niet langer over het rooster gelopen hoeft te worden. Er hoeft dus geen gebruik gemaakt te worden van een rooster. De afstand kan berekend worden met behulp van de stelling van Pythagoras.

Opdracht 4

a) Teken in onderstaand rooster een route tussen punt A en punt B waarmee je de Manhattan afstand kunt berekenen en een route waarmee je de Euclidische afstand kunt berekenen.



b) Bereken de Manhattan afstand en de Euclidische afstand tussen punt A en B.

Stel punt C heeft coördinaten (x_C, y_C) en punt D heeft coördinaten (x_D, y_D)

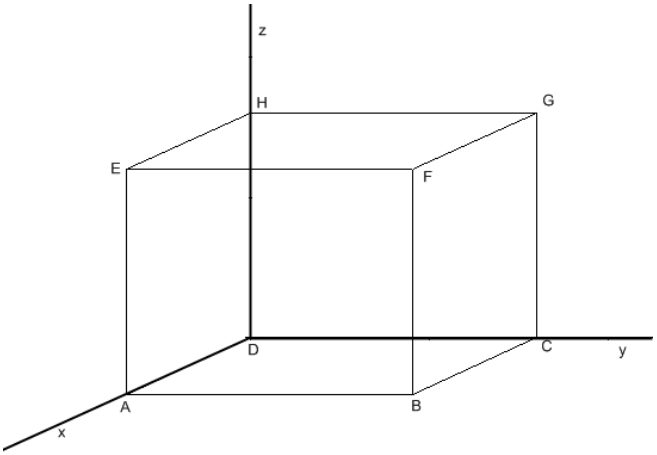
c) Stel een formule op waarmee je de Euclidische afstand kunt bereken tussen $C(x_C, y_C)$ en $D(x_D, y_D)$. ■

Je hebt zojuist twee verschillende afstanden berekend tussen twee punten die beide twee coördinaten hebben. Je kunt ook de afstand berekenen tussen twee punten met meer dan twee coördinaten. Je zou bijvoorbeeld kunnen kijken naar de coördinaten die behoren bij een film. Elk

coördinaat geeft aan of de film tot een bepaald genre behoort (met de afspraak dat een 1 aangeeft dat de film wel tot dit genre behoort en een 0 niet).

We kijken nu een situatie met drie coördinaten. De eerste coördinaat geeft aan of de film tot het genre Animatie behoort, het tweede coördinaat of een film tot het genre Comedy behoort en het derde coördinaat de de film tot het genre Children behoort. De film Toy Story heeft in dit geval de coördinaten $(1,1,1)$.

Drie dimensionale punten kunnen weergeven worden in een *xyz*-assenstelsel. In onderstaand figuur zie je dit assenstelsel. In het assenstelsel is de kubus ABCD.EFGH getekend. De ribben van deze kubus hebben lengte 1.



Opdracht 5

Punt D ligt in de oorsprong.

a) Geef een zelfbedacht voorbeeld van een film waarvan de coördinaten overeenkomen met punt D.
b) Welk van de acht hoekpunten correspondeert met de film Toy Story?
c) Bereken de Euclidische afstand tussen punt D en het punt dat hoort bij Toy Story. ■

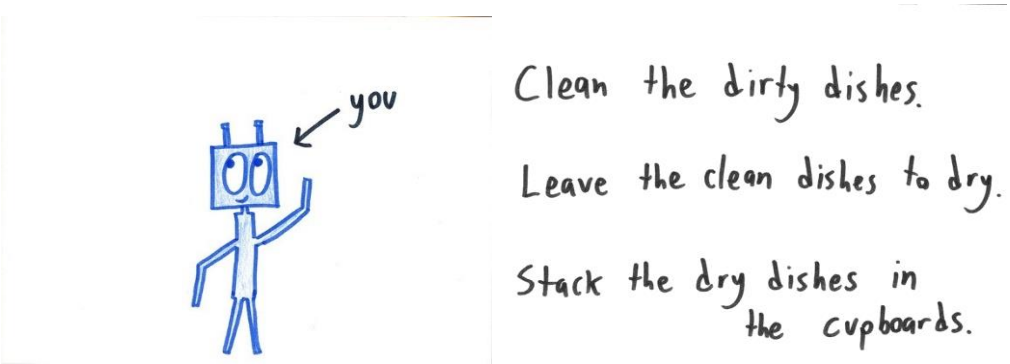
4. Algoritme

Een algoritme is een vaste reeks instructies of stappen die je van een startpunt naar een eindpunt leiden. Voor veel dingen die je doet maak je eigenlijk gebruik van een algoritme. Denk maar eens aan het koken aan de hand van een recept of het doen van de afwas. Ook in de wiskundeles maak je veel gebruik van algoritmes. Een voorbeeld is het opstellen van een raaklijn aan een grafiek. Om dat te doen volg je altijd hetzelfde stappenplan.

Opdracht 6

a) Beschrijf een ander algoritme dat je gebruikt hebt in de wiskundeles.

Het is belangrijk dat de stappen in het algoritme duidelijk zijn.
Stel je voor dat je een huishoudrobot bent. Jouw taak is om de afwas te doen. Hierbij voer je het algoritme uit dat jou gegeven is, zonder daarbij na te denken. In onderstaand figuur vind je het algoritme dat bestaat uit drie stappen.



Er kunnen nogal veel dingen misgaan. In de bijlage staan vier afbeeldingen op chronologische volgorde. In elke afbeelding is het algoritme voor de afwasrobot verbeterd, nadat duidelijk werd dat de robot zijn taak niet zoals verwacht uitvoerde.

b) Geef bij elk afbeelding aan wat de robot fout gedaan zou kunnen hebben, voordat het algoritme aangepast moest worden. ■

5. K-means clustering

Nu je weet wat een algoritme is gaan we één van de bekendste clusteringalgoritmes bestuderen: K-means clustering. Het is een methode die *N* punten verdeelt over *k* clusters. Het doel is om de *N* punten op te delen in *k* clusters, zo dat elk cluster een betekenisvolle groep (films die op elkaar lijken) representeert die we van een goed advies over hun filmkeuze kunnen voorzien. Belangrijk is dat vooraf het aantal clusters bepaald moet worden. Het k-means algoritme is een iteratieve methode, wat aangeeft dat het gaat om een zich herhalend proces. Het algoritme gaat door tot de clusters stabiel zijn, dat wil zeggen dat de inhoud ervan niet meer verandert. Een iteratie is het doorlopen van één cyclus van het proces (stap 3 t/m 5).

Het algoritme werkt met behulp van clustercentrums. Deze clustercentrum worden eerst gekozen. Hoe deze gekozen worden, dat bepaal je zelf of laat je willekeurig bepalen. Het clustercentrum mag een datapunt zijn, maar ook elk ander willekeurig punt. In elke iteratie worden de clustercentrums opnieuw bepaald.

Het *K-means* algoritme werkt als volgt:

Voorbereiding (ook wel initialisatie)
Stap 1. Kies *k,* het aantal clusters dat je wilt maken
Stap 2. Kies de positie van de clustercentrums.

Toewijzing
Stap 3: Wijs elk punt toe aan het cluster waarvan het centrum het dichtst bij is.

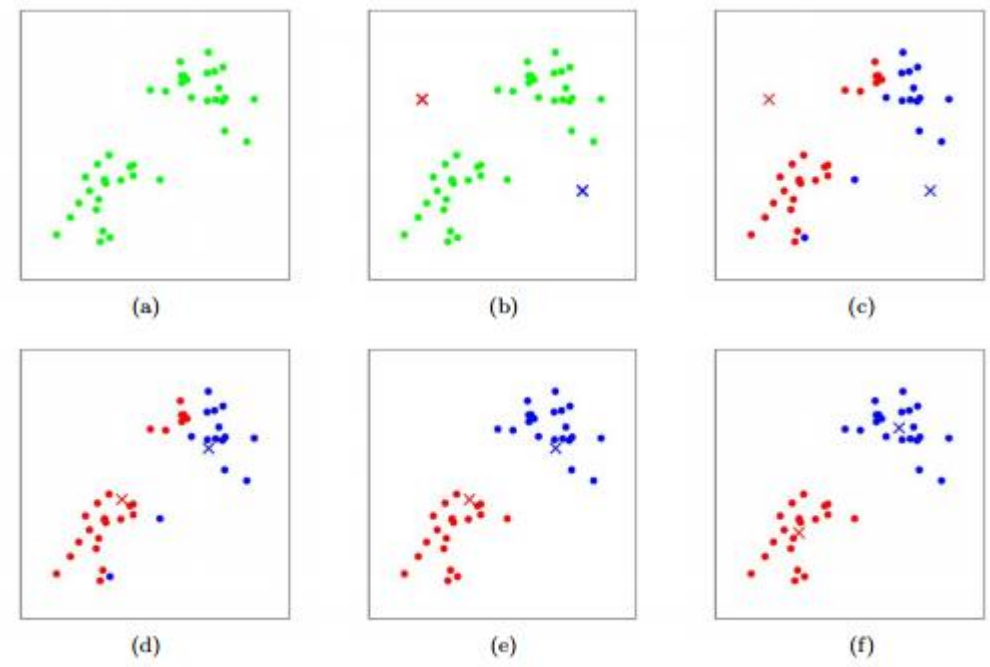
Updaten clustercentrums
Stap 4. Bepaal van elk cluster het nieuwe clustercentrum door voor elke coördinaat het gemiddelde te nemen van alle punten in het cluster.

Controle
Stap 5. Controleer of er clustercentrums zijn die minstens één nieuwe coördinaat hebben.

a. Zo ja; ga terug naar stap 3.
b. Zo nee; je bent klaar.

Voorbeeld

Stap 1 In onderstaand figuur zie je een grafisch voorbeeld van het K-means algoritme met 2 clustercentrums ($k = 2$). In figuur (a) zijn alle datapunten gegeven.



Stap 2. In figuur (b) worden twee clustercentrums gekozen die gegeven worden door een rood kruisje en een blauw kruisje. In dit geval zijn de clustercentrums willekeurig gekozen.
Stap 3. In figuur (c) wordt elk punt toegewezen aan het dichtstbijzijnde clustercentrum. De punten die het dichtst bij het rode clustercentrum zijn worden rood gemaakt, de punten die het dichtst bij het blauwe clustercentrum worden blauw gekleurd.
Stap 4. In figuur (d) zijn de clustercentrums opnieuw berekend.
Stap 5. De locatie van de clustercentrums is veranderd, dus ga weer verder met stap 3.
Stap 3. In figuur (e) worden punten opnieuw toegewezen aan de clustercentrums.

Het algoritme stopt nadat de clustercentrums onveranderd blijven. Dat is na figuur (f).

Opdracht 7

Zorg dat je begrijpt hoe het algoritme werkt door te experimenteren op http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html. Je hoeft over deze opdracht niets in het verslag te schrijven. ■

We bekijken nu een getallenvoorbeeld. Je gaat dit getallenvoorbeeld in de volgende opdracht zelf afmaken.

Gegeven zijn de volgende punten: (1,1), (2,1), (4,3) en (5,4).
Stap 1 Kies k=2.
Stap 2 Kies $(1,1)$ *en* $(2,1)$ *als clustercentrums.*
Stap 3 Bereken voor elk punt de Euclidische afstand tot de clustercentrums en wijs elk punt toe aan het dichtstbijzijnde clustercentrum.

Punt	*Afstand tot cluster 1*	*Afstand tot cluster 2*	*Toegewezen cluster*
(1,1)	*0*	*1*	*1*
(2,1)	*1*	*0*	*2*
(4,3)	$\sqrt{13}$	$\sqrt{8}$	*2*
(5,4)	5	$\sqrt{18}$	*2*

Stap 4
De coördinaten van clustercentrum 1 *zijn nu:* $(1,1)$
De coördinaten van clustercentrum 2 *zijn nu:* $\left(\frac{2+4+5}{3}, \frac{1+3+4}{3}\right) = (3\frac{2}{3}, 2\frac{2}{3})$

Stap 5 Ja, clustercentrum 2 heeft nieuwe coördinaten. Ga weer verder met stap 3.

Opdracht 8
Maak bovenstaand voorbeeld af door het K-means algoritme verder uit te voeren. Doe dit op dezelfde manier als in het voorbeeld. Schrijf alle stappen uitgewerkt op. ■

Opdracht 9
Open het Excel-bestand (). In de puntenwolk staan 17 punten: de blauwe punten zijn de datapunten, de andere kleuren stellen de clustercentrums voor. Er ontbreken nog een aantal formules voordat er een K-means clustering met $k = 3$ uitgevoerd kan worden.

a) Vul in Cel C25, D25 en E25 de formule in waarmee je de afstand tussen punt 1 en het clustercentrum berekent. Trek de formules door naar beneden zodat van elk punt de afstand tot de clustercentra berekend wordt. Schrijf de formule die je in cel A25 hebt ingevoerd op in je verslag.

Hint: bij verwijzingen naar de cellen van het clustercentrum gebruik je $-tekens, daarmee zet je de celverwijzing vast als je de formule doortrekt. Voorbeeld voor cel B20: B20.

b) In cel H25 staat de volgende formule: =ALS(XXX(C25:E25)=C25;"Cluster 1";ALS(XXX(C25:E25)=D25;"Cluster 2";"Cluster 3")). Vul op de plaats van XXX de juiste Excel-functie in. Schrijf in je verslag welke functie je hebt gebruikt.

c) Voer de K-means clustering uit met de initialisatie van de clustercentrums (4,7), (2,9) en (7,5). Neem het diagram dat ontstaan is na convergentie van het algoritme op in je verslag. ■

Opdracht 10
a) Het K-means algoritme stopt nadat de clustercentrums niet meer verschuiven. Je kunt er zeker van zijn dat dit na een eindig aantal stappen gebeurt. Geef een reden waarom je hier zeker van kunt zijn.

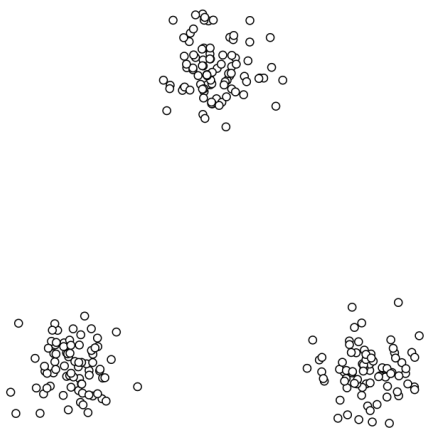
b) Onderzoek met behulp van de website http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html of de clusters die resulteren na uitvoeren van het algoritme afhankelijk zijn van de keuze van de locatie van de clustercentrums in stap 2.

c) Het aantal iteraties voordat convergentie (de clustercentrums veranderen niet meer) optreedt is niet altijd hetzelfde. Noem ten minste drie factoren die van invloed zijn op het aantal iteraties dat nodig is. ■

6. Eigenschappen van K-means clustering

In de volgende reeks opdrachten bekijk je verschillende datasets.

We starten met de Gaussian mixture dataset in figuur X. Natuurlijk is het eigenlijk niet nodig om te clusteren bij dit voorbeeld, op het oog is het al overduidelijk wat de drie clusters zijn: de drie groepjes punten die bij elkaar liggen. Toch gaan we met dit voorbeeld aan de slag en beperken we ons niet tot $k = 3$, omdat deze dataset uitermate geschikt is om een aantal negatieve eigenschappen van het k-means algoritme te onderzoeken waar je rekening mee zou moeten houden als je het k-means algoritme gebruikt bij grotere datasets (waarbij je niet direct ziet hoeveel clusters er zijn en hoe die eruit zouden moeten zien).



Figuur X. Gaussian mixture dataset

Je gaat nu zelf op zoek naar initialisaties waarmee je verschillende eigenschappen van de k-means clustering kunt laten zien. Ga naar de volgende website: https://www.naftaliharris.com/blog/visualizing-k-means-clustering/. Kies in het eerste scherm de optie dat jij zelf de clustercentra wilt initialiseren (‘ I’ll Choose’) en kies vervolgens voor de ‘Gaussian mixture’-verdeling. Je kunt nu zelf clustercentra kiezen door in het venster op een locatie te klikken. In de visualisatie neemt elk datapunt de kleur aan van het cluster waartoe het toegewezen is.

Opdracht 11: Een leeg cluster
Bedenk een initialisatie met $k = 4$ waarbij er na het uitvoeren van het k-means algoritme één van de clusters leeg is (het bevat geen datapunten). Maak een screenshot van het resultaat en zet dat in het verslag. ■

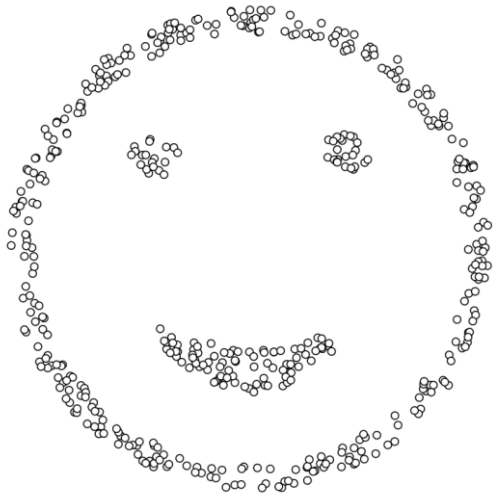
Opdracht 12: Verschillende uitkomsten bij hetzelfde aantal clusters.
Laat zien dat na initialisatie met $k = 3$ en het uitvoeren van het algoritme de volgende twee gevallen kunnen ontstaan:

1. Elk cluster bestaat uit een groepje punten.
2. Eén cluster bestaat uit twee groepjes punten.

Maak van beide gevallen een screenshot. ■

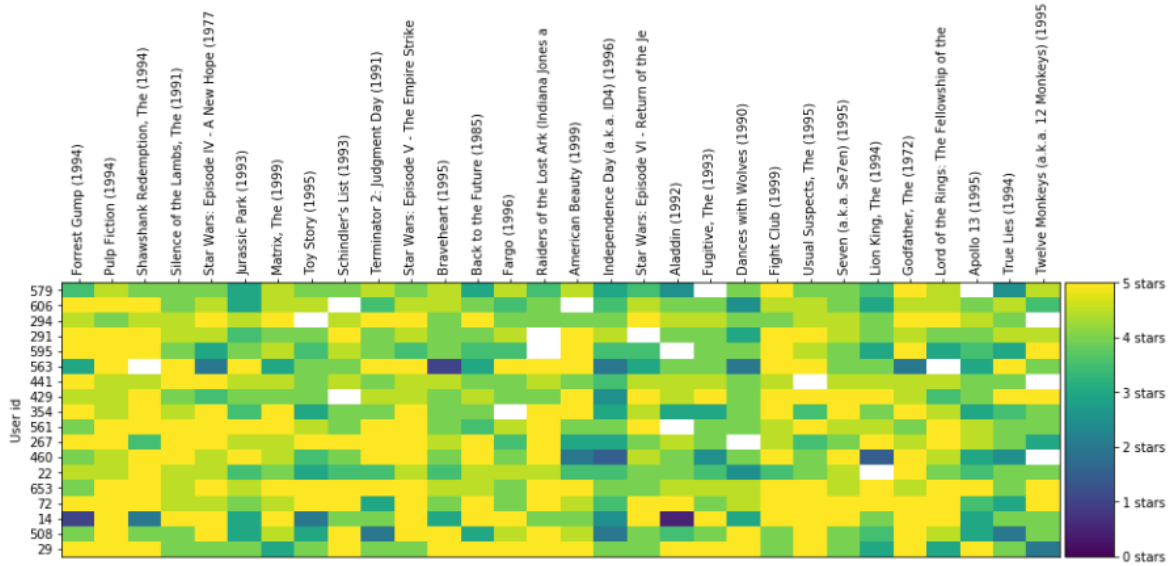
Opdracht 13: Er is geen goede clustering mogelijk
We bekijken nu de Smiley Face dataset. Zie de figuur hieronder.

a) Welke clustering zou jij zonder gebruik te maken van het algoritme doen?
b) Welk deel van het k-means algoritme zorgt ervoor dat er bij de Smiley Face dataset geen goede clustering zal ontstaan? ■



7.Film level clustering

Tot slot bekijken we de resultaten van een k-means clustering bij een grote dataset met gegevens van Netflix. In onderstaand figuur is een deel van de dataset grafisch weergeven. Elke kolom is een film, elke rij een gebruiker. De kleur geeft de beoordeling die de gebruiker aan de film geeft aan. Zie hiervoor de kleurenschaal naast het figuur.

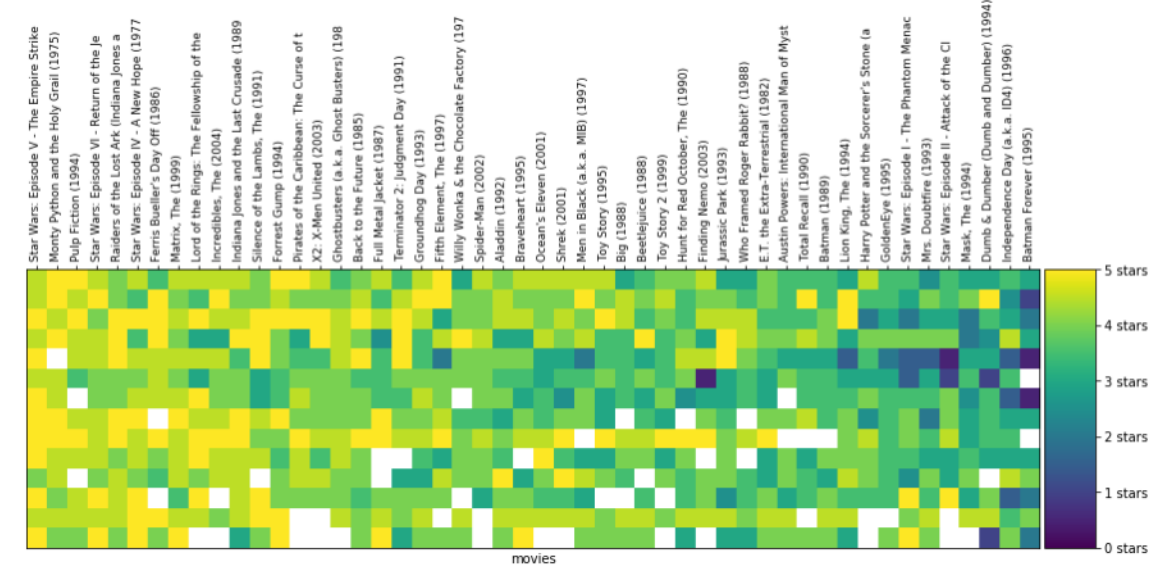


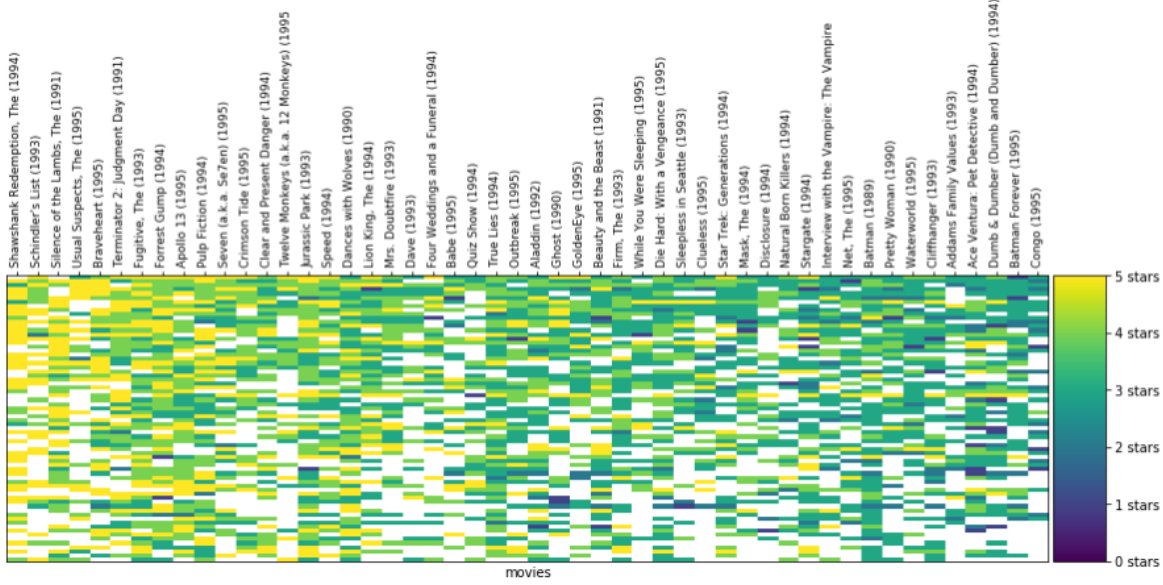
Opdracht 14

a) Er zijn enkele cellen wit. Wat zou dit kunnen betekenen?

Het k-means algoritme wordt toegepast op deze dataset. Zo ontstaan 20 clusters van gebruikers met overeenkomstige filmsmaak. Deze clusters kunnen dan gebruikt worden om films te adviseren aan gebruikers.

In onderstaande twee figuren zie je twee verschillende clusters die zijn ontstaan na toepassing van het k-means algoritme.





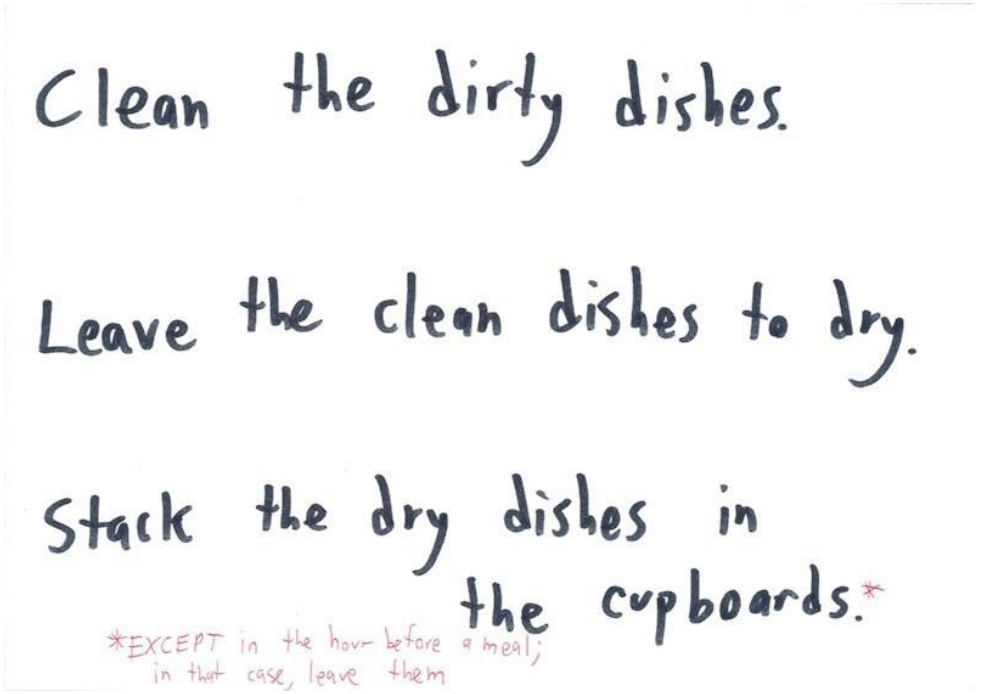
b) Streep in de volgende zin bij de onderstreepte woorden door wat niet van toepassing is: Hoe meer gelijk de smaak is van gebruikers in het cluster, hoe meer horizontale/verticale lijnen in hetzelfde/verschillende kleur er zichtbaar zijn.

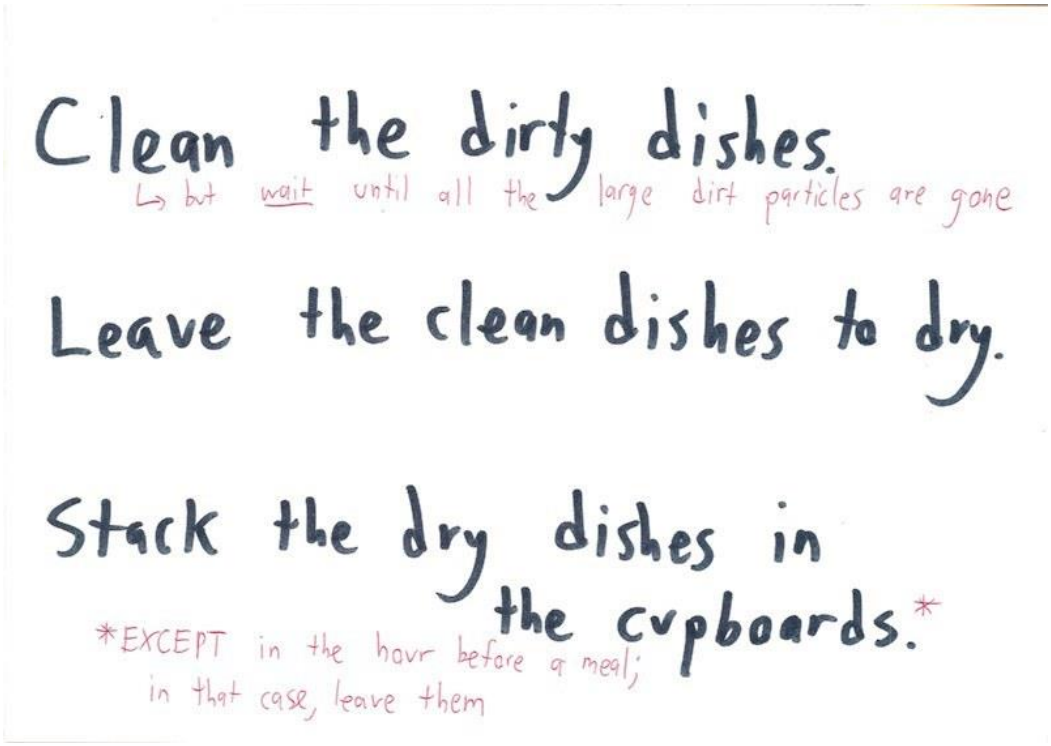
c) Sommige clusters zijn meer geel en andere clusters zijn meer groen/blauw. Wat is hiervan de praktische betekenis?

Er zijn nu clusters van gebruikers met gelijke smaak. Dit kun je gebruiken om films aan te bevelen die gebruikers nog niet hebben gezien.

d) Beschrijf hoe je dit zou doen op basis van de beoordelingen van de films van andere gebruikers in dit cluster. ■

Bijlage bij opdracht 5





Clean the dirty dishes.
↳ but wait until all the large dirt particles are gone*
*OR until removal of particles has stopped

Leave the clean dishes to dry.

Stack the dry dishes in the cupboards.*
*EXCEPT in the hour before a meal;
in that case, leave them

Clean the dirty dishes.
↳ but wait until all the large dirt particles are gone*
*OR until removal of particles has stopped
FOR AT LEAST 5 MINUTES

Leave the clean dishes to dry.
↳ UNLESS it's right before a meal, and there aren't enough dishes, in which case manually dry them

Stack the dry dishes in the cupboards.*
45 minutes
*EXCEPT in the ~~hour~~ before a meal;
in that case, leave them
assuming they're on the table or the counter, ones in the drying rack can go