Uvod u znanost o podacima

Uvod u nenadzirano strojno učenje

Prof. dr. sc. Bojana Dalbelo Bašić

9. predavanje, 14. prosinca 2021.

ak. god. 2021./2022.

Sadržaj

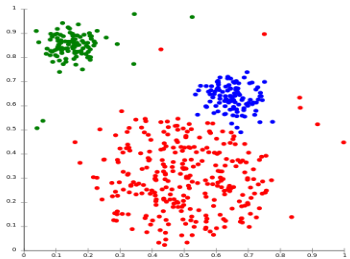
- Uvod i primjeri
- Grupiranje podataka
- Hijerarhijsko grupiranje
- Algoritam k srednjih vrijednosti
- DBSCAN

Strojno učenje

- **Nadzirano:** Dani su parovi ulaz/izlaz (X, y) (tj. uzorak) pomoću kojih tražimo funkciju y = f(X). Naučenu funkciju f evaluiramo na novim podacima. Vrste:
  - **Klasifikacija:** output y je diskretan (oznake klasa)
  - **Regresija:** output y je kontinuiran (linearna regresija)
- **Nenadzirano:** Dani su samo podaci X, oblikujemo funkciju f tako da je y = f(X) *jednostavnija* reprezentacija podataka.
  - **Diskretan y: grupiranje**
  - Kontinuirani y: redukcija dimenzionalnosti (matrična faktorizacija, nenadzirane NN)

Definicija problema grupiranja

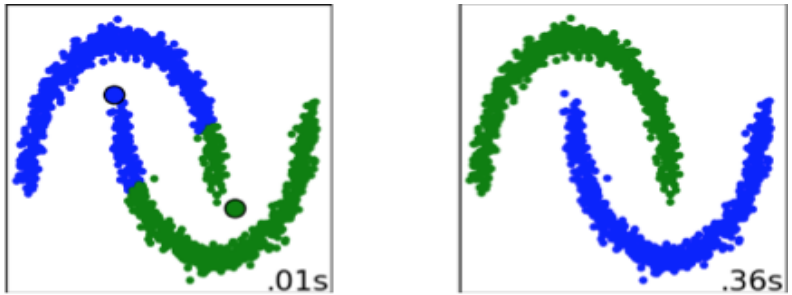
- Ako je dan **skup točaka**, zajedno s pojmom **udaljenosti** između njih, grupiraj točke u neki broj grupa tako da:
  - članovi iste grupe su blizu (tj. slični) jedan drugom
  - članovi različitih grupa su daleko jedan od drugoga
- **Obično je slučaj:**
  - Točke su u visoko dimenzijskom prostoru
  - Sličnost je definirana pomoću mjere udaljenosti
    - Euklidska, cosinus, Jaccard, *edit distance*, …



Svojstva metoda grupiranja

Kvantitativne: skalabilnost (puno uzoraka), dimenzionalnost (puno značajki)

Kvalitativne: tipovi varijabli (numeričke i nenumeričke), oblici (poliedar, hiperravnine i sl.)



Svojstva metoda grupiranja

Robustnost: osjetljivost na redoslijed grupiranja

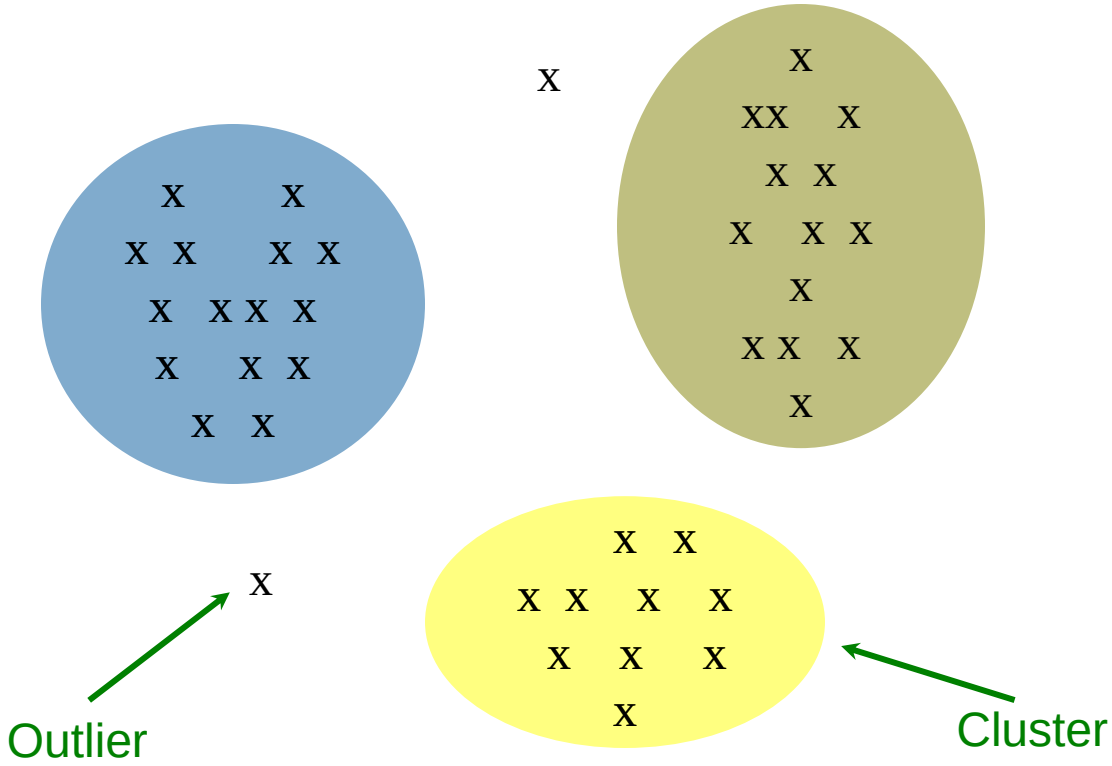
Korisnička interakcija: uvažavanje korisnikovih uvjeta - broj grupa, maksimalna veličina grupe, interpretabilnost, upotrebljivost

Kako interpretiramo ovo grupiranje?

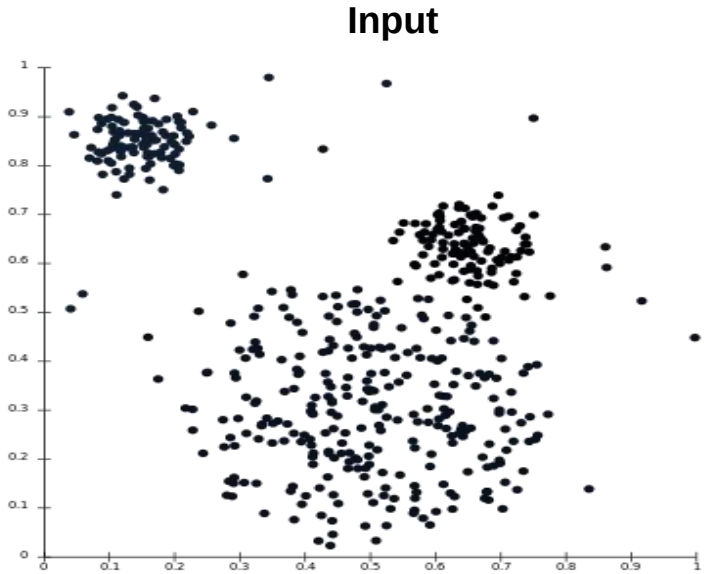
In simplified, seven - dimensional factorial space, we **further summarised original variables** (more than 60% variability) **by clustering** on factor loadings.

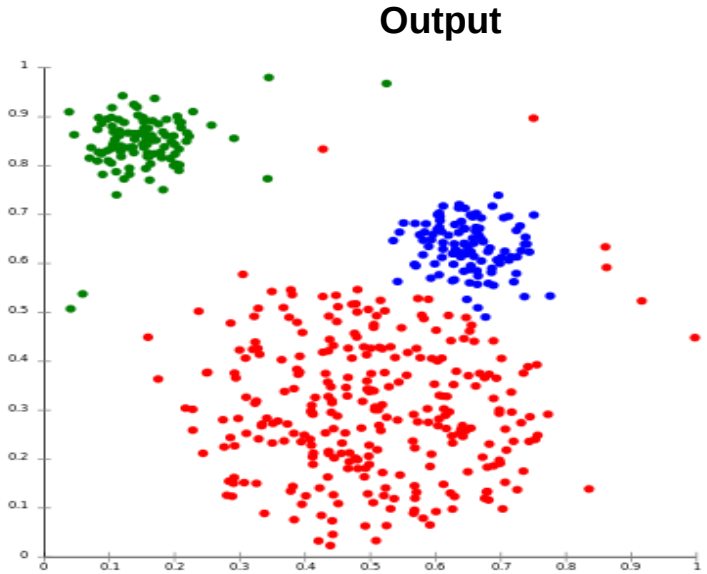
Mala slova – kategorije, Velika slova -EMOCIJE

Primjer: grupe i stršeće vrijednosti



Tipični primjer grupiranja



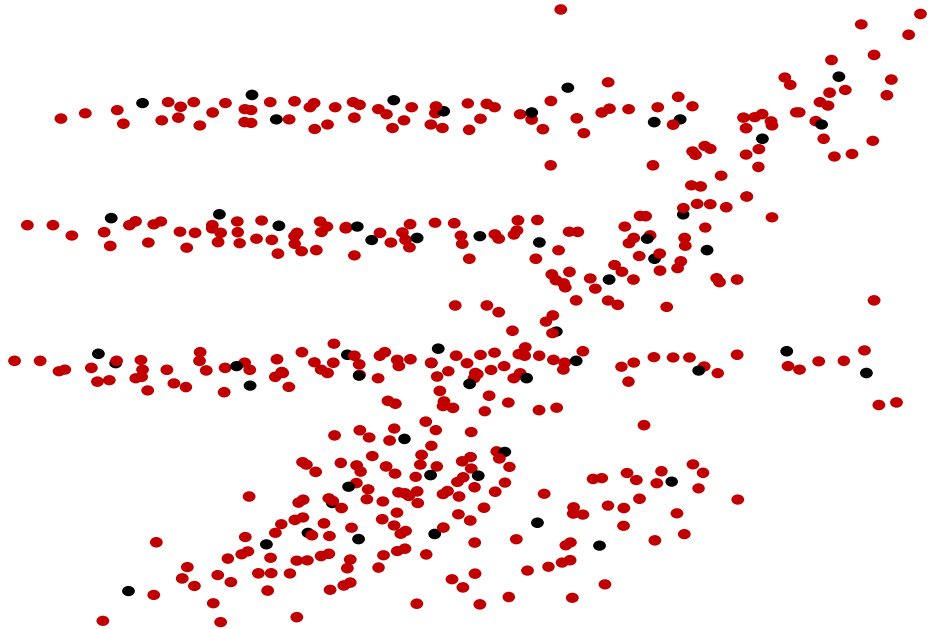


NB: Ovo je jednostavan 2D scenario, obično puno više dimenzija.
Na primjer: 10,000 dimenzija za 100x100 sliku.

Neke primjene grupiranja

- Istraživanje podataka (posebno za visokodimenzijske podatke gdje nema vizualizacije)
- Grupiranje podataka za slijedeću analizu
- Marketing: izrada profila grupe korisnika
- Potpora označavanju podataka za nadzirano učenje
- Sažimanje podataka …

Grupiranje za sažimanje/kompresiju



U ovom slučaju ne zahtijevamo da grupiranje oblikuje strukturu nego da daju grubu sliku podataka

Budite svjesni “pristranosti grupiranja”!

- Ljudi konceptualiziraju svijet kroz predstavljene primjere *exemplars* (Rosch 1973, Estes 1994).



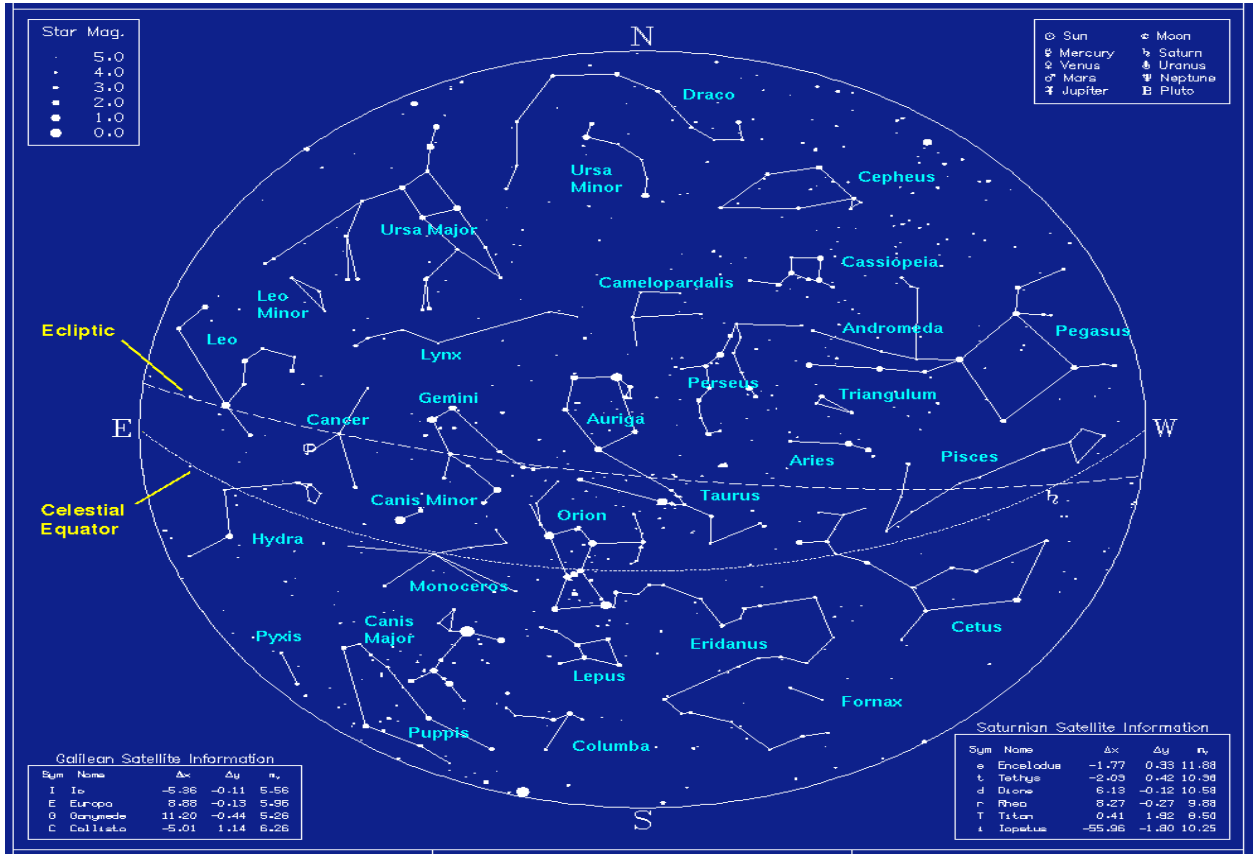
- Skloni smo vidjeti strukturu neovisno o tome je li prisutna ili nije
- OK na slikama pasa, ali …

Pristranost grupiranja

This is the **clustering illusion bias**. It is the tendency to "*erroneously consider the inevitable "streaks" or "clusters" arising in small samples from random distributions, to be non-random.*" (**Wikipedia**)

In London, during World War II, Germany was sending V1 every day …

Pristranost grupiranja



Pristranost grupiranja

- **Grupiranje je korišteno više nego bi trebalo biti** zato jer ljudi pretpostavljaju da domena od interesa ima diskretne klase
- Posebno istinito za osobine ljudi, napr. *Myers-Briggs personality types*
- Često su u stvarnosti podaci **kontinuirani**.

What's Your Personality Type?

Use the questions on the outside of the chart to determine the four letters of your Myers-Briggs type.
For each pair of letters, choose the side that seems most natural to you, even if you don't agree with every description.

1. Are you outwardly or inwardly focused? If you:

- Could be described as talkative, outgoing
- Like to be in a fast-paced environment
- Tend to work out ideas with others, think out loud
- Enjoy being the center of attention

then you prefer

E
Extraversion

- Could be described as reserved, private
- Prefer a slower pace with time for contemplation
- Tend to think things through inside your head
- Would rather observe than be the center of attention

then you prefer

I
Introversion

2. How do you prefer to take in information? If you:

- Focus on the reality of how things are
- Pay attention to concrete facts and details
- Prefer ideas that have practical applications
- Like to describe things in a specific, literal way

then you prefer

S
Sensing

- Imagine the possibilities of how things could be
- Notice the big picture, see how everything connects
- Enjoy ideas and concepts for their own sake
- Like to describe things in a figurative, poetic way

then you prefer

N
Intuition

3. How do you prefer to make decisions? If you:

- Make decisions in an impersonal way, using logical reasoning
- Value justice, fairness
- Enjoy finding the flaws in an argument
- Could be described as reasonable, level-headed

then you prefer

T
Thinking

- Base your decisions on personal values and how your actions affect others
- Value harmony, forgiveness
- Like to please others and point out the best in people
- Could be described as warm, empathetic

then you prefer

F
Feeling

4. How do you prefer to live your outer life? If you:

- Prefer to have matters settled
- Think rules and deadlines should be respected
- Prefer to have detailed, step-by-step instructions
- Make plans, want to know what you're getting into

then you prefer

J
Judging

- Prefer to leave your options open
- See rules and deadlines as flexible
- Like to improvise and make things up as you go
- Are spontaneous, enjoy surprises and new situations

then you prefer

P
Perceiving

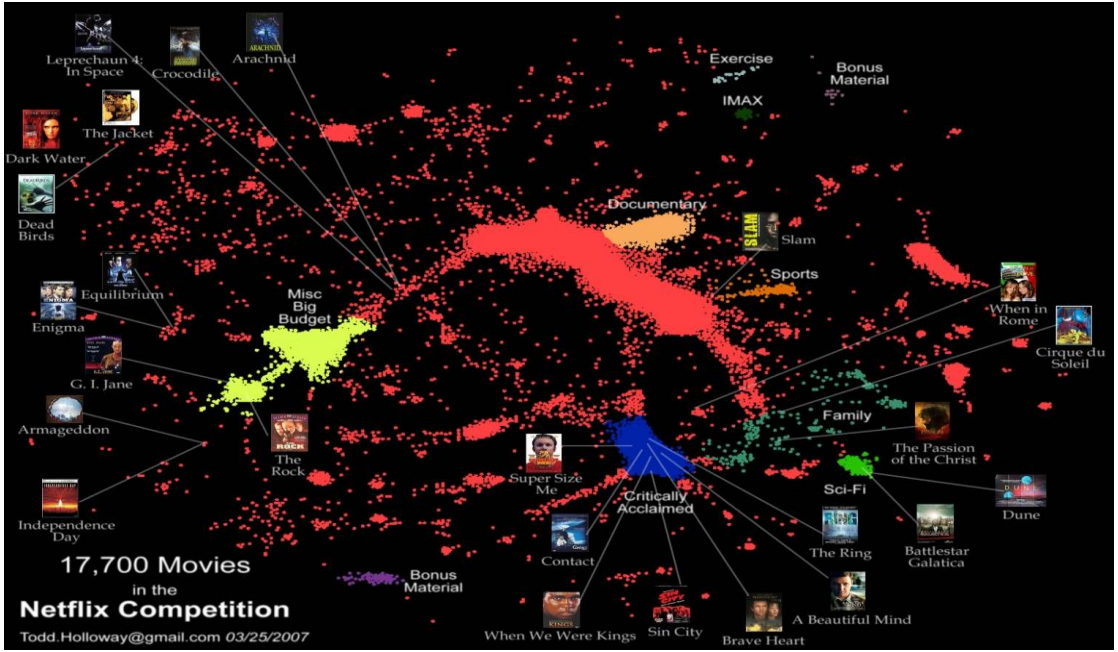
ISTJ Responsible, sincere, analytical, reserved, realistic, systematic. Hardworking and trustworthy with sound practical judgment.	**ISFJ** Warm, considerate, gentle, responsible, pragmatic, thorough. Devoted caretakers who enjoy being helpful to others.	**INFJ** Idealistic, organized, insightful, dependable, compassionate, gentle. Seek harmony and cooperation, enjoy intellectual stimulation.	**INTJ** Innovative, independent, strategic, logical, reserved, insightful. Driven by their own original ideas to achieve improvements.
ISTP Action-oriented, logical, analytical, spontaneous, reserved, independent. Enjoy adventure, skilled at understanding how mechanical things work.	**ISFP** Gentle, sensitive, nurturing, helpful, flexible, realistic. Seek to create a personal environment that is both beautiful and practical.	**INFP** Sensitive, creative, idealistic, perceptive, caring, loyal. Value inner harmony and personal growth, focus on dreams and possibilities.	**INTP** Intellectual, logical, precise, reserved, flexible, imaginative. Original thinkers who enjoy speculation and creative problem solving.
ESTP Outgoing, realistic, action-oriented, curious, versatile, spontaneous. Pragmatic problem solvers and skillful negotiators.	**ESFP** Playful, enthusiastic, friendly, spontaneous, tactful, flexible. Have strong common sense, enjoy helping people in tangible ways.	**ENFP** Enthusiastic, creative, spontaneous, optimistic, supportive, playful. Value inspiration, enjoy starting new projects, see potential in others.	**ENTP** Inventive, enthusiastic, strategic, enterprising, inquisitive, versatile. Enjoy new ideas and challenges, value inspiration.
ESTJ Efficient, outgoing, analytical, systematic, dependable, realistic. Like to run the show and get things done in an orderly fashion.	**ESFJ** Friendly, outgoing, reliable, conscientious, organized, practical. Seek to be helpful and please others, enjoy being active and productive.	**ENFJ** Caring, enthusiastic, idealistic, organized, diplomatic, responsible. Skilled communicators who value connection with people.	**ENTJ** Strategic, logical, efficient, outgoing, ambitious, independent. Effective organizers of people and long-range planners.

Pristranost grupiranja

- Da li se potpuno nalazite u jednoj kategoriji ili smatrate da ste djelomično u jednoj, a djelomično u drugoj ili ponekad u jednoj, a ponekad u drugoj?

- U takvim slučajevima, kontinuirani modeli daju bolje rezultate (matrična faktorizacija, meko grupiranje…)

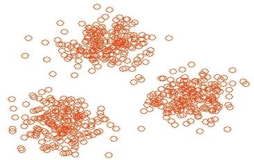
Netflix



- Više kontinuirane strukture nego diskretne
- Druge metode (matrična faktorizacija, k-NN) mogu dati bolji opis strukture nego diskretne grupe

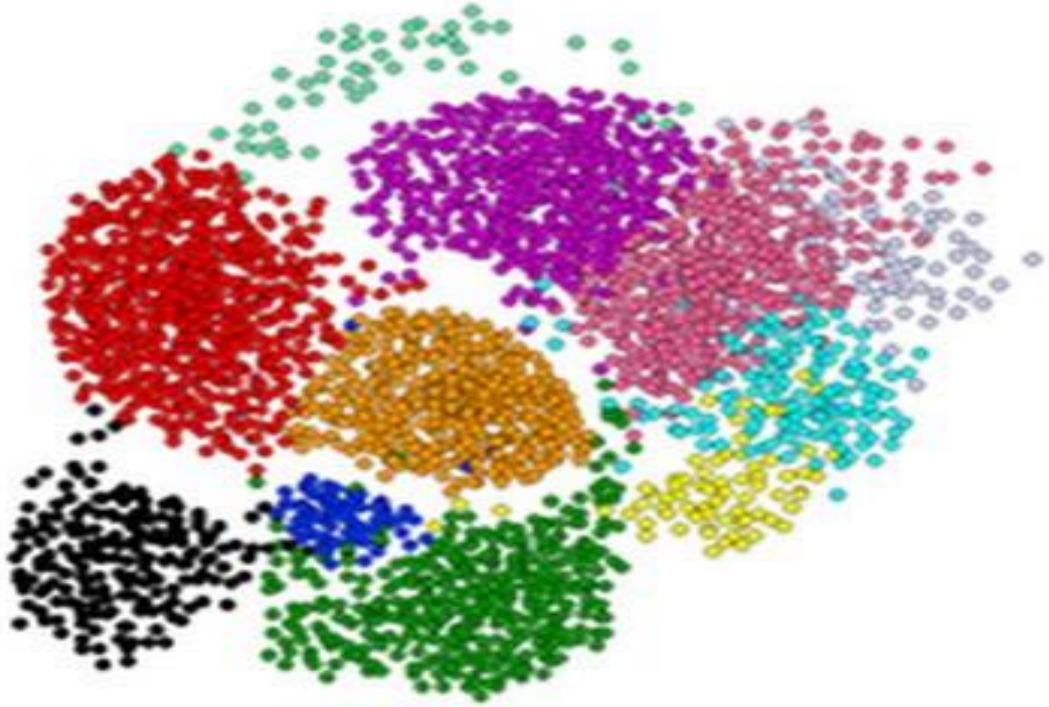
Vrste grupiranja

- **Hijerahijsko grupiranje:** grupe stablaste hijerarhijske strukture. Može se računati *bottom-up* ili *top-down*.
- **Ravno grupiranje:** nema strukture između grupa.



- **Čvrsto grupiranje:** svaki objekt pripada samo jednoj grupi
- **Meko grupiranje:** pripadnost grupi se modelira pomoću teorije vjerojatnosti ili neizrazitom logikom

Grupiranje je težak problem!



Zašto?

Grupiranje je težak problem!

- Grupiranje u dvije dimenzije izgleda lako
- Grupiranje malo podataka izgleda lako
- I u takvim specijalnim slučajevima jest uglavnom lako,
- ali...

Grupiranje je težak problem!

- Grupiranje u 2 dimenzije izgleda lako
- Grupiranje malo podataka izgleda lako
- I u takvim specijalnim slučajevima jest uglavnom lako,
- ali...
- … većina primjena podrazumijeva ne dvije, nego 10 ili 10,000 dimenzija (i velike količine podataka)

Grupiranje je težak problem!

- Grupiranje u 2 dimenzije izgleda lako
- Grupiranje malo podataka izgleda lako
- I u takvim specijalnim slučajevima jest uglavnom lako,
- ali...
- … većina primjena podrazumijeva ne 2, nego 10 ili 10,000 dimenzija (i velike količine podataka)
- **Visokodimenzionalni prostori izgledaju drugačije: Gotovo svi parovi točaka su otprilike jednako udaljeni** (“*Curse of dimensionality*”, predavanje 7)

Problem grupiranja: galaksije

- Katalog od 2 milijarde nebeskih objekata predstavljaju objekte u 7 dimenzija (frekvencije)
- Problem: Grupiranje sličnih objekata, e.g., galaksije, obližnje zvijezde, kvazari etc.
- Sloan Digital Sky Survey [link]

Problem grupiranja: muzika na CD-ovima

- **Intuitivno:** Muzika se dijeli u kategorije, a slušatelji preferiraju neke kategorije
  - Što su kategorije zapravo?
  - —> *take a data-driven approach*!

- Predstavljanje CD-a sa skupom kupaca tog CD-a (“*collaborative filtering*”)
- Slični CD-i imaju slične kupce i *vice-versa*

Problem grupiranja: muzika na CD-ovima

Prostor svih CD-ova:

- Prostor – jedna dimenzija za svakog kupca
  - Vrijednosti mogu biti samo između 0 or 1
  - CD je točka u prostoru (x_1, x_2,…, x_k),
    gdje je x_i = 1 akko je i-ti kupac kupio CD

- Za Amazon, dimenzionalnost desetine miliona

- **Zadatak:** Naći grupe sličnih CD-ova

Problem grupiranja: dokumenti

Nalaženje tema:

- Dokument je predstavljen vektorom (x_1, x_2,…, x_k), gdje je x_i = 1 akko se *i*-ta riječ pojavljuje u dokumentu (na bilo kojoj poziciji)

- **Ideja: dokumenti sa sličnim skupovima riječi govore o istoj temi**

Cosinus, Jaccard, Euklidska udaljenost…

U oba primjera (CD, dokumneti) imamo izbor kada razmišljamo o točkama kao skupovima značajki (korisnici, riječi):

- **Skupovi kao vektori: Euklidska, cosinus, …**
- **Skupovi kao skupovi: mjera sličnosti/udaljenosti Jaccard distance**

Mjere udaljenosti

- **Euclidean Distance:** Simplest, fast to compute

$$d(x, y) = \|x - y\|$$

- **Cosine Distance:** Good for documents, images, etc.

$$d(x, y) = 1 - \frac{x \cdot y}{\|x\|\|y\|}$$

- **Jaccard Distance:** For set data:

$$d(X, Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}$$

- **Hamming Distance:** For string data:

$$d(x, y) = \sum_{i=1}^{n} (x_i \neq y_i)$$

Mjere udaljenosti

- **Manhattan Distance:** Coordinate-wise distance

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

- **Edit Distance:** for strings, especially genetic data.

- **Mahalanobis Distance:** Normalized by the sample covariance matrix – unaffected by coordinate transformations.

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^{\top} \mathbf{S}^{-1} (\vec{x} - \vec{y})}.$$

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{s_i^2}}$$

Mjere udaljenosti

I druge mjere, ovisno istraživačkim pitanjima.

Pearsonova **korelacijska udaljenost** (parametarska)

$$r(x, y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2 \sum_{i=1}^{n}(y_i - \mu_y)^2}}$$

Udaljenost d = 1 - r

ili Spearman ili Kendall (neparametarske mjere)

Mjere udaljenosti

I druge mjere, ovisno istraživačkim pitanjima..

Pearsonova **korelacijska udaljenost** (parametarska)

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2 \sum_{i=1}^{n}(y_i - \mu_y)^2}}$$

Udaljenost d = 1 - r

ili Spearman ili Kendall (neparametarske mjere)

- Dva objekta su slična ako su njihove značajke korelirane (iako mogu biti daleko u smislu E. metrike)

Mjere udaljenosti

I druge mjere, ovisno istraživačkim pitanjima..

Pearsonova **korelacijska udaljenost** (parametarska)

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2 \sum_{i=1}^{n}(y_i - \mu_y)^2}}$$

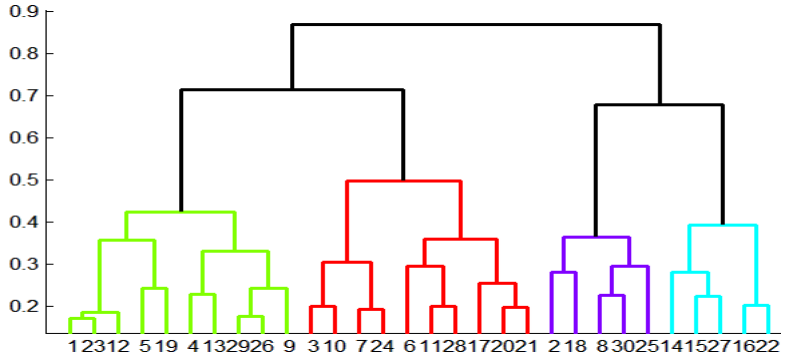
Udaljenost d = 1 - r

ili Spearman ili Kendall (neparametarske mjere)

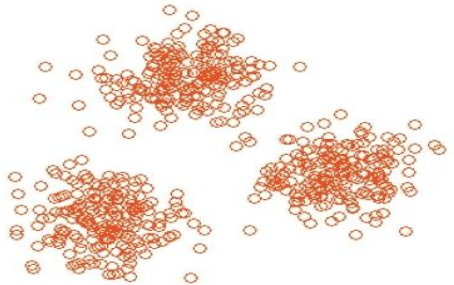
- Dva objekta su slična ako su njihove značajke korelirane
  (iako mogu biti daleko u terminima Euklidske udaljenosti)
- Za identifikaciju grupa objekata sa zajedničkim profilom, bez obzira na magnitudu
  (geni, kupci, *up and down zajedno*)
- Korelacijska udaljenost je osjetljiva na stršeće vrijednosti -> tada bolje neparametrski

Pregled metoda grupiranja

- **Hijerarhijsko (rezultat dendrogram):**
  - **Aglomerativno** (*bottom up*):
  - **Divizivno** (*top down*):



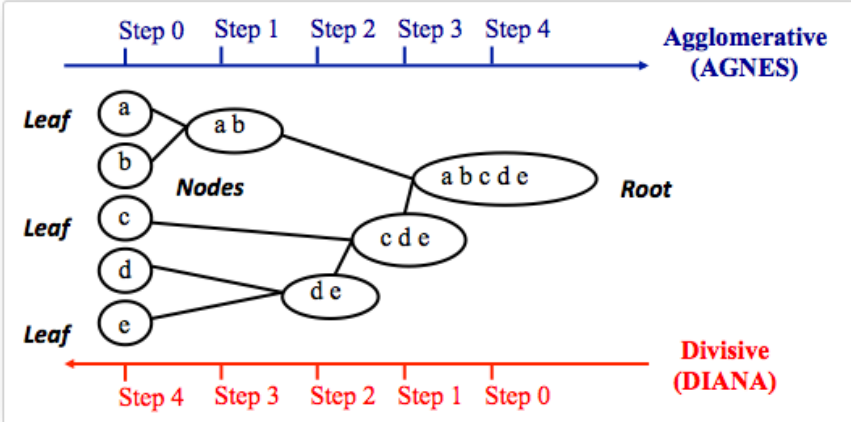
- **Pridruživanje točaka**



KOJE GRUPIRANJE JE DOBRO ZA KOJU SVRHU?

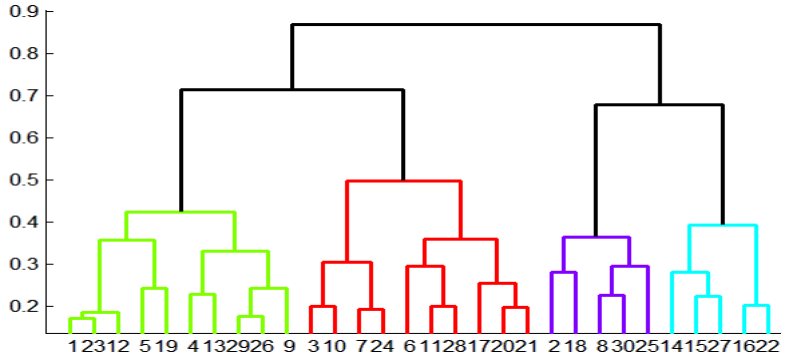
Pregled metoda grupiranja

- **Hijerarhijsko (rez. dendrogram):**
  - **Aglomerativno** (*bottom up*):
    - Inicijalno, svaka točka je grupa
    - Ponavljamo kombiniranje „najbližih” grupa u jednu - do jedne grupe
  - **Divizivno** (*top down*):
    - Počinje jednom grupom i rekurzivno je dijelimo najheterogenije grupe

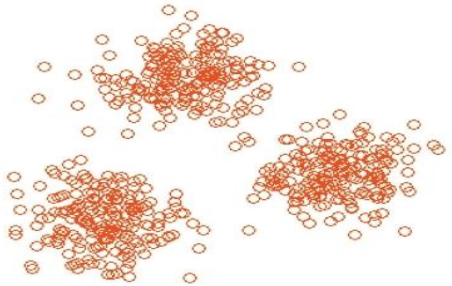


Pregled metoda grupiranja

- **Hijerarhijsko (rez. dendrogram):**
  - **Aglomerativno** (*bottom up*):
    - Inicijalno, svaka točka je grupa
    - Ponavljamo kombiniranje „najbližih” grupa u jednu - do jedne grupe
  - **Divizivno** (*top down*):
    - Počinje jednom grupom i rekurzivno je dijelimo najheterogenije grupe



- **Pridruživanje točaka:**
  - Zadržavamo broj grupa
  - Točke pripadaju najbližoj grupi



Prednosti i mane

PREDNOSTI

- Nema pretpostavke o broju grupa
- Mogu odgovarati prirodnim taksonomijama

MANE

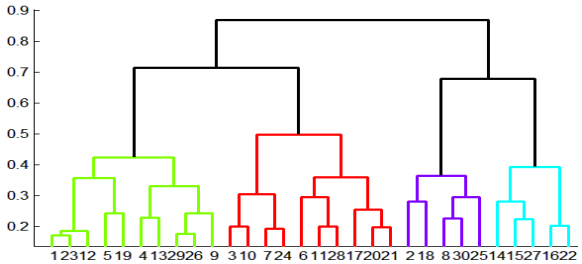
- Jedan put kombinirane grupe ne mogu se raščlaniti
- Sporo za velike skupove podatka

Važan praktični savjet

- Podatke je potrebno standardizirati prije analize!
- Standardizacija omogućuje usporedbu varijabli koje su mjerene na različitim skalama

$$Z = \frac{(x - \bar{x})}{s}$$

```
##            Murder Assault UrbanPop     Rape
## Alabama    1.2426   0.783   -0.521 -0.00342
## Alaska     0.5079   1.107   -1.212  2.48420
## Arizona    0.0716   1.479    0.999  1.04288
## Arkansas   0.2323   0.231   -1.074 -0.18492
## California 0.2783   1.263    1.759  2.06782
## Colorado   0.0257   0.399    0.861  1.86497
```



Koraci u aglomerativnom grupiranju

1. Priprema podataka
2. Računanje udaljenosti između svakog para točaka

   -> što daje…?
3. Korištenje funkcije povezivanja (LINKAGE)
4. Određivanje gdje odrezati stablo i odrediti broj grupa

Koraci u aglomerativnom grupiranju

1. Priprema podataka

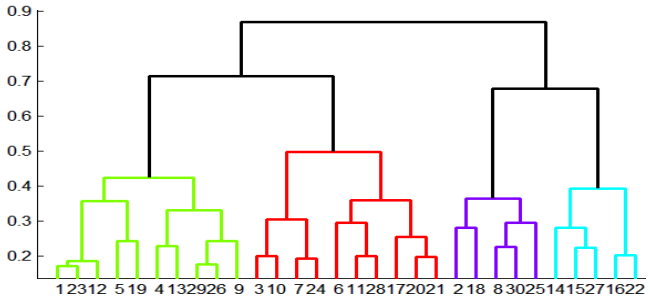
2. Računanje udaljenosti između svakog para točaka -> što daje?

	Alabama	Alaska	Arizona	Arkansas	California	Colorado
Alabama	0.00	2.70	2.29	1.29	3.26	2.65
Alaska	2.70	0.00	2.70	2.83	3.01	2.33
Arizona	2.29	2.70	0.00	2.72	1.31	1.37
Arkansas	1.29	2.83	2.72	0.00	3.76	2.83
California	3.26	3.01	1.31	3.76	0.00	1.29
Colorado	2.65	2.33	1.37	2.83	1.29	0.00

3. korištenje funkcije povezivanja (LINKAGE)

4. Određivanje gdje presiječi stablo i odrediti broj grupa

Agglomerativno hijerarhijsko grupiranje



- **Ključna operacija:**

  Ponavljano kombiniranje bliskih grupa

  (linkage)

- **Tri važna pitanja:**
  - **1)** Kako predstaviti grupu sa više od tri točke?
  - **2)** Kako odrediti „bliskost” grupa?
  - **3)** Kada prestati kombinirati grupe?

Definiranje bliskosti grupa

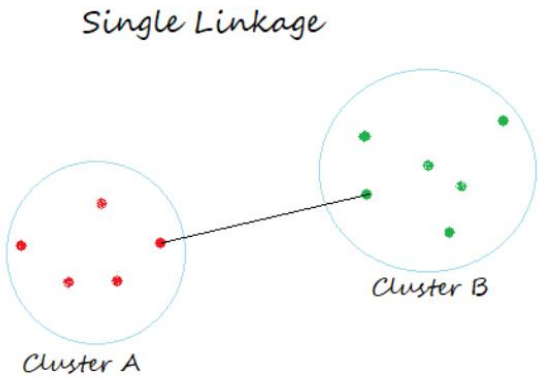
Pristupi :

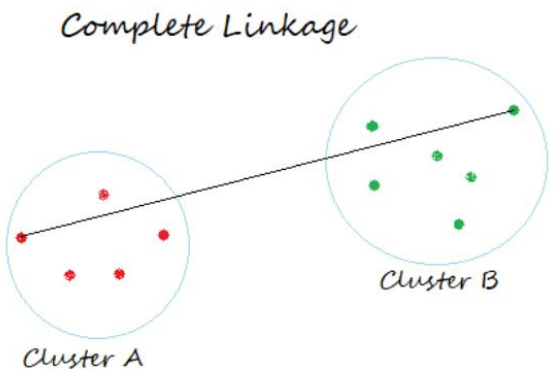
1. ***Međugrupna*** **udaljenost** = minimum/maksimum/prosjek udaljenosti svih točaka u grupi

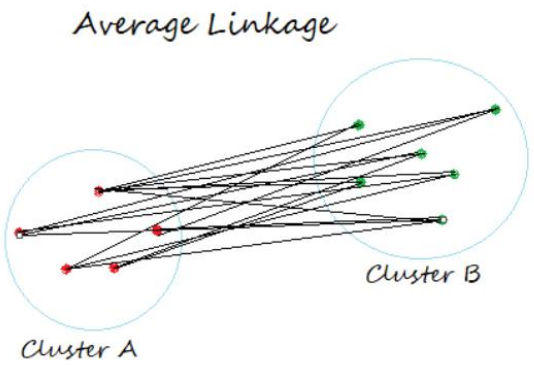
2. ***Pojam kohezije*** **(povezanosti unutar grupe)**

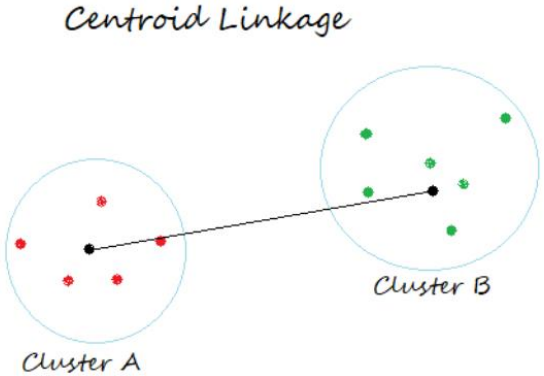
Vrste povezivanja grupa (*linkage*)

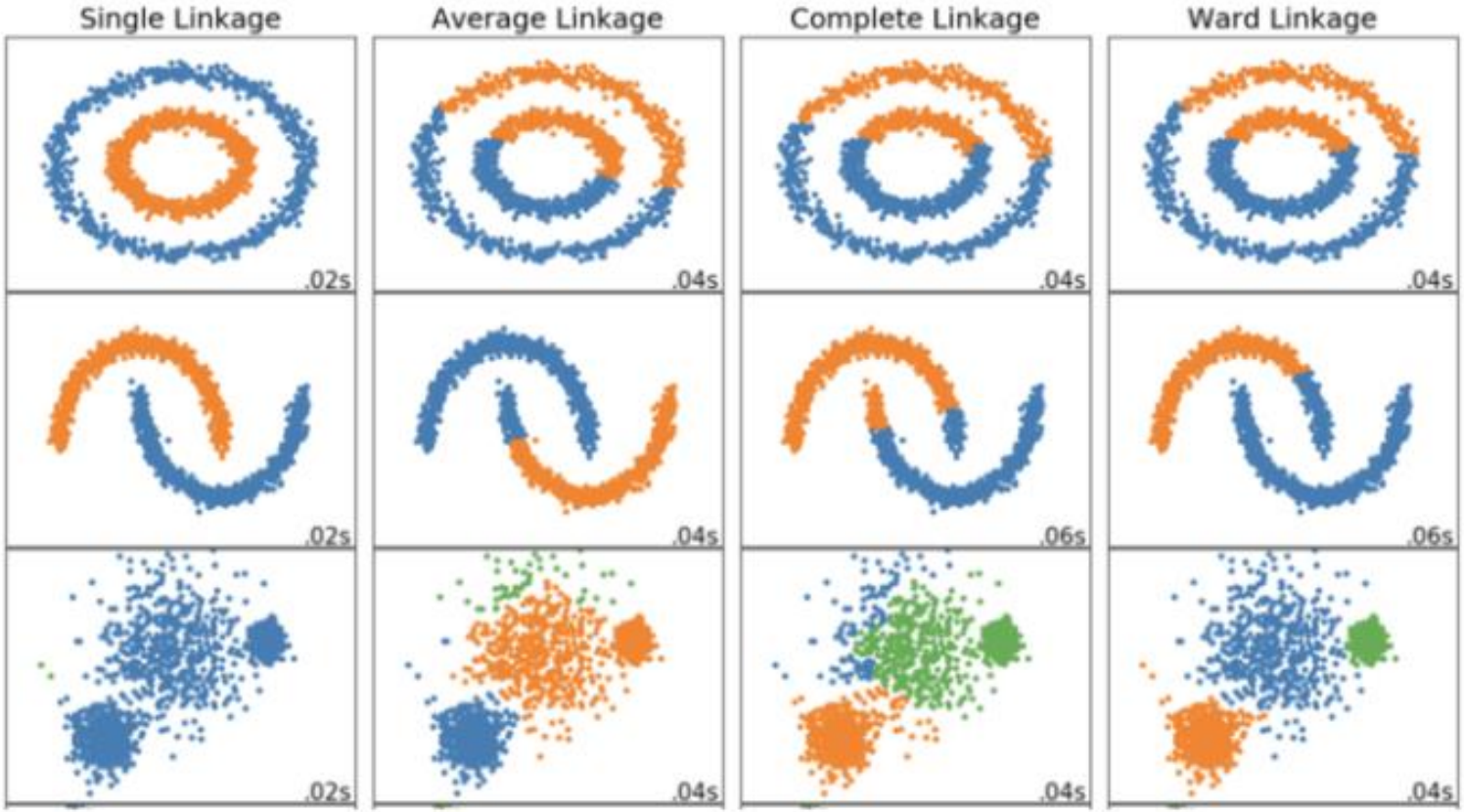
1. Maksimum ili *complete linkage*
2. Minimum ili *single linkage* – (izdužene grupe)
3. Srednja vrijednost ili *average linkage*
4. Centroid ili *centroid linkage*
5. Wardova metoda ili *minimum variance method*


Single Linkage
Cluster B
Cluster A

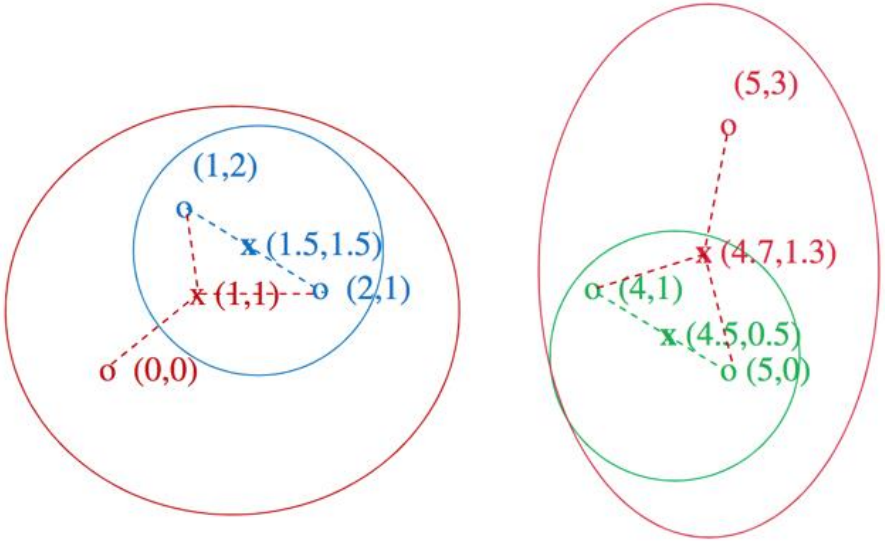

Complete Linkage
Cluster B
Cluster A


Average Linkage
Cluster B
Cluster A

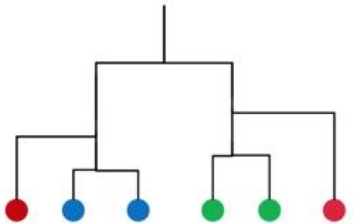

Centroid Linkage
Cluster B
Cluster A


Single Linkage
Average Linkage
Complete Linkage
Ward Linkage
.02s
.04s
.04s
.04s
.02s
.04s
.06s
.06s
.02s
.04s
.04s
.04s

Primjer: Hijerarhijsko grupiranje



Data:
o … data point
x … centroid



Dendrogram

Dendrogram u potpunosti omogućuje rekonstrukciju slijeda grupiranja

Što je s kombinacijom kvalitativnih i kvantitativnih podataka?

Gower udaljenost

$$d(x,y) = \frac{\sum_{j=1}^{n} d_{x,y}^{j}}{n}$$

gdje je za norminalne i binarne (dihotomne) varijable

$$d_{x,y}^{j} = \begin{cases} 1, & x_j = y_j \\ 0, & x_j \neq y_j \end{cases}$$

te za numeričke i ordinalne varijable

$$d_{x,y}^{j} = 1 - |x_j - y_j|$$

Implementacija

- **Naivna implementacija hijerarhijskog grupiranja**

U svakom koraku izračunaj udaljenost između svakog para točaka, zatim spoji

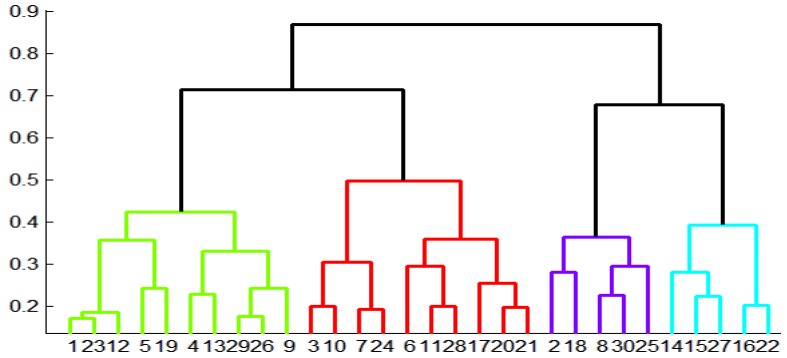
  - $O(N^3)$, gdje je N broj točaka

- **Pažljivo implementiraj koristeći red prioriteta – to može reducirati** vrijeme na $O(N^2 \log N)$
  - Još uvijek preskupo za zaista velike skupove podataka koji ne stanu u memoriju

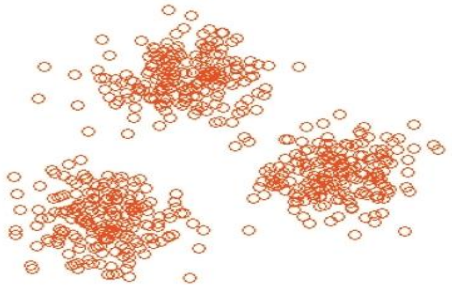
https://www.cs.utah.edu/~jeffp/teaching/cs5140-S17/cs5140/L8-hierarch.pdf

Pregled metoda grupiranja

- **Hijerarhijsko:**
  - **Aglomerativno** (*bott*
    - Inicijalno, svaka … grupa
    - …avljamo … iranje „najbližih”
    - … je
  - **Divizivn** … *n*):
    - Počinje … nom grupom i rekurzivno je dijelimo
- **Pridruživanje točaka:**
  - Zadržavamo broj grupa
  - Točke pripadaju najbližoj grupi



sljedeće



Algoritam k-srednjih vrijednosti

Najvažniji među algoritmima grupiranja koji se temelje na pridruživanju točaka

Algoritam k-srednjih vrijednosti

- Cilj: pridružiti svaku točku jednom od ***k*** grupa tako da je udaljenost točaka od centroida minimalna.
- Jednostavan, pohlepan algoritam, optimizira srednju udaljenost članova u grupi

$$\sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} \|x_{ki} - x_{kj}\|^2$$

Algoritam k-srednjih vrijednosti

- Cilj: pridružiti svaku točku jednom od **k** grupa tako da je udaljenost točaka od centroida minimalna.
- Jednostavan, pohlepan algoritam, optimizira srednju udaljenost članova u grupi

$$\sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} \|x_{ki} - x_{kj}\|^2$$

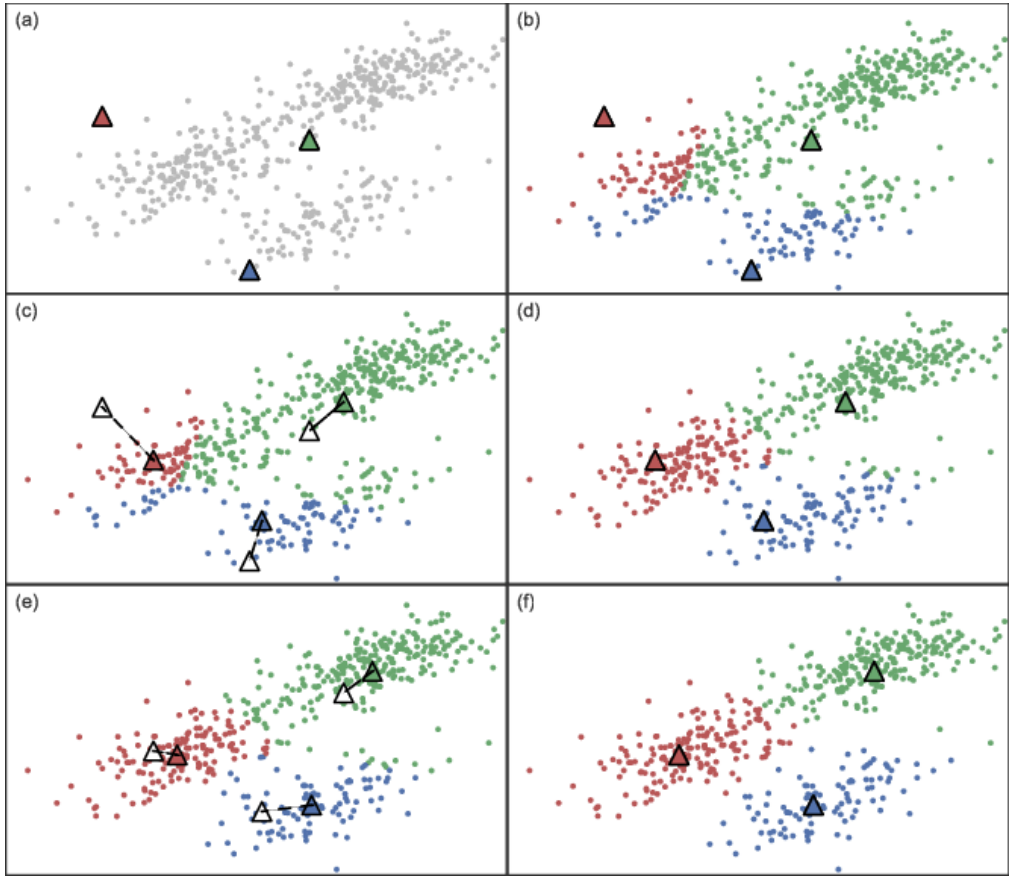
- Minimizira udaljenost (kvadrat Euklidske udalj.) od podatka do centroida

$$SSE = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|x_{ki} - \mu_k\|^2$$

Algoritam k-srednjih vrijednosti

Nađi najbliži centroid grupe za svaki element i pridruži element toj grupi

Ponovno izračunaj nove centroide



Algoritam k-srednjih vrijednosti

Koliko dugo iterirati?

- Fiksni broj iteracija
- ili dok nema promjena u pridruživanju
- ili dok su samo male promjene u povezanosti grupa (suma kvadrata udaljenosti od svake točke do centroida).

Algoritam k-srednjih vrijednosti

Algoritam k-srednjih vrijednosti (engl. k-means) sastoji se od niza koraka:

1. Izaberi broj grupa k.
2. Inicijaliziraj *k* centara klastera (slučajnim odabirom).
3. Svaki od n objekata pridruži najbližem centroidu.
4. Promijeni centre klastera pretpostavljajući da su objekti stavljeni u točne klastere.

……

Inicijalizacija

Za početak potrebno je k točaka:

- **Slučajan uzorak k točaka iz skupa**

- **Algoritam k-means++:** Iterativno konstruiraj slučajni uzorak sa dobrim razmakom

Nalaženje optimalnog *k-means* grupiranja je NP-težak problem.

K-means++ [link]

Početak: Prvi centar grupe (od njih *k*) odaberi slučajno

Iteriraj:

- Za svaku drugu točku x, izračunaj udaljenost od x do najbližeg prethodno izabranog centra, D(x).

- Novi centar grupe je slučajno odabrana točka s vjerojatnošću proporcionalnom s $D(x)^2$.

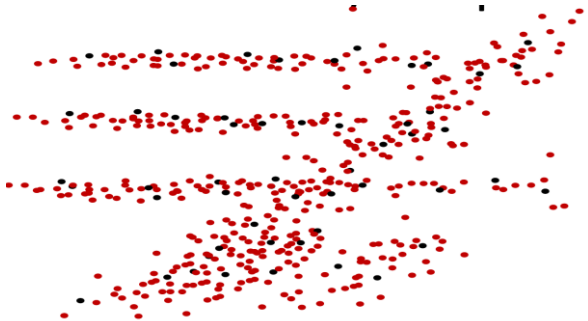
Svojstva algoritma k-srednjih vrijednosti

- Pohlepan algoritam sa elementima slučajnosti – rješenje nije optimalno i znatno varira s ovisno o inicijalnim uvjetima.
- Jednostavan dokaz konvergencije.
- **Složenost je O(nk) po iteraciji** — nije loše, može se unaprijediti heuristikama

Svojstva algoritma k- srednjih vrijednosti

- Pohlepan algoritam sa elementima slučajnosti – rješenje nije optimalno i znatno varira s ovisno o inicijalnim uvjetima.
- Jednostavam dokaz konvergencije.
- **Složenost je O(nk) po iteraciji** — nije loše, može se unaprijediti heuristikama

- Puno varijanti, na primjer:
  - Grupe fiksne veličine
  - Mek grupiranje
  - …
- Radi dobro kompresiju podataka.



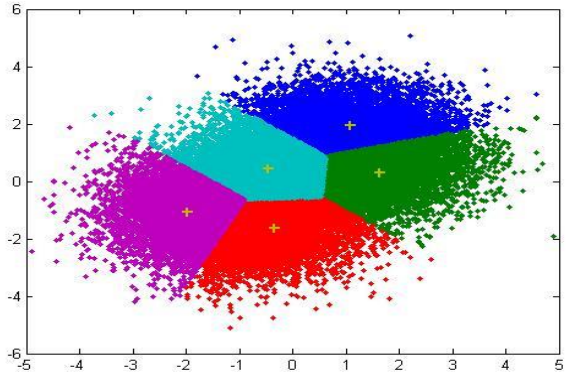
Nedostaci algoritma k-srednjih vrijednosti

- Obično završava u **lokalnom optimumu** (poboljšanje -pametne inicijalizacije k-means++, višestruko pokretanje s različitim inicijalizacijama)
- Potrebno je odrediti **broj k (broj grupa) unaprijed**

Nedostaci algoritma k- srednjih vrijednosti

- Obično završava u **lokalnom optimumu** (pametne inicijalizacije k-mens++, višetruko pokretanje s različitim inicijalizacijama)
- Potrebno je odrediti **broj k (broj grupa) unaprijed**

- NE ponaša se dobro na podacima sa **šumom i stršećim vrijednostima**
- Grupe imaju samo **konveksne oblike**



Kako izabrati *k*?

b(i): avg. distance to points in closest other cluster

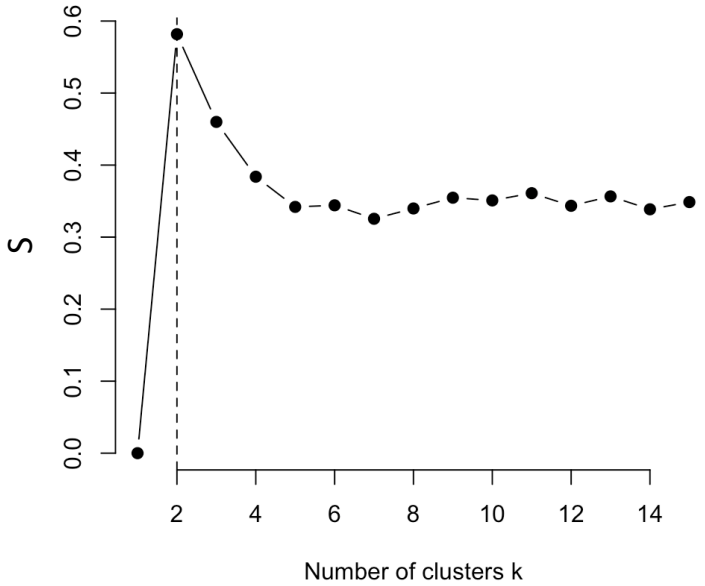
a(i): avg. distance to points in own cluster

Radi *k-means* za k = 1, 2, 3, …

Za svaku točku i računaj siluetu (***silhouette)*** $s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$

S = prosjek s(i) po svim i

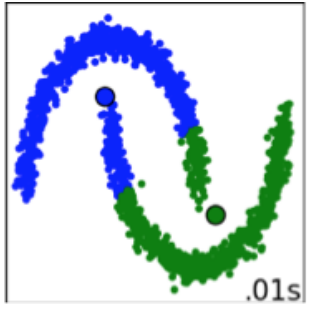
Odaberi **k** za koji je **S** najveći.

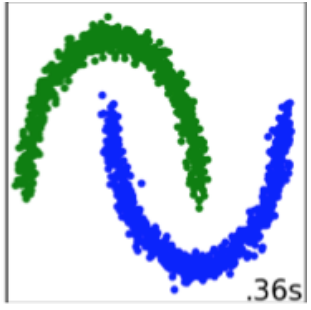


DBSCAN

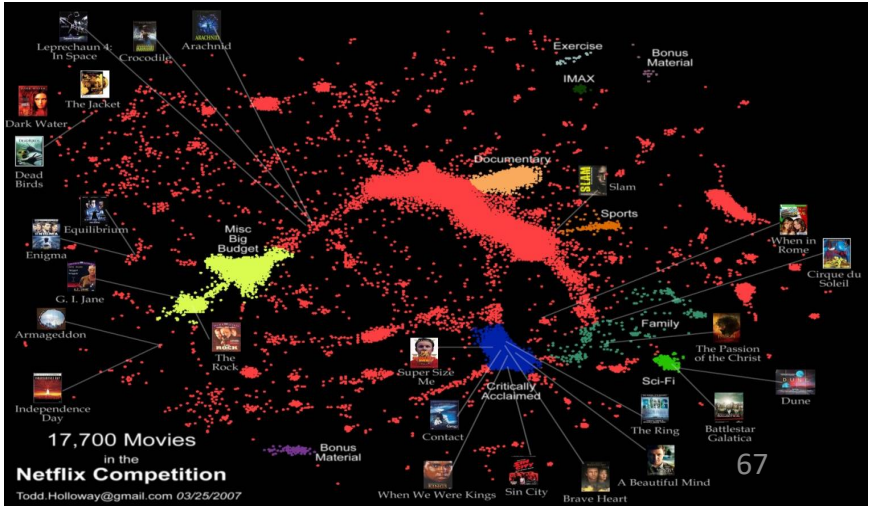
"**D**ensity-**b**ased **s**patial **c**lustering of **a**pplications with **n**oise"

- Centroid-based, sličan k-means-u, preferira sferične grupe



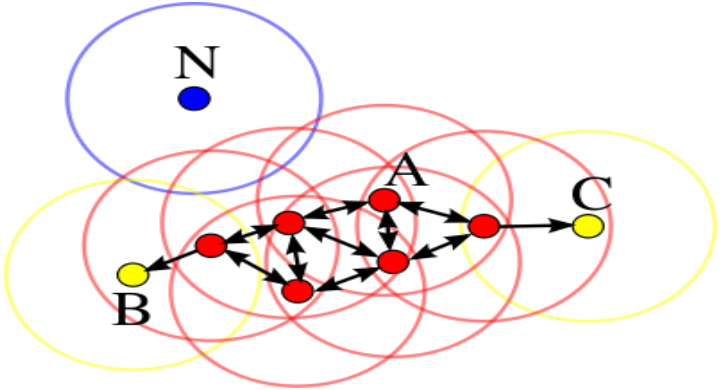


- Sa stvarnim podacima:



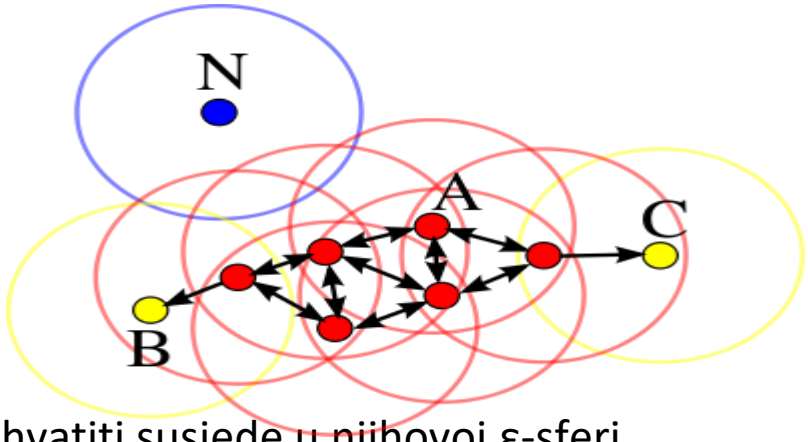
DBSCAN

- DBSCAN izvodi *density-based clustering*, i slijedi oblik gustog susjedstva točaka. PARAMETRI: **ε i *minPts***



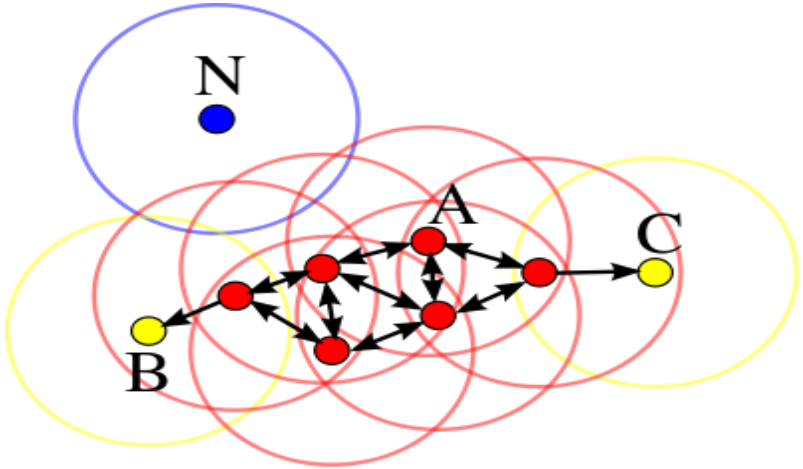
- ***Core points*** imaju najmanje ***minPts*** susjeda u sferi dijametra ε oko njih.
- Crvene točke su core points sa barem *minPts* = 3 susjeda u **ε-sphere**.

DBSCAN

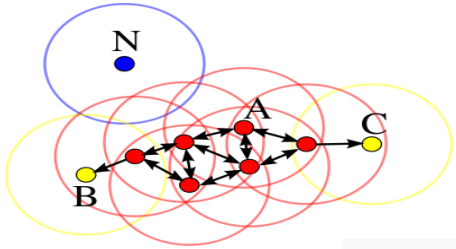


- **Core points** mogu direktno dohvatiti susjede u njihovoj ε-sferi
- Za točke koje nisu core points ne mogu s dohvatiti druge točke
- Točka q je **density-reachable** ako postoji niz točaka $p = p_1, \ldots, p_n = q$ tako da je p_{i+1} direktno dohvatljiva iz p_i
- Sve točke koje nisu *density reachible* iz bilo koje druge točke su stršeće vrijednosti (**outliers)**

DBSCAN grupe

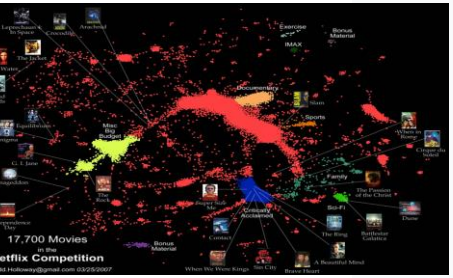


- Točke *p*, *q* su **density-connected** ako postoji točka *o* tako da su obje *p* i *q* *density-reachable* iz *o*.
- Grupa (**cluster) je skup točaka koje su međusobno *density-connected***.
- To znači ako je neka točka *density-reachable* iz točke neke grupe, onda je i ona također član grupe.
- Crvene točke su međusobno *density reachable*; B i C su *density-connected*; N je *outlier*.



DBSCAN algorithm

```
DBSCAN(DB, dist, eps, minPts) {
   C = 0                                                  /* Cluster counter */
   for each point P in database DB {
      if label(P) ≠ undefined then continue               /* Previously processed in inner loop */
      Neighbors N = RangeQuery(DB, dist, P, eps)          /* Find neighbors */
      if |N| < minPts then {                              /* Density check */
         label(P) = Noise                                 /* Label as Noise */
         continue
      }
      C = C + 1                                           /* next cluster label */
      label(P) = C                                        /* Label initial point */
      Seed set S = N \ {P}                                /* Neighbors to expand */
      for each point Q in S {                             /* Process every seed point */
         if label(Q) = Noise then label(Q) = C            /* Change Noise to border point */
         if label(Q) ≠ undefined then continue            /* Previously processed */
         label(Q) = C                                     /* Label neighbor */
         Neighbors N = RangeQuery(DB, dist, Q, eps)       /* Find neighbors */
         if |N| ≥ minPts then {                           /* Density check */
            S = S ∪ N                                     /* Add new neighbors to seed set */
         }
      }
   }
}
```



DBSCAN performanse

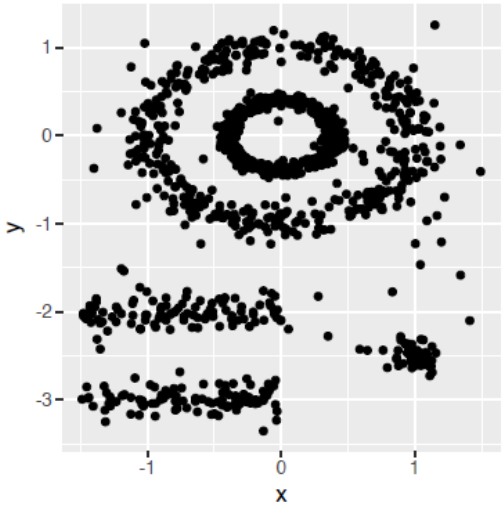
- DBSCAN koristi udaljenosti svih parova, ali korištenjem efikasne indeksne strukture, svaki RangeQuery (za pronalaženje susjeda u ε-sferi) treba samo O(log n) vremena
- Algoritam je složenosti **O(n log n)**
- Može pronaći grupe bilo kojih oblika (ne nužno kružni)
- Može identificirati stršeće vrijednosti

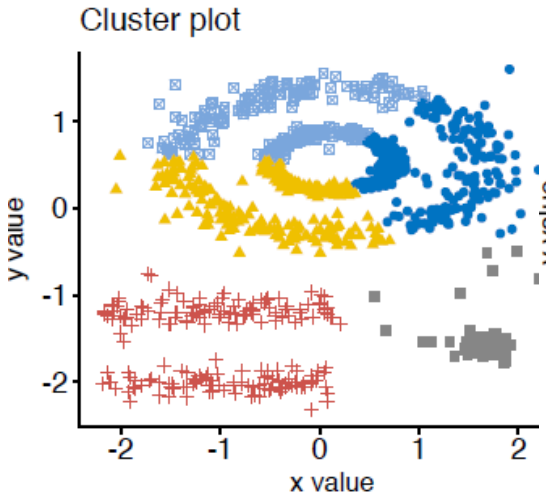
DBSCAN performanse

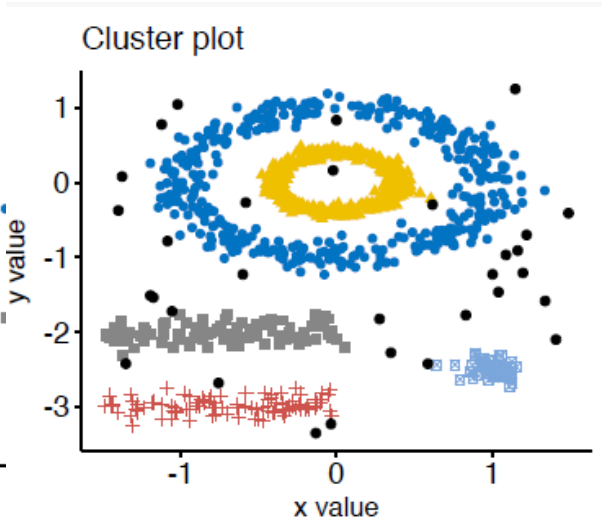
- DBSCAN koristi udaljenosti svih parova, ali korištenjem efikasne indeksne strukture, svaki RangeQuery (za pronalaženje susjeda u ε-sferi) treba samo O(log n) vremena
- Algoritam je složenosti **O(n log n)**
- Može pronaći grupe bilo kojih oblika (ne nužno kružni)
- Može identificirati stršeće vrijednosti
- Osjetljiv je na izbor eps, posebno ako su grupe različite gustoće
- Brzo pretraživanje susjedstva postaje teže u višim dimenzijama

DBSCAN performanse

Originali skup, k-means i DBSCAN







Ovo predavanje temelji se na nastavnim materijalima predmeta *Applied Data Analysis* (ADA) EPFL-a autora Roberta Westa.