

## Grupiranje (engl. Clustering)

- do sada bavateli "nadgledanom klasifikacijom" ili učili (klasifikator) s učiteljem
  - Grupiranje → "neoznačeni uzoreci"  
(ne znamo pripadnosti takvih uzoraka niti znamo broj razreda iz kojih uzoreci dolaze)
- Zadatak: Otkriti organizaciju uzoraka i grupirati ih u "smislene" ("prirodne") grupe koje će nam omogućiti otkrivanje sličnosti i različitosti između uzoraka i time dopustiti izvođenje korisnih zaključaka o njima.

Ovakva se zamisao obilato rabi u :

- biologiji i zoologiji
- psihijatriji i patologiji
- sociologiji
- arheologiji
- zemljopisu
- geologiji
- tehniči

Grupiranje → nenadgledano učenje  
 (engl. unsupervised learning)

učenje bez učitelja

} PR

→ numerička taksonomija

biologija i  
ekologija

→ tipologija i društvene  
znanosti

Primjer :

Razmotrimo sljedeće životinje :

ovca, pas, mačka (sisarci)

vrabac, galeb. (ptice)

zmiјa, gušter (reptili)

elatnaribica, cipla (ribe)

žaba (vodozemac)

cipla =  
cipalj

Organizirajmo ih u grupe !

- Kriterij grupiranja ?

a) Npr. da li ženke nose svoje (buduće) mladuncade?

Grupe :

ovca  
pas  
mačka

cipla  
vrabac  
galeb  
gušter  
zmiјa  
elatna ribica  
žaba

a)

b) da li imaju pluća?

elatna ribica  
cipla

5)

ovca pas  
mačka gusten  
vrabac galeb  
žabka žabka  
zmija

c) okoliš u kojem žive?

ovca  
pas  
mačka zmija  
galeb  
vrabac  
gusten

žabka

elatna ribica  
cipla

c)

Osnorni koraci u postupku grupiranja:

- Izbor znatljiki
- Izbor mjere sličnosti (ili različitosti)
- Kriterij grupiranja  
(zavisi od interpretacije eksperta  
čemu daje naglasak u "smislenom"  
razvrstavanju neoznacenih  
uporaba)
- Algoritam grupiranja
- Validacija rezultata
- Interpretacija rezultata

U brojnim slučajevima rabi se još jedan dodatni korak:

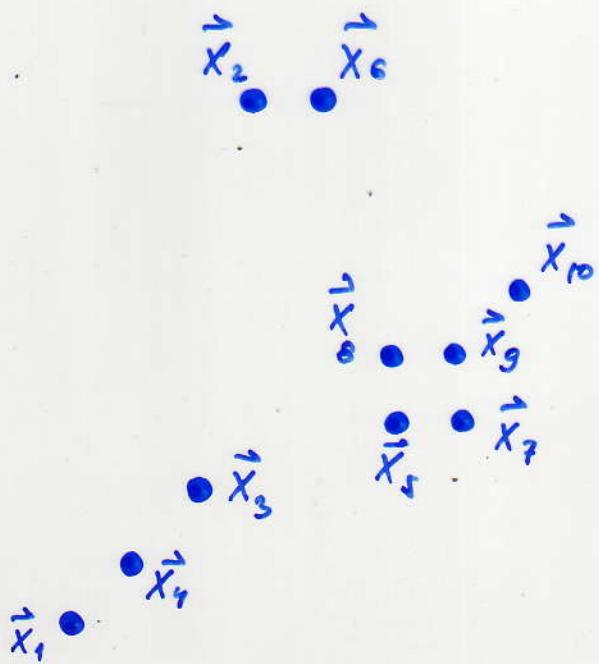
- težnja grupiranju
- podrazumijeva različite testove koji pokazuju da li raspoloživi podaci imaju strukturu grupe

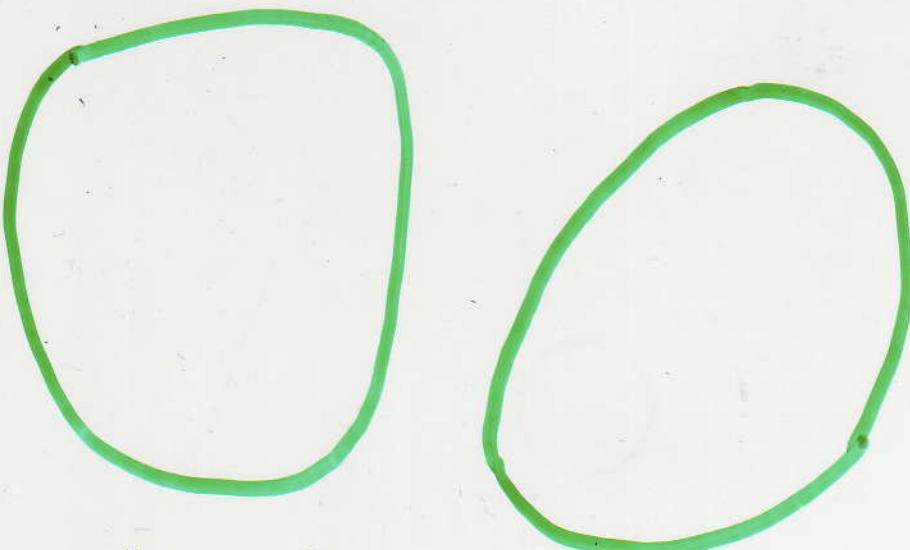
(npr. skup podataka može biti potpuno slučajne prirode te pokušaj otkrivanja "smislenih" grupa je besmislen)

Svi osnovni koraci su podložni subjektivnosti eksperta!

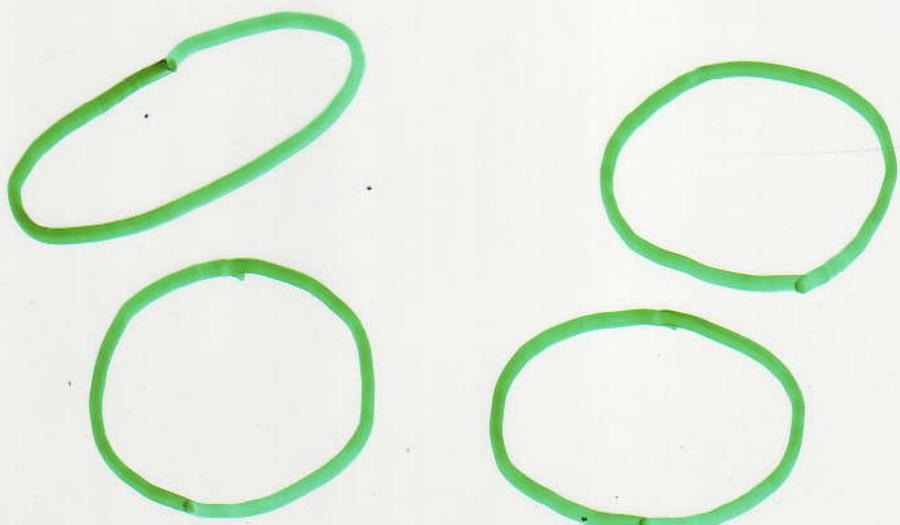
(Subjectivity is a reality we have to live with from now on)

Primjer:





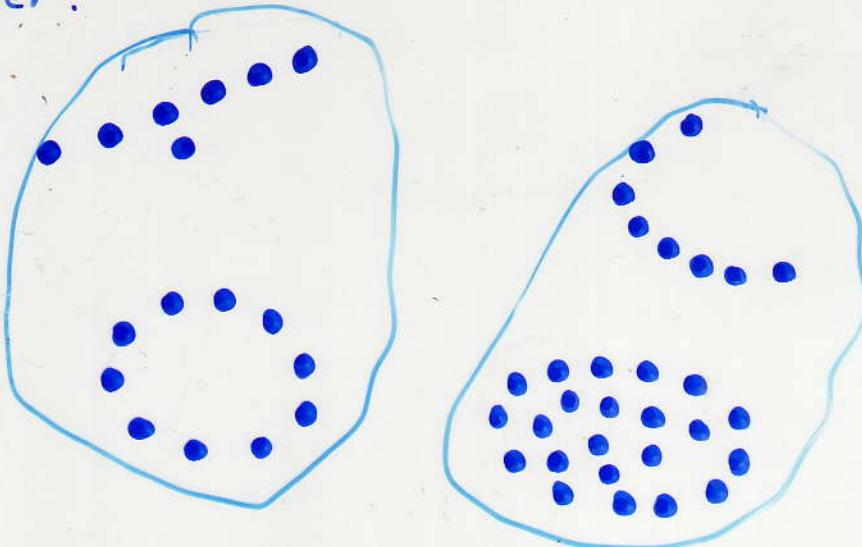
a) "Grublje" grupiranje u dvije grupe



b) "Finije" grupiranje

\* Primjer:

gr 5 \*



## Primjeri uporabe grupiranja:

### - Redukcija podataka

- $N$  raspoloživih podataka  $N \gg 1$
- postupkom grupiranja (u "smislene" grupe) dobivamo  $m \ll N$  grupa

### - Generiranje hipoteza

- uporabljamo analizu grupa (grupiranje) u cilju utvrđivanja i zaključivanja u vezi prirode podataka

Grupiranje  $\rightarrow$  poticaj za postavljanje hipoteza

### - Ispitivanje hipoteza

### - Predviđanje na temelju grupe

Npr. analiza grupa je primijenjena na skupu podataka o pacijentima koji su oboljeli od iste bolesti

rezultat - broj grupa pacijenata prema njihovoј reakciji na određene lijekove

- novi pacijent - za njega identificiramo odg. grupu

## Vrste značajki

- značajke mogu zauzimati vrijednosti iz nekog kontinuiranog opsega (podskup od  $R$ ) ili iz nekog konačnog diskretnog skupa.
- ako je konačan skup diskretan i ima samo DVA elementa tada se značajka naziva BINARNA ili DIHOTOMNA (dichotomous)

Klasifikacija značajki na temelju relativnog značenja vrijednosti koje one zauzimaju:

- nominalne (a)
- uređene (b)
- intervalno skalirane (c)
- skalirane omjerom (d)

(a) spol osoba : npr. 1 za muškarce

0 za žene  
(ili obratno)

/ kvantitativno uspoređivanje između tih vrijednosti je besmisleno/

(b) karakterizacija sposobnosti

5, 4, 3, 2, 1

odličan  
vrlo dobar  
dobar  
dovoljan, nedovoljan

c) Mjerenje temperature u °C

Npr. ako je Paris  $10^{\circ}\text{C}$   
London  $5^{\circ}\text{C}$

Smisleno je reći da je temperatura u Parizu za  $5^{\circ}\text{C}$  viša od one u Londonu

Besmisleno (ili skoro besmisleno), je reći da je Pariz dvostruko topliji od Londona

d) Omjer između enacijački ima smisla.

Npr. osoba koja ima  $120\text{kg}$  je dvaput teža (i debja) od osobe koja ima  $60\text{kg}$ .

## Definicije grupiranja

Definicije temelje na "labavo" definiranim razima kao što su "slični", "različiti"  
(odnose se na uzorce u pojedinim rezedima)

Everitt, 1981:

Vektori → točke u  $l$ -dimensionalnom prostoru

Grupe → kontinuirana područja prostora koja imaju relativno visoku gustoću točaka i odvojena su od drugih kontinuiranih prostora visokih gustoća s područjima relativno malih gustoća točaka

Grupe opisane na taj način  
vrlo često se nazivaju i  
"prirodne grupe" (engl. natural  
clusters)

Definicija grupiranja:

- Neka je  $X$  skup podataka

$$X = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_N\}$$

m - grupiranje  $X$ -a odgovara  
dijeljenju  $X$  u m skupova (grupa)

$C_1, C_2, \dots, C_m$  C - cluster

tako da su zadovoljena sljedeća  
tri uvjeta:

i)  $C_i \neq \emptyset, i=1, 2, \dots, m$

ii)  $\bigcup_{i=1}^m C_i = X$

iii)  $C_i \cap C_j = \emptyset, i \neq j, i, j = 1, 2, \dots, m$

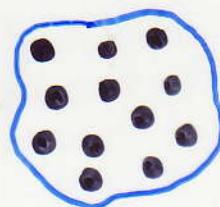
Važno:

Vektori sadržani u grupi  $C_i$  su  
"sličniji" jedan drugome i  
"manje sličniji" vektorima  
iz drugih grupa.

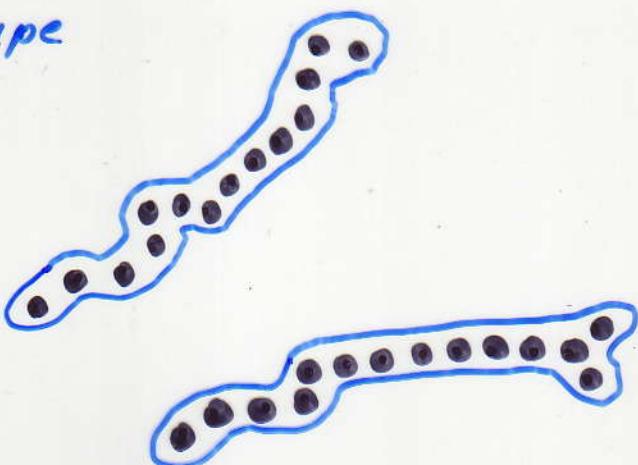
Kvantifikacija i crata  
 "slican" i "ratlican" zavisni od  
 tipa grupe.

Primjer:

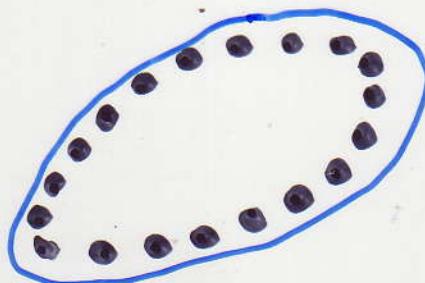
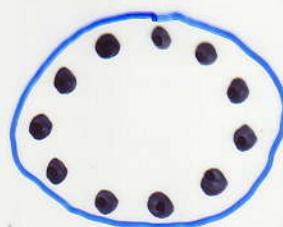
a) kompaktne grupe



b) "izduzene" grupe



c) sferične i elipsoidne grupe



U definiciji grupe zahtijevamo da svaki vektor pripada samo jednoj grupi  
 → takva vrsta grupiranja naziva se

### crisp grupiranje

(čerazito, jasno)

Alternativni pristup grupiranju:

### nečerazito grupiranje

(fuzzy clustering)

Nečerazito grupiranje skupa podataka  $X$  u  $m$  grupe određeno je s  $m$  funkcijama  $u_j$ :

$$u_j : X \rightarrow [0, 1], \quad j=1, 2, \dots, m$$

$$\sum_{j=1}^m u_j(\vec{x}_i) = 1 \quad i=1, 2, \dots, N$$

$$0 < \sum_{i=1}^N u_j(\vec{x}_i) < N \quad j=1, 2, \dots, m$$

$u_j, j=1, 2, \dots, m \rightarrow$  funkcije pripadnosti  
 (engl. membership functions)

U slučaju nečrtežitog grupiranja svaki vektor  $\vec{x}_i$  pripada više od jedne grupi istodobno, s nekim stupnjem ići mjerom pripadnosti (iz intervala  $[0, 1]$ ).

### Mjere bliskosti (sljčnosti) (engl. proximity measures)

- Mjera različitosti (dissimilarity measure DM)

DM je funkcija  $d$  od  $X$

$$d : X \times X \rightarrow \mathbb{R},$$

gdje je  $\mathbb{R}$  skup realnih brojeva takav da :

$$\exists d_0 \in \mathbb{R} : -\infty < d_0 \leq d(\vec{x}, \vec{y}) < +\infty$$

$$\forall \vec{x}, \vec{y} \in X$$

$$d(\vec{x}, \vec{x}) = d_0 \quad \forall \vec{x} \in X$$

$$d(\vec{x}, \vec{y}) = d(\vec{y}, \vec{x}) \quad \forall \vec{x}, \vec{y} \in X$$

$$d(\vec{x}, \vec{y}) = d_0 \text{ iff } \vec{x} = \vec{y}$$

i još dodatni uvjet:

$$d(\vec{x}, \vec{z}) \leq d(\vec{x}, \vec{y}) + d(\vec{y}, \vec{z})$$

$$\forall \vec{x}, \vec{y}, \vec{z} \in X$$

**d se naziva metrika DM**

$d_0$  - minimalna moguća različitost između bilo koja dva vektora iz  $X$  (dobiva se kada su oni identični!)

• Mjera sličnosti (similarity measure SM)

SM je funkcija s:

$$s : X \times X \rightarrow \mathbb{R}$$

tako da je:

$$\exists s_0 \in \mathbb{R} : -\infty < s(\vec{x}, \vec{y}) \leq s_0 < +\infty$$
$$\forall \vec{x}, \vec{y} \in X$$

$$i) s(\vec{x}, \vec{y}) = s(\vec{y}, \vec{x})$$

$$s(\vec{x}, \vec{x}) = s_0 \quad \forall \vec{x} \in X$$

$$i) s(\vec{x}, \vec{y}) = s_0 \text{ iff } \vec{x} = \vec{y}$$

$$s(\vec{x}, \vec{y}) s(\vec{y}, \vec{z}) \leq [s(\vec{x}, \vec{y}) + s(\vec{y}, \vec{z})] \cdot s(\vec{x}, \vec{z})$$

$\forall \vec{x}, \vec{y}, \vec{z} \in X$

**$s$  se nativa metrika  $SM$**

Primjer:

Euklidska udaljenost  $d_2(\vec{x}, \vec{y})$

$$d_2(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2},$$

gdje  $\vec{x}, \vec{y} \in X$

$x_i : y_i$  su  $i$ -te koordinate od  $\vec{x}$  i  $\vec{y}$

DM je definiran s  $d_0 = 0$

(minimalna moguća udaljenost između dvaju vektora iz  $X$  je 0.

Euklidska udaljenost je metrika DM.

**Pozor:** svih algoritmi grupiraju ne temelje na mjeri bliskosti između vektora.

Neki (hjерархијски) algoritmi grupiraju) računaju udaljenosti između parova skupova vektora iz  $X$ .

Bliškost između podskupova od  $X$ :

Neka je  $U$  skup koji sadrži podskupove od  $X$ .

$$D_i \subset X, i = 1, 2, \dots, k$$

$$U = \{D_1, D_2, \dots, D_k\}$$

Mjera bliskosti  $\rho$  definisana nad  $U$  je:

$$\rho : U \times U \rightarrow \mathbb{R}$$

gde je  $\mathbb{R}$  skup realnih brojeva takav da

$$\exists d_0 \in \mathbb{R} : -\infty < d_0 \leq d(D_i, D_k) < +\infty$$

$$\forall D_i, D_k \in U$$

$$d(D_i, D_i) = d_0 \quad \forall D_i \in U$$

i

$$d(D_i, D_k) = d(D_k, D_i)$$

$$\forall D_i, D_k \in U$$

$$i \quad d(D_i, D_k) = d_0 \text{ iff } D_k = D_i$$

i

$$d(D_i, D_\ell) \leq d(D_i, D_k) + d(D_k, D_\ell)$$

$$\forall D_i, D_k, D_\ell \in U$$

## KRITERIJI GRUPIRANJA

Postupci:

- i) Heuristički - rodeni intuicijom i iskustvom
- ii) Oni koji se temelje na minimizaciji (ili maksimizaciji) neke kriterijske funkcije ili (performance-index) indeksa

Najčešći kriterij

$$J = \sum_{j=1}^{N_c} \sum_{\vec{x} \in S_j} \|\vec{x} - \vec{m}_j\|^2$$

$N_c$  - broj grupa

$S_j$  - skup uzoraka koji pripadaju  $j$ -toj grupi

$$\vec{m}_j = \frac{1}{N_j} \sum_{\vec{x} \in S_j} \vec{x}$$

$N_j$  - broj uzoraka u grupi  $S_j$

- iii) Kombinacija heurističkog pristupa i onog u ii)

i) Heuristički postupci grupiranja  
— jednostavan algoritam grupiranja  
Imamo  $N$  neoznačenih učoraka

$$C = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_N\}$$

1) Korak :

Izaberi nenegativan broj  
 $T$  (prag)

2) Korak :

Izaberi bilo koji učorak iz  
 $C = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_N\}$  i proglaši  
ga središtem grupe  $\vec{z}_1$

Pretpostavimo da smo izabrali

$$\vec{x}_1 = \vec{z}_1$$

3) Korak :

Iračunamo udaljenost

$$d_2(\vec{x}_1, \vec{x}_2) = d_2(\vec{z}_1, \vec{x}_2)$$

i uspoređujemo je s pragom  $T$ :

a) ako je  $d_2(\vec{z}_1, \vec{x}_2) > T$

proglašavamo  $\vec{x}_2$  središtem  
nove grupe  $\vec{z}_2 = \vec{x}_2$

b) ako je  $d_2(\vec{z}_1, \vec{x}_2) \leq T$

$\vec{x}_2$  dodjeljujemo grupi sa središtem  $\vec{z}_1$

Pretpostavimo da je  $d_2(\vec{z}_1, \vec{x}_2) > T$

tada  $\vec{z}_2 = \vec{x}_2$

4) Korak:

Računamo udaljenosti

$d_2(\vec{z}_1, \vec{x}_3)$  i  $d_2(\vec{z}_2, \vec{x}_3)$

a) Ako su  $d_2(\vec{z}_1, \vec{x}_3)$  i  $d_2(\vec{z}_2, \vec{x}_3)$   
 $> T$  formiramo središte nove  
 grupe  $\vec{z}_3 = \vec{x}_3$

b) Ako nije a)  $\vec{x}_3$  se dodjeljuje  
 grupi čijem je središtu  
 najbliže

Postupak se provodi dok se ne  
 obradi svih  $N$  uzoraka

Prednost algoritma:

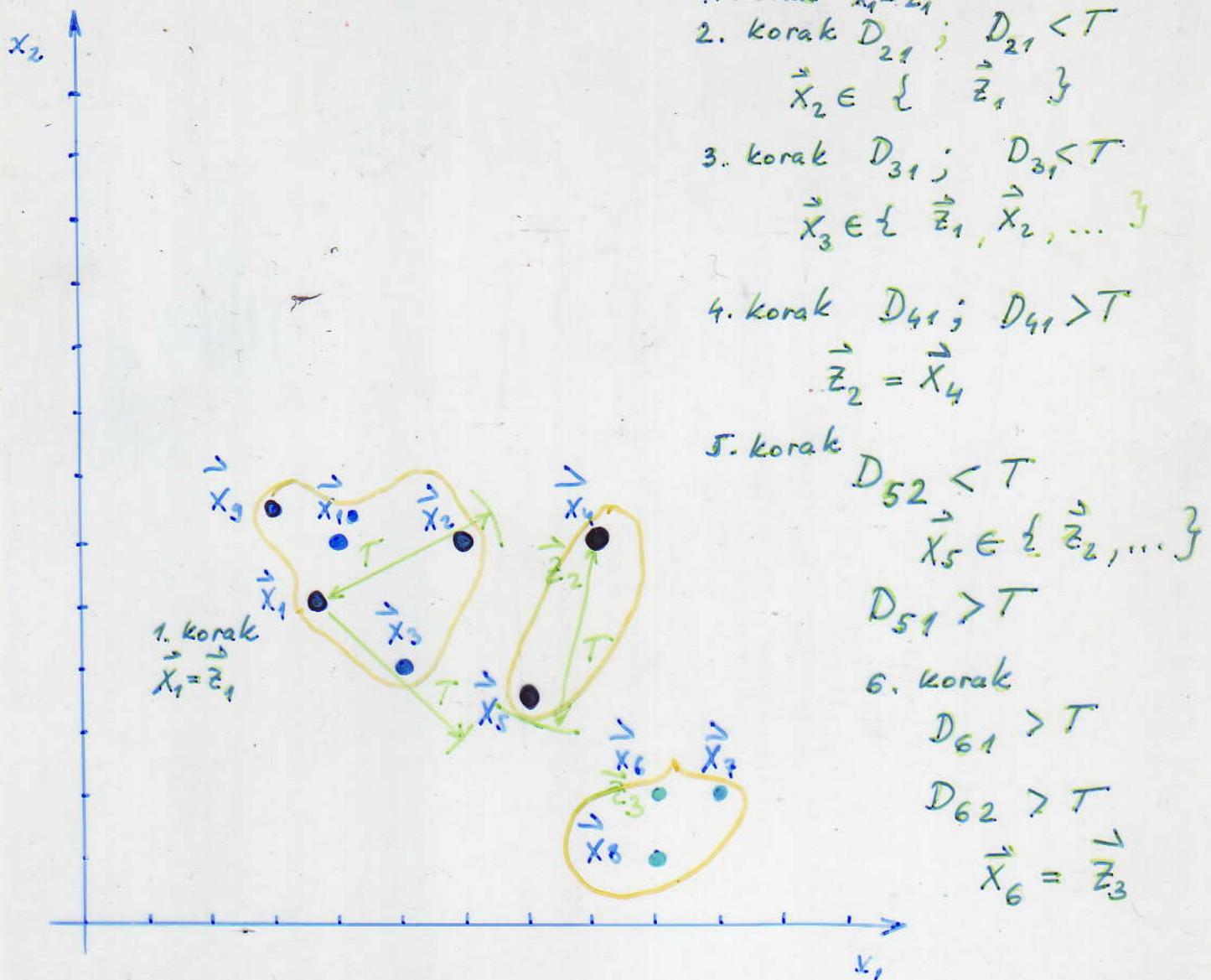
- njegova jednostavnost  
rezultat se dobiva jednim  
prolazom kroz skup  
uzoraka  $C$

Nedostatak:

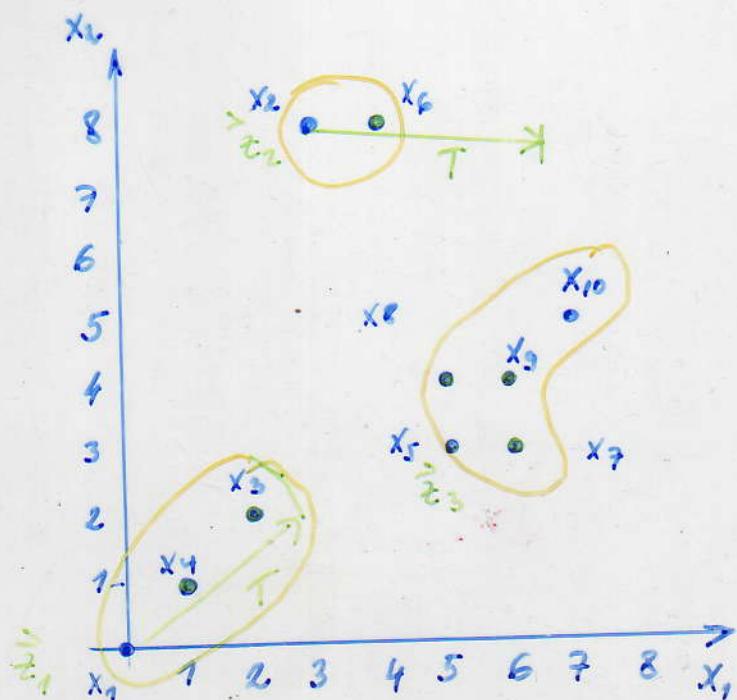
- rezultat zavisi od vrijednosti  
praga  $T$
- zavisi od izbora prvog središta  
grupe
- zavisi od redoslijeda uzimanja  
uzoraka i  $\in C$ .

PRIMER JEDNOSTAVNOG HEURISTIČKOG  
ALGORITMA

Ruza  
građevna

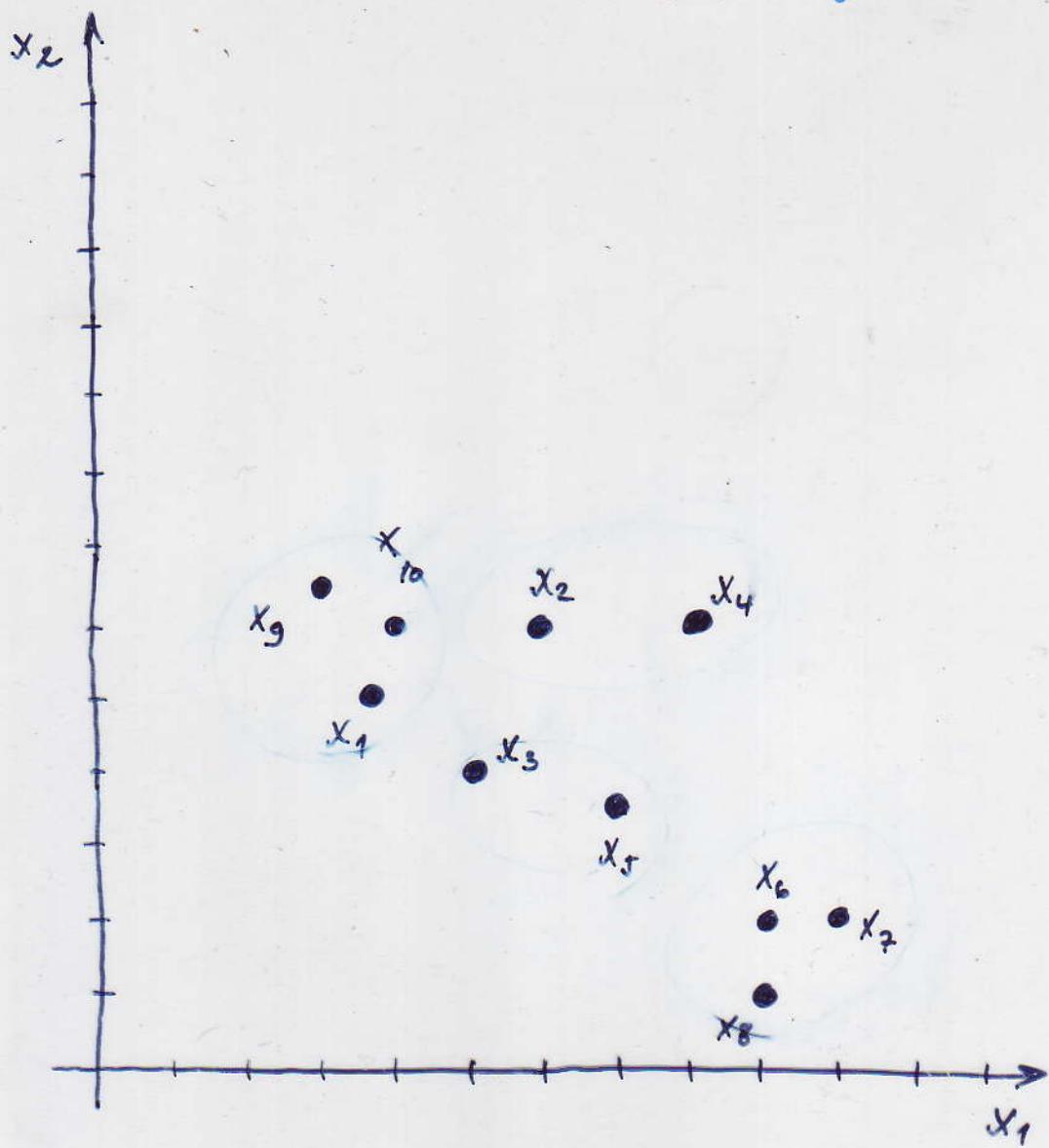


$T$  vrijednost praga  $T$

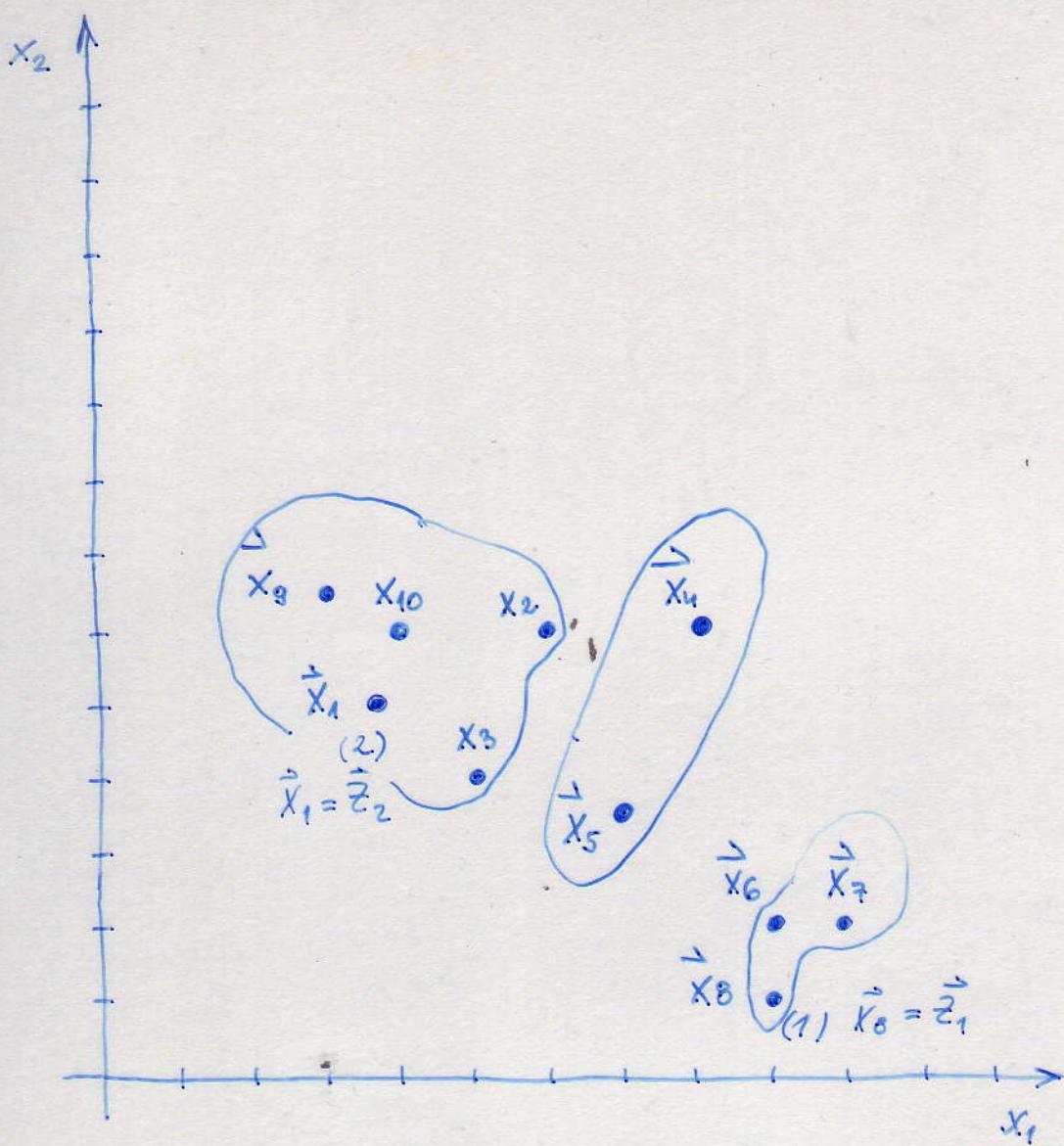


PRIMER : Jednostavan algoritam  
grupiranja

RE 2a1  
gr 27



$\overline{T}$  vrijednost praga  $T$



1. Korak  $\vec{x}_8 = \vec{z}_1$

2. Korak  $D_{18} > T$   
 $\vec{x}_1 = \vec{z}_2$

3. Korak  $D_{28}$   $D_{28} > T$   
 $D_{21}$   $D_{21} < T$   
 $\vec{x}_2 \in \{\vec{z}_2, \dots\}$

4. Korak  $D_{38}$   $D_{38} > T$   
 $D_{31}$   $D_{31} < T$   
 $\vec{x}_3 \in \{\vec{z}_2, \vec{x}_2, \dots\}$

4. Korak  
 $D_{48} > T$   
 $D_{41} > T$   
 $\vec{x}_4 = \vec{z}_3$

5. Korak  
 $D_{58} > T$   
 $D_{51} > T$   
 $D_{54} < T$   
 $\vec{x}_5 \in \{\vec{z}_3, \dots\}$

6. Korak  
 $D_{68} < T$   $\vec{x}_6 \in \{z_1, \dots\}$   
 $D_{61} > T$   
 $D_{64} > T$

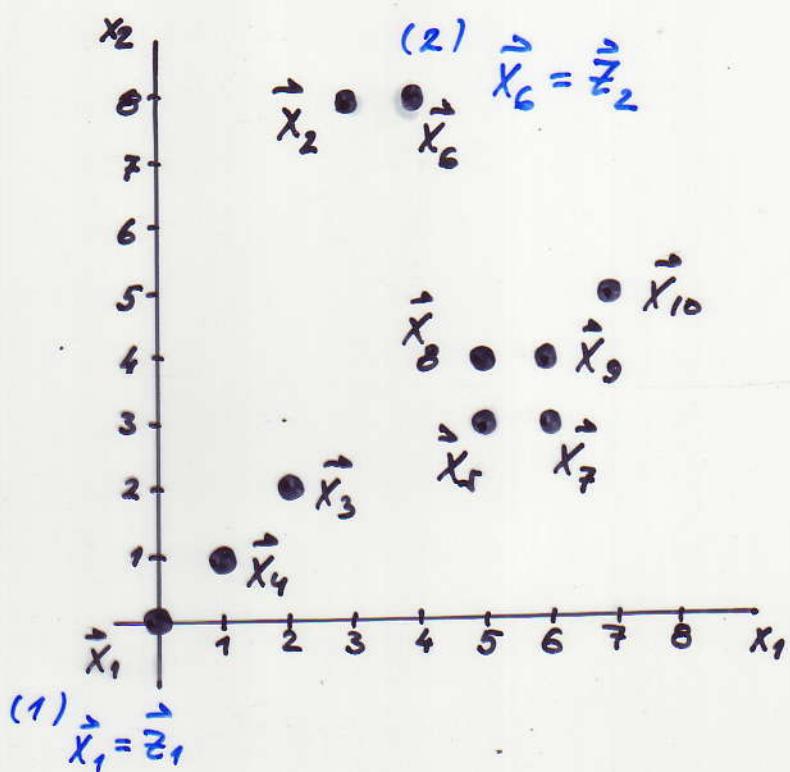
# Heuristički algoritam grupiranja

## MAXMIN udaljenosti

(engl. maximum-minimum distance)

- Sličan prethodnom algoritmu ali s tom razlikom što se prvo identificiraju područja grupe koje su najudaljenija
- Temelji se na Euklidskoj udaljenosti

Primjer: MAXMIN algoritam



### 1. korak

Izaberemo  $\vec{z}_1 = \vec{x}_1$

$\vec{x}_1$  - središte prve grupe

### 2. korak

Odredimo najudaljeniji uzorak od  $\vec{z}_1 = \vec{x}_1$   
 ( u našem slučaju je to  $\vec{x}_6$  )  
 i proglašavamo ga središtem  $\vec{z}_2$  !

### 3. korak

$$\vec{z}_1 = \vec{x}_1 ; \quad \vec{z}_2 = \vec{x}_6$$

Izračunavamo udaljenosti između preostalih uzoraka i uzoraka  $\vec{z}_1$  i  $\vec{z}_2$

označava	$D_{12}$	$D_{13}$	$D_{14}$	$D_{15}$	$D_{17}$	$D_{18}$	$D_{19}$	$D_{1,10}$
$\vec{z}_1$	$D_{22}$	$D_{23}$	$D_{24}$	$D_{25}$	$D_{27}$	$D_{28}$	$D_{29}$	$D_{2,10}$

↑  
označava  $\vec{z}_2$

Za svaki par izaberemo i pohranimo  
MINIMALNU VRJEDNOST :

$$D_{22} \quad D_{13} \quad D_{14} \quad D_{25} \quad D_{27} \quad D_{28} \quad D_{29} \quad D_{2,10}$$

### 4. korak

Izaberemo MAKSIMUM od tih minimalnih vrijednosti !  
 $(D_{2,7})$

### 5. korak

Ako je ta udaljenost signifikantna u odnosu na udaljenost između  $\vec{z}_1$  i  $\vec{z}_2$  (npr. najmanje 1/2 udaljenosti), uzorak

koji odgovara toj udajenosti  $\underline{\text{proglavaju}}$   
se središtem NOVE GRUPE  $\vec{z}_3$ ,

u drugim slučajevima algoritam završava.

$$(\vec{x}_2 = \vec{z}_3)$$

### 6. korak

Izračunavamo udajenosti uzoraka od

$$\vec{z}_1, \vec{z}_2 \text{ i } \vec{z}_3$$

$$\left( \begin{array}{lll} D_{1,2}, & D_{1,3}, & \dots \\ D_{2,1}, & D_{2,3}, & \dots \\ D_{3,1}, & D_{3,2}, & \dots \\ \downarrow & & \\ \text{označava} & & \\ \vec{z}_3 & = & \vec{x}_2 \end{array} \right)$$

Postupak se ponavlja - traži se minimum trojki udajenosti, pobrajuje se - bira se maksimum i uspoređuje sa 1/2 udajenosti  $\vec{z}_1, \vec{z}_2$  -

### 7. korak

Preostali uzorci se dodjeljuju najblizim središnima grupa!

# Heurištički algoritam grupiranja

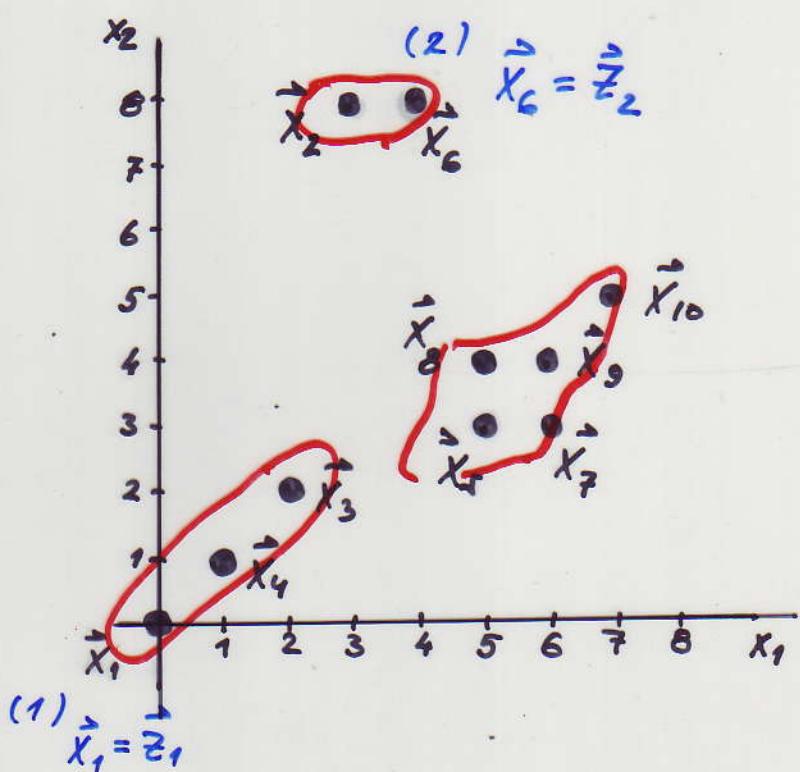
## MAXMIN udaljenosti

(engl. maximum-minimum distance)

- Sličan prethodnom algoritmu ali s tom razlikom što se prvo identificiraju područja grupe koje su najudaljenija
- Temelji se na Euklidskoj udaljenosti

Primjer: MAXMIN algoritam

Rezultat grupiranja MAXMIN udaljenosti



# GRUPIRANJE NA TEMEYU TEORIJE GRAFOVA

Postupci grupiranja koji se temelje na teoriji grafova imaju neke zajedničke točke:

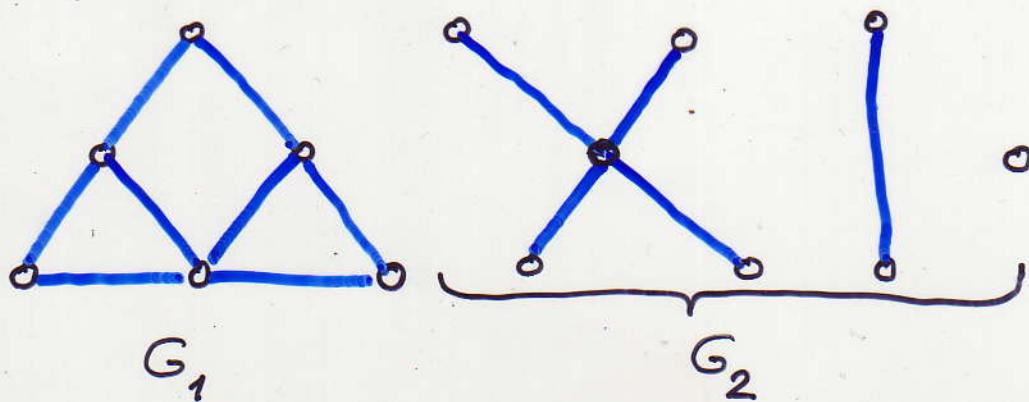
- svakom se uzorku iz skupa uzoraka  $S_N$  dodijeli čvor povezanog grafa  $G$
- grupe se određuju traženjem komponenata povezanosti grafa  $G$

**Def. 1.** Neorientirani graf je povezan ako se njegova dva proizvoljna čvora mogu povezati putem koji čine lukovi grafa;

- ako postoji čvorovi koji se ne mogu povezati putem  $\rightarrow$  graf je nepovezan;

**Def. 2.** Nepovezani graf se sastoji od dva ili više odvojenih dijelova. Odvojeni dijelovi grafa nativaju se Komponente povezanosti grafa;

Primjer



- Većina postupaka grupiranja na temelju teorije grafova koristi matricu sličnosti koja se generira pomoću udaljenosti između uzorka u  $S_N$ :

$$D_{kl} = \|\vec{x}_k - \vec{x}_l\| \quad k=1, 2, \dots, N \\ l=1, 2, \dots, N$$

/Euklidска udaljenost - može se koristiti i neka druga mjera /

Matrica sličnosti dimenzija  $N \times N$  je binarna matrica čiji elementi su

$$s_{kl} = \begin{cases} 1, & \text{ako je } D_{kl} \leq \Theta \\ 0, & \text{ako je } D_{kl} > \Theta \end{cases}$$

$\Theta$  - prag

Matrica sličnosti  $S$  definira graf sličnosti u kojem čvorovi odgovaraju uzorcima iz  $S_N$ , a grane grafa (lukovi) povezuju čvorove  $i$  i  $j$  samo ako je  $s_{ij} = 1$ .

**Postupak grupiranja (R.O. Duda, P.E. Hart, D.G. Stork, 2001)** tzv. single-linkage algoritam: dva uzorka  $\vec{x}_i$  i  $\vec{x}_j$  su u istoj grupi akko postoji niz  $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_k$  za koji vrijedi da je  $\vec{x}_i$  sličan  $\vec{x}_1, \vec{x}_2$  sličan  $\vec{x}_2, \dots$ , itd., ...  $\vec{x}_{j-1}$ , sličan  $\vec{x}_j$  - tako da

grupa odgovara komponentama povezanosti  
grafa sličnosti.

Jednostavan postupak grupiranja  
određivanjem komponenti povezanosti  
grafa sličnosti (W. S. Meisel, 1972)

### 0. korak.

Generiramo matricu sličnosti  $S$

### 1. korak

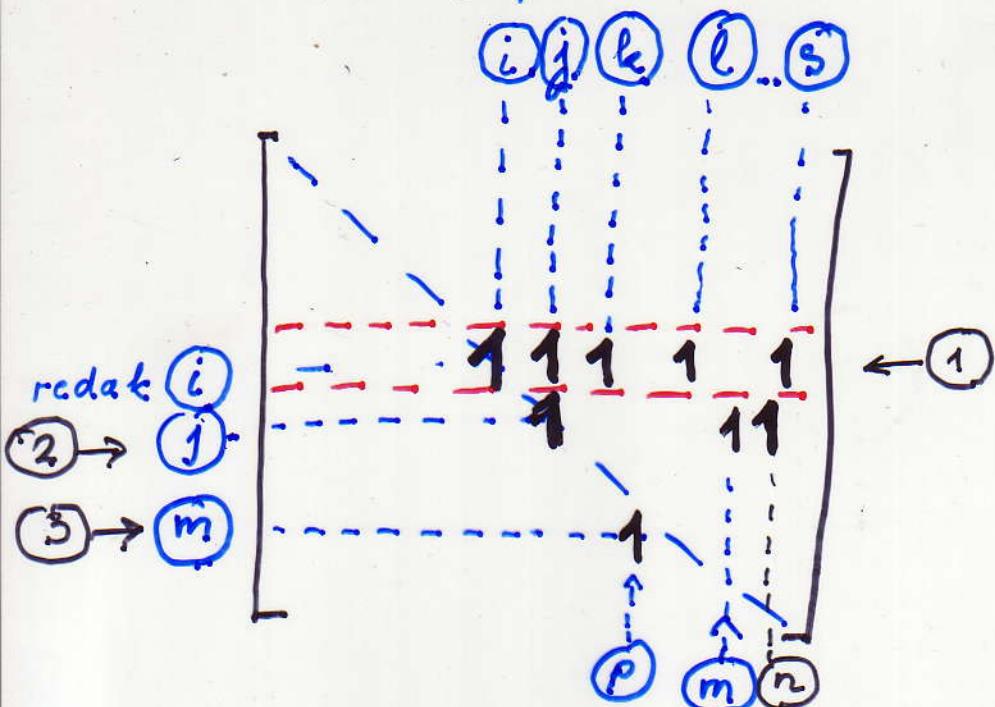
Izaberimo redak matrice sličnosti koja ima najveći broj jedinica. Pretpostavimo da je najviše jedinica u  $i$ -tom retku.

### 2. korak

Tvorimo grupu od uzoraka koji odgovaraju  $i$ -tom retku. Pretpostavimo da su jedinice u stupcima  $j, k, l, \dots$ . Toj grupi pridodajemo još uzorke koji odgovaraju jedinicama u  $j, k, l, \dots$  - tom retku.

U slučaju da su u tim redima jedinice i izvan  $j, k, l, \dots$  - tog stupca, npr. u stupcima  $m, n, o, \dots$ , pridodamo toj grupi i uzorke koji odgovaraju jedinicama u  $m, n, o, \dots$  - tom retku. Ako sada u  $m, n, o, \dots$  - tom nemaju jedinica izvan  $i, j, k, l, m, n, o, \dots$  - tog stupca postupak pridodavanja uzoraka u tu grupu završavamo. U suprotnom, pridodajemo toj grupi i uzorke koji odgovaraju jedinicama u  $p, r, s, \dots$  - tom retku i.t.d.

stupci



$$\textcircled{1} \quad S^1 = \{\vec{x}_i, \vec{x}_j, \vec{x}_k, \vec{x}_l, \vec{x}_s\}$$

$$\textcircled{2} \quad S^1 = \{\vec{x}_i, \vec{x}_j, \vec{x}_k, \vec{x}_l, \vec{x}_s, \vec{x}_m, \vec{x}_n\}$$

$$\textcircled{3} \quad S^1 = \{\vec{x}_i, \vec{x}_j, \vec{x}_k, \vec{x}_l, \vec{x}_s, \vec{x}_m, \vec{x}_n, \vec{x}_p\}$$

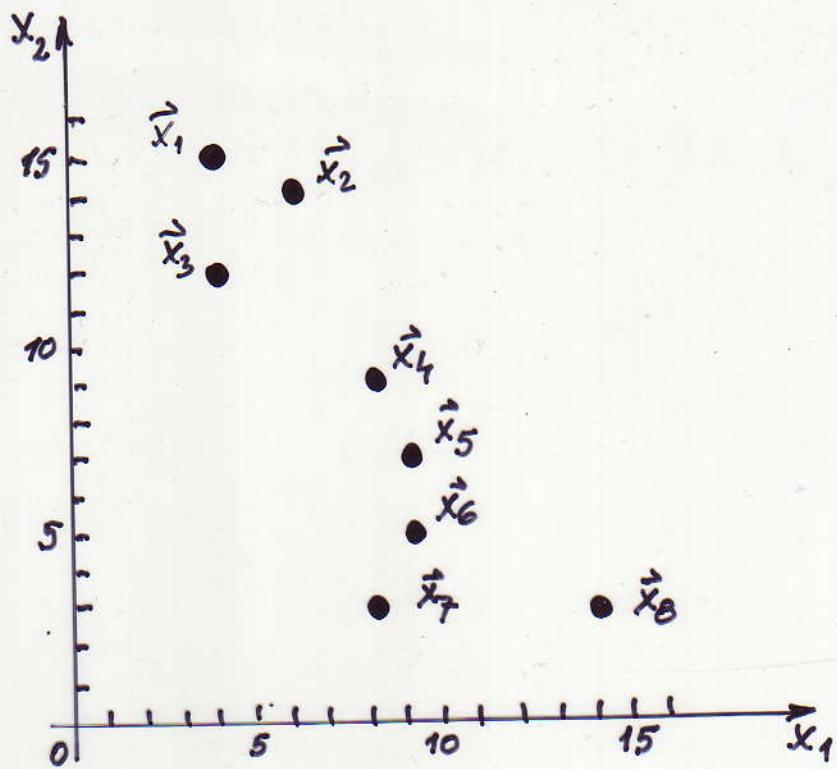
- provjeravamo redak  $p$  ako on ima jedinicu na poziciji stupca  $w$  pridodajemo  $\vec{x}_w$  grupi  $S^1$

i.t.d

### 3. korak

Brišemo redak i stupac uzorka, odnosno uzoraka koji smo već grupirali.  
Postupak ponavljamo za tako reduciraniu matricu sličnosti i nastavljamo sve dok matriča sličnosti ne nestane.

Primjer



Izračunajmo najprije  $N(N-1)/2$   
udaljenosti između uzoraka

$$\begin{aligned}
 D_{11} &= 0,00 & D_{12} &= 2,24 & D_{13} &= 3,00 & D_{14} &= 7,2 & D_{15} &= 9,43 & D_{16} &= 11,18 & D_{17} &= 13 & D_{18} &= 15,62 \\
 D_{22} &= 0,00 & D_{23} &= 2,83 & D_{24} &= 5,39 & D_{25} &= 7,62 & D_{26} &= 9,49 & D_{27} &= 11,18 & D_{28} &= 13,60 \\
 D_{33} &= 0,00 & D_{34} &= 5,00 & D_{35} &= 7,07 & D_{36} &= 8,60 & D_{37} &= 9,85 & D_{38} &= 13,45 \\
 D_{44} &= 0,00 & D_{45} &= 2,24 & D_{46} &= 4,12 & D_{47} &= 6,00 & D_{48} &= 8,40 \\
 D_{55} &= 0,00 & D_{56} &= 2,00 & D_{57} &= 4,12 & D_{58} &= 6,40 \\
 D_{66} &= 0,00 & D_{67} &= 2,24 & D_{68} &= 5,39 \\
 D_{77} &= 0,00 & D_{78} &= 6,00 \\
 D_{88} &= 0,00
 \end{aligned}$$

Ako izaberemo  $\text{H}=4$  dobivamo sljedeću matricu sličnosti:

$$S = \left[ \begin{array}{cccc|c}
 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
 \hline
 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\
 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\
 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
 \hline
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
 \end{array} \right] \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{matrix}$$

### 1. korak

Za redak koji ima najveći broj jedinica izaberimo prvi redak matrice  $S$ .  $i=1$

### 2. korak

Tvorimo grupu iz uzoraka  $\vec{x}_1, \vec{x}_2$  i  $\vec{x}_3$ . Budući da u recima 2 i 3 nemamo dodatnih jedinica izvan stupaca 1, 2, 3 grupate ravnopravno iz uzoraka  $\vec{x}_1, \vec{x}_2$  i  $\vec{x}_3$ .

$$S^1 = \{\vec{x}_1, \vec{x}_2, \vec{x}_3\}$$

3. KORAK

Tvorimo reducirana matricu tako da brišemo retke i stupce 1, 2, 3 :

$$S' = \left[ \begin{array}{cccc|c} 4 & 5 & 6 & 7 & 8 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array} \right] \begin{matrix} 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{matrix}$$

1. KORAK

Redak 5 ima najveći broj jedinica

2. KORAK

U grupu  $S^2$  uvrštavamo azorke  $\vec{x}_4, \vec{x}_5, \vec{x}_6$   
 Zato što redak 6 ima u sedmom stupcu ima jedinice uvrštavamo  $\vec{x}_7$  u grupu  $S_2$ .

Dobivamo:  $S^2 = \{\vec{x}_4, \vec{x}_5, \vec{x}_6, \vec{x}_7\}$ .

3. KORAK

Tvorimo reduciraniu matricu tako da brišemo retke i stupce 4, 5, 6 i 7. Dobivamo:

$$S'' = [1]_8$$

1. KORAK

Briamo redak 8

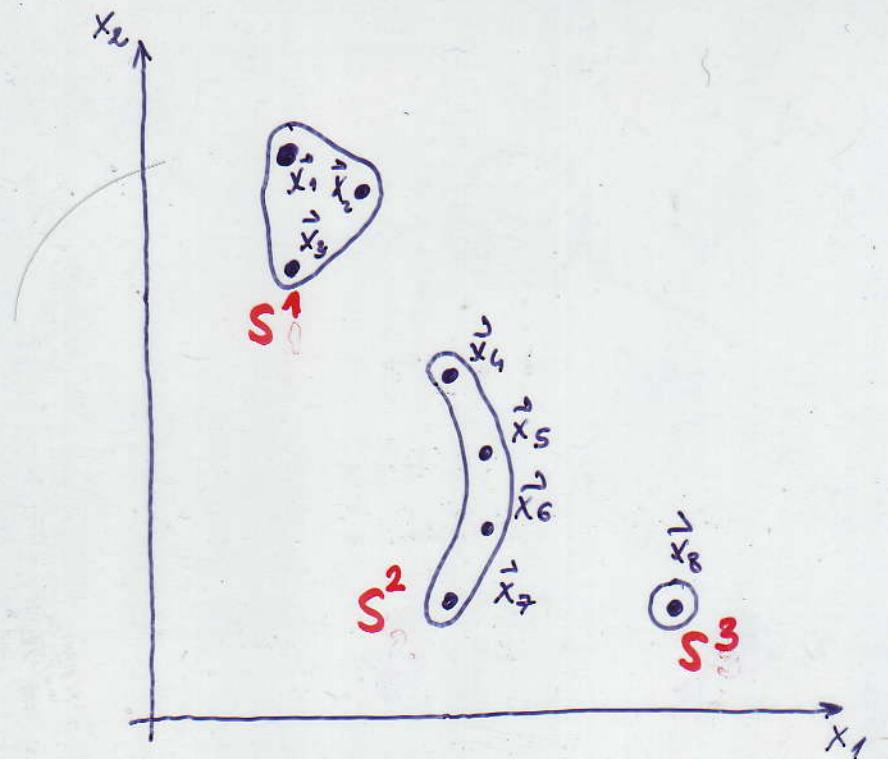
2. KORAK

Tvorimo novu grupu u koju uvrštavamo  $\vec{x}_8$ .  $S^3 = \{\vec{x}_8\}$ .

3. KORAK

Matriča  $S$  je sastavljena 8. reka nekada.

REZULTAT GRUPIRANJA:  
(za  $\Theta = 4$ )



$$S^1 = \{\vec{x}_1, \vec{x}_2, \vec{x}_3\}$$

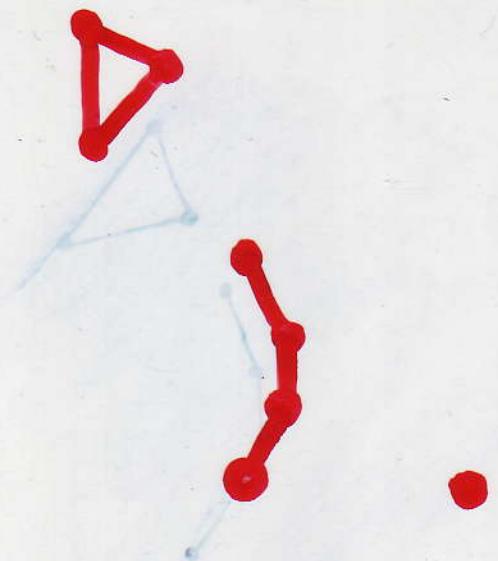
$$S^2 = \{\vec{x}_4, \vec{x}_5, \vec{x}_6, \vec{x}_7\}$$

$$S^3 = \{\vec{x}_8\}$$

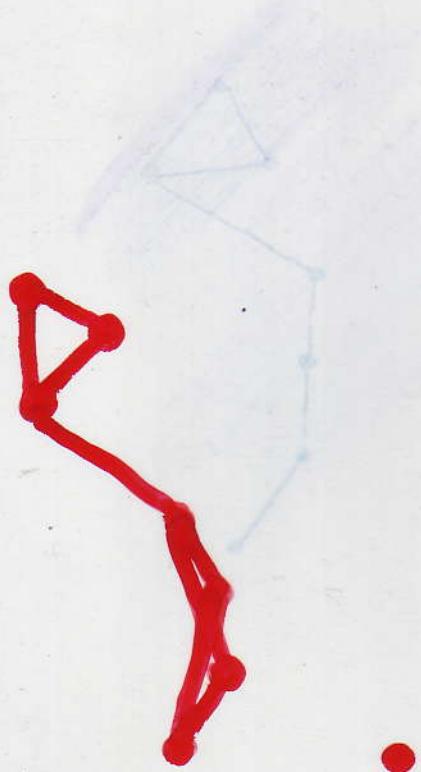
- rezultat grupiranja zavisi od praga  $\Theta$
- Prag  $\Theta$  moze se izabratи na temelju statistike uzorka;  
npr.  $\Theta = f(\mu, \sigma)$   
 $\mu$  - srednja vrijednost uzorka  
 $\sigma$  - standardna devijacija
- ako je  $\sigma$  relativno mala vrijednost u odnosu na  $\mu \rightarrow$  vjerojatnost uspjehnosti grupiranja raste (Zaku, 1971).

999

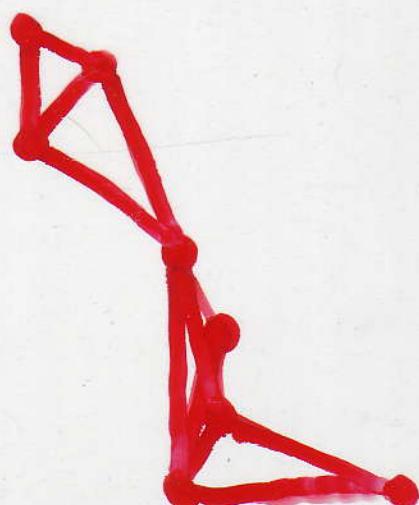
GRAFOVI GENERIRANI ZA ISTI SKUP UZORAKA  
 (U ZAVISNOSTI OD IZABRANOG PRAGA  $\Theta$ )



$$\Theta = 4$$



$$\Theta = 5$$

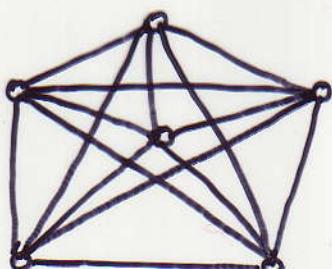


$$\Theta = 6$$

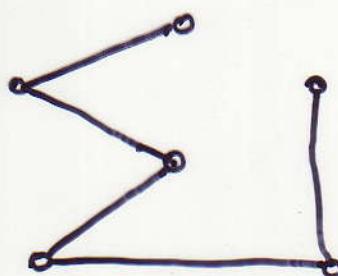
## Postupak grupiranja pomoću povezanih stabla minimalne duljine (minimal spanning tree)

- problem izbora praga  $\oplus$
- iterativno obradujemo matričnu udaljenost između uzoraka
- $N$  uzoraka iz  $S_N$  predstavljaju čvorove punog neusmjerenog grafa
- svakoj grani (luku) grafa dodijelimo težinsku vrijednost koja odgovara udaljenosti između odgovarajućih čvorova
- tvorimo povezano stablo grafa (spanning tree)  $\rightarrow$  djelomični graf koji povezuje sve čvorove grafa i ne sadrži petlju (konturu)

Primjer:



Graf



Povezano stablo grafa

Povezano stablo minimalne duljine - povezano stablo kod kojeg je ukupna duljina svih njegovih grana najmanja.

stablo - povezan graf s  $n$  ( $n \geq 1$ ) čvorova  
i  $m = n - 1$  grana

stablo je graf koji ne sadrži nijednu  
konturu (petlju)

Kontura (petlja) - konacan, povezan  
i regularan graf stupnja 2

- Graf stupnja  $r$  ima  $m = \frac{1}{2}nr$

broj  
grana

broj čvorova

- graf  $r=2$

$$m = \frac{1}{2}n \cdot 2 \rightarrow m = n$$

$$\begin{aligned} m &= 1 \\ n &= 1 \end{aligned}$$

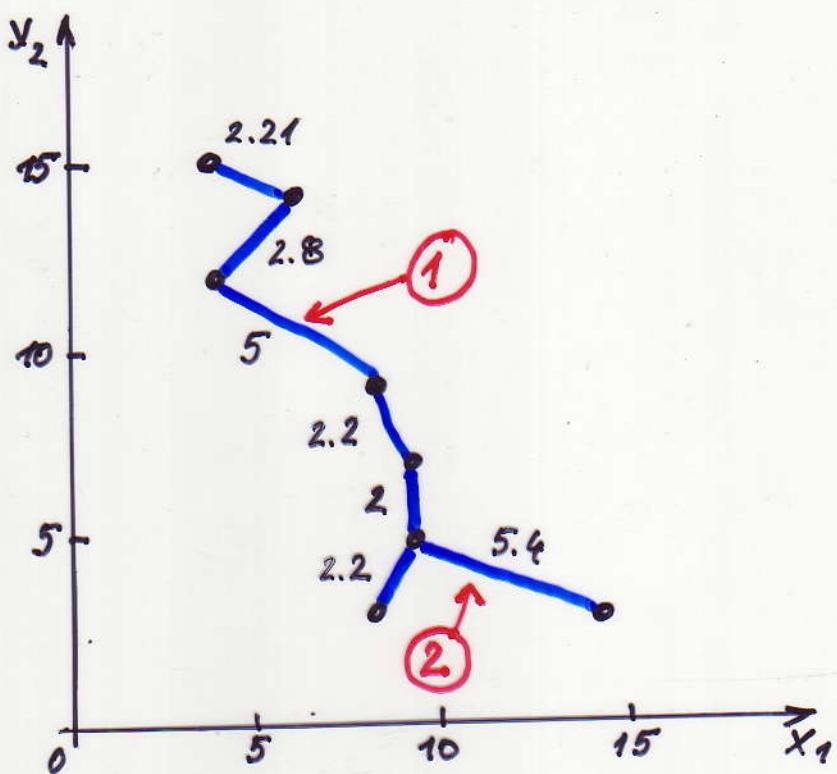


$$\begin{aligned} m &= 2 \\ n &= 2 \end{aligned}$$



- Nadimo stablo povezanosti minimalne  
dužine (Algoritam C. Zahu, 1971)

Primjer:

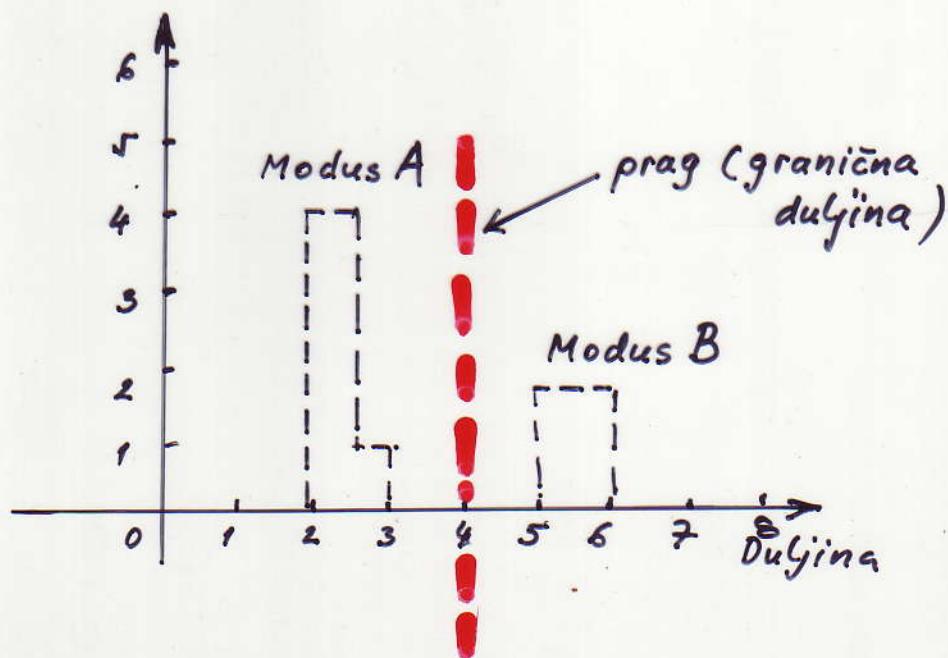


Problem grupiranja: Otkriti suvišne grane  
u stablu povezanosti minimalne  
duljine

Kako otkriti suvišne grane?

- Odrediti distribuciju duljina u stablu povezanosti minimalne duljine!
- Histogram duljina  $\rightarrow$  očekujemo bimodalni histogram
  - jedan vrh histograma odgovara duljinama koje se javljaju unutar grupe
  - drugi vrh histograma odgovara duljinama između grupa

Prag - granica duljina NALAZI SE IZMEĐU DVA MODUSA :



POSTUPCI GRUPIRANJA NA TEMELJU  
MINIMIZACIJE KRITERIJSKE FUNKCIJE  
(engl. performance index)

- Algoritam K-srednjih vrijednosti  
(engl. K-Means Algorithm)

- kriterijska funkcija:

$$J = \sum_{j=1}^{N_c} J_j,$$

gdje je:  $J_j = \sum_{\vec{x} \in S_j} \|\vec{x} - \vec{z}_j\|^2$

$N_c$  - broj grupa,  $K$

### 1. korak

zaberimo  $K \leq N$  središta grupa

$$\vec{z}_1(1), \vec{z}_2(1), \dots, \vec{z}_K(1).$$

$N$  - je broj uzoraka

### 2. korak

u  $k$ -tom koraku (iteraciji) razdijelimo uzorke  $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_N$  u  $K$  grupa pomoću relacije:

$$\vec{x} \in S_j(k) \text{ ako je } \|\vec{x} - \vec{z}_j(k)\| < \|\vec{x} - \vec{z}_i(k)\| \quad i = 1, 2, \dots, K \quad i \neq j$$

$S_j(k)$  - označava skup uzoraka čiji je centar  $\vec{z}_j(k)$ .

3. korak

Izračunavamo nova središta grupa

$$\vec{z}_j(k+1), \quad j=1, 2, \dots, K,$$

tako da je kriterijska funkcija

$$J = \sum_{j=1}^K \sum_{\vec{x} \in S_j(k)} \|\vec{x} - \vec{z}_j(k+1)\|^2 \quad j=1, 2, \dots, K$$

minimalna.

Središta grupa koja minimiziraju kriterijsku funkciju u  $k$ -toj iteraciji su ARITMETIČKE SREDNJE VRIVEDNOSTI UZORAKA POSEDINIH GRUPA

$$\vec{z}_j(k+1) = \frac{1}{N_j} \sum_{\vec{x} \in S_j(k)} \vec{x} \quad \text{za } j=1, 2, \dots, K$$

$N_j$  - broj uzoraka u grupi

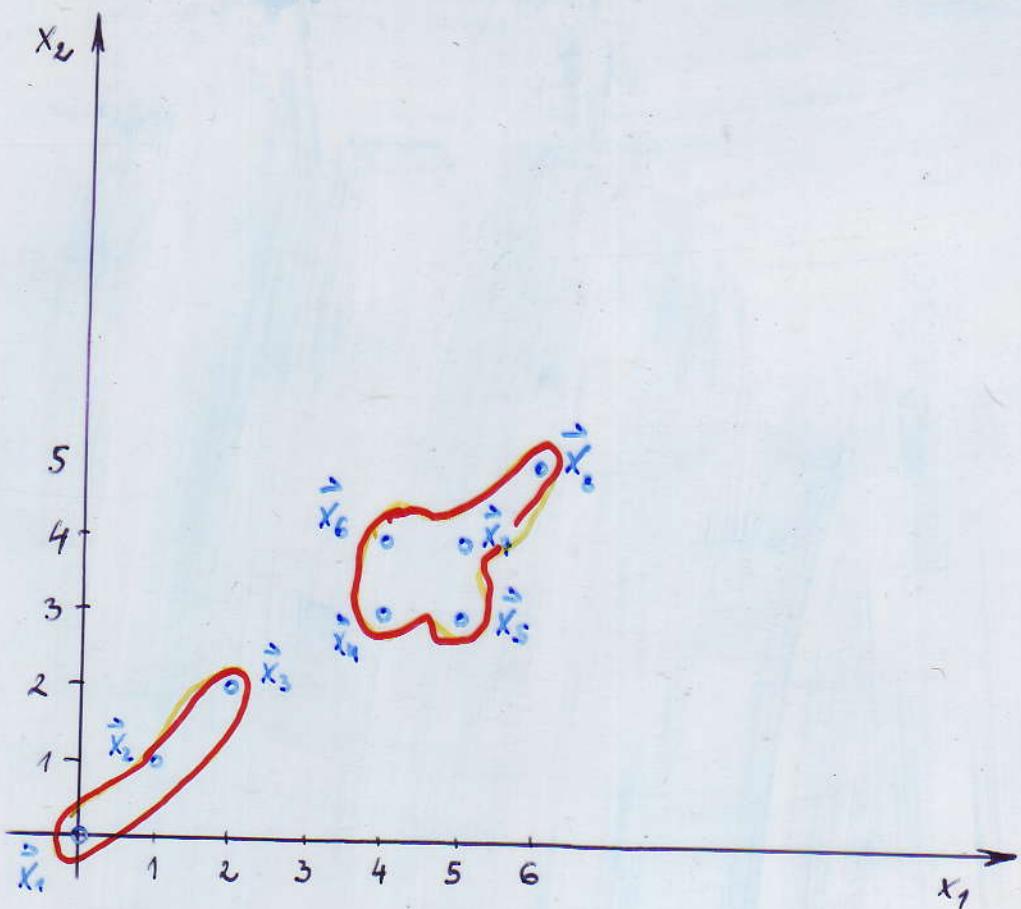
4. korak

$$\text{Ako je } \vec{z}_j(k+1) = \vec{z}_j(k) \text{ za sve}$$

$$j=1, 2, \dots, K$$

postupak završava.

Ukoliko nije, ponavljamo postupak od 2. koraka



Na rezultat grupiranja pomoću algoritma k-predjih vrijednosti utječe:

- broj grupa
- izbor početnih mrežista grupa
- redoslijed kojim se uzorci uzimaju
- geometrijska raspjata podataka

Problem konvergencije?

NEMA OPĆENITOG DOKAZA  
O KONVERGENCIJI ALGORITMA

Algoritam zahvaljujući eksperimentiranju sa različitim vrijednostima  $K$  i različitim početnim konfiguracijama!

Primjer:

$$\begin{aligned}\vec{x}_1 &= (0, 0)', & \vec{x}_5 &= (2, 1)', & \vec{x}_9 &= (6, 6)', & \vec{x}_{13} &= (7, 7)' \\ \vec{x}_2 &= (1, 0)', & \vec{x}_6 &= (1, 2)', & \vec{x}_{10} &= (3, 6)', & \vec{x}_{14} &= (8, 7)' \\ \vec{x}_3 &= (0, 1)', & \vec{x}_7 &= (2, 2)', & \vec{x}_{11} &= (8, 6)', & \vec{x}_{15} &= (9, 2)' \\ \vec{x}_4 &= (1, 1)', & \vec{x}_8 &= (3, 2)', & \vec{x}_{12} &= (6, 7)', & \vec{x}_{16} &= (2, 8)' \\ \vec{x}_{17} &= (3, 8)', & \vec{x}_{19} &= (8, 9)', \\ \vec{x}_{18} &= (9, 8)', & \vec{x}_{20} &= (9, 9).'\end{aligned}$$

1. KORAK

$$\underline{k=2}; \quad \vec{z}_1(1) = \vec{x}_1 = (0, 0)' ; \quad \vec{z}_2(1) = \vec{x}_2 = (1, 0)'$$

2. KORAK

Buduci da je  $\|\vec{x}_1 - \vec{z}_1(1)\| < \|\vec{x}_1 - \vec{z}_i(1)\|$  i

$$\|\vec{x}_3 - \vec{z}_1(1)\| < \|\vec{x}_3 - \vec{z}_i(1)\|$$

$i=2$  imamo :

$$S_1(1) = \{\vec{x}_1, \vec{x}_3\}$$

$$S_2(1) = \{\vec{x}_2, \vec{x}_4, \dots, \vec{x}_{20}\};$$

3. KORAK

Racunamo nova središta grupa:

$$\vec{z}_1(2) = \frac{1}{N_1} \sum_{\vec{x} \in S_1(1)} \vec{x} = \frac{1}{2} (\vec{x}_1 + \vec{x}_3) = \begin{pmatrix} 0.0 \\ 0.5 \end{pmatrix}$$

$$\begin{aligned}\vec{z}_2(2) &= \frac{1}{N_2} \sum_{\vec{x} \in S_2(1)} \vec{x} = \frac{1}{18} (\vec{x}_2 + \vec{x}_4 + \dots + \vec{x}_{20}) \\ &= \begin{pmatrix} 5.67 \\ 5.33 \end{pmatrix}\end{aligned}$$

#### 4. KORAK

Budući da je  $\vec{z}_j(2) \neq \vec{z}_j(1)$ ;  $j=1, 2$   
 vracamo se na KORAK 2.

#### 2'. KORAK

$$\|\vec{x}_e - \vec{z}_1(2)\| < \|\vec{x}_e - \vec{z}_1(1)\| \text{ eq}$$

$$e = 1, 2, \dots, 8 \quad i$$

$$\|\vec{x}_e - \vec{z}_2(2)\| < \|\vec{x}_e - \vec{z}_2(1)\| \text{ eq}$$

$$e = 9, 10, \dots, 20$$

$$S_1(2) = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_8\}$$

$$S_2(2) = \{\vec{x}_9, \vec{x}_{10}, \dots, \vec{x}_{20}\}$$

#### 3'. KORAK

Obnovimo vrijednosti centara:

$$\vec{z}_1(3) = \frac{1}{N_1} \sum_{\vec{x} \in S_1(2)} \vec{x} = \frac{1}{8} (\vec{x}_1 + \vec{x}_2 + \dots + \vec{x}_8)$$

$$= \begin{pmatrix} 1.25 \\ 1.13 \end{pmatrix}$$

$$\vec{z}_2(3) = \frac{1}{N_2} \sum_{x \in S_2(2)} \vec{x} = \frac{1}{12} (\vec{x}_9 + \vec{x}_{10} + \dots + \vec{x}_{20})$$

$$= \begin{pmatrix} 7.62 \\ 7.33 \end{pmatrix}$$

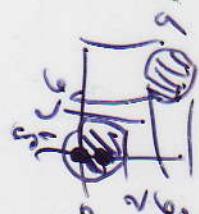
4. KORAK

Budući da  $\vec{z}_j(3) \neq \vec{z}_j(2)$  za  $j=1, 2$   
vraćamo se na Korak 2.

2. KORAK ; 3. KORAK

Daje isti rezultat kao u prethodnoj

iteraciji :  $\vec{z}_1(4) = \vec{z}_1(3)$   
 $\vec{z}_2(4) = \vec{z}_2(3)$



4. KORAK  $\vec{z}_j(4) = \vec{z}_j(3)$  za  $j=1, 2$

algoritam je konvergirao i dao slijedeće

centre grupa :

$$\vec{z}_1 = \begin{pmatrix} 1.25 \\ 1.13 \end{pmatrix} ; \vec{z}_2 = \begin{pmatrix} 7.62 \\ 7.33 \end{pmatrix}$$

T. Kohonen, The "Neural" Phonetic Typewriter, IEEE Computer Vol. 21, No. 3, March 1988, pp. 11-22

### Shortcut learning algorithm

- slučajno izaberimo početne vrijednosti  $m_i$ :  $m_i(0)$
- za  $t = 0, 1, 2, \dots$  izračunajmo

(1) centar mjehanica ( $c$ ):

$$\|x(t) - m_c(t)\| = \min_c \{\|x(t) - m_i(t)\|\}$$

(2) "popravimo" vrijednost težiških vektora

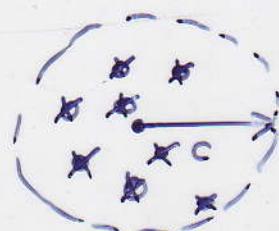
$$m_i(t+1) = m_i(t) + \alpha(t) (x(t) - m_i(t))$$

za  $i \in N_c$ .

$N_c$  - ~~slo~~ (broj elemenata) u radijusu  $c$  "mjehanic":

$$m_i(t+1) = m_i(t)$$

za sve ostale  $i$



$\alpha = \alpha(t)$  ;  $N_c = N_c(t)$  su empirijske funkcije vremena



Monotonu padačicu funkcija vremena

$c \rightarrow$  pada monotono

$t_1$  - threshold values  
 $t_2$  -  
 $t_3$  -

## ISODATA

$t_1$  - splitting,  
 $t_2$  - merging,  
 $t_3$  - discarding,  
(D. W. Petterson, 1990)

1. Select  $m$  samples as seed points for initial cluster centers.

This can be done by taking the first  $m$  points, selecting random points or by taking the first  $m$  points which exceed some mutual minimum separation distance  $d$ .

2. Group each sample with its nearest cluster center.

3. After all samples have been grouped, compute new cluster centers for each group. The center can be defined as the centroid (mean value of the attribute vectors) or some similar centre measure.

4. If the split threshold  $t_1$  is exceeded for any cluster, split it into two parts and recompute new cluster centers.

5. If the distance between two cluster centers is less than  $t_2$ , combine the clusters and recompute new cluster centers.

6. If a cluster has fewer than  $t_3$  members, discard the cluster. It is ignored for the remainder of the process.
7. Repeat steps 3 through 6 until no change occurs among cluster groupings or until some iteration limit has been exceeded.

## - ALGORITAM ISODATA

( Iterative - Self - Organizing Data A )

( B.H. Ball, 1965 )

- sličan algoritmu K-srednjih vrijednosti

RAZLIKA:

- OMOGUĆAVA DA SE U FAZI IZVOĐENJA

ALGORITMA MIENJA BROJ GRUPA

Postupak ISODATA povezuje iterativnu minimizaciju kriterijske funkcije s heurističkim postupcima grupiranja.

### 1. KORAK

Izaberimo skup od  $N_c$  početnih središta grupa:

$$\vec{m}_1, \vec{m}_2, \dots, \vec{m}_{N_c}$$

Taj skup ne treba biti jednak broju željenu središta grupa i tvori se izborom uzoraka iz skupa podataka.

### 2. KORAK

Izaberimo vrijednosti sljedećih parametara

K - proizvoljan konacni broj. (željenu) grupa,

$\textcircled{H}_N$  - najmanji, još dozvoljen, broj uzoraka u grupi,

$\textcircled{H}_s$  - granicna vrijednost standardne devijacije,

$\textcircled{H}_c$  - parametar objedinjavanja središta grupa,

L - najveci broj parova grupa koje mozemo objediti;

I - dozvoljen broj ponavljanja (iteracija),

3. KORAK

Grupiramo uzorke  $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_N$  na slijedeći način:

$\vec{x} \in S_j$ , ako je  $\|\vec{x} - \vec{m}_j\| < \|\vec{x} - \vec{m}_i\|$  za  $i = 1, 2, \dots, N_c$  i  $i \neq j$

4. KORAK

Istegćujemo grupe s manje od  $N_N$  uzorka,

odnosno zanemarujemo  $S_j$  ako je  $N_j < N_N$

za svaki  $j$ , te smajćujemo  $N_c$  za 1.

Ako u ovom koraku postupka nastupa istegćivanje grupa nastavljamo s korakom 3, a suprotnom nastavljamo sa 5. korakom!

5. KORAK

Računamo aritmetičke sredine grupa  $S_j$ :

$$\vec{m}_j = \frac{1}{N_j} \sum_{\substack{\vec{x} \\ \vec{x} \in S_j}} \vec{x} \quad ; \quad j = 1, 2, \dots, N_c$$

6. KORAK

Računamo srednje vrijednosti srednjih udaljenosti između uzorka u grupi i aritmetičkih sredina

grupa:

$$\bar{D} = \frac{1}{N} \sum_{j=1}^{N_c} N_j \bar{D}_j, \text{ gde je } j \in$$

overall average  
deviance of the  
samples from  
their respective  
cluster center

$$\bar{D}_j = \frac{1}{N_j} \sum_{\substack{\vec{x} \\ \vec{x} \in S_j}} \|\vec{x} - \vec{m}_j\| \quad j = 1, 2, \dots, N_c$$

srednja udaljenost uzorka u grupi od odgovarajućeg  
medijita grupe

7. KORAK

Imao na raspolaganju četiri mogućnosti:

a) ako je to zadnja iteracija postavljamo  $H_c = 0$   
i nastavljamo sa 11. korakom

b) ako je  $N_c \leq K/2$  nastavljamo postupak sa 8. kor.

c) ako je  $N_c > K$  ili ako je to parna iteracija  
nastavljamo sa korakom 11.

d) ako nije ni jedan uvjet (od gore uvedenih)  
ispunjeno nastavljamo postupak sa 8. korakom

8. KORAK

za svaku grupu odredujemo vektor standardne  
devijacije

$$\vec{G}_j' = (\tilde{G}_{1j}, \tilde{G}_{2j}, \dots, \tilde{G}_{ij}, \dots, \tilde{G}_{nj})^T,$$

$$\text{gdje je } \tilde{G}_{ij} = \sqrt{\frac{1}{N_j} \sum_{\vec{x} \in S_j} (x_{ik} - m_{ij})^2}$$

dimenzionalnost  
vektora  
↓  
 $i = 1, 2, \dots, n$   
 $j = 1, 2, \dots, N_c$

pri tome je  $x_{ik}$  - i-ta komponenta k-tog uzorka  
u grupi  $S_j$ ,  $m_{ij}$  i-ta komponenta aritmetičke  
sredine  $\vec{m}_j$  uzorka u grupi  $S_j$ .

Svaka komponenta  $\tilde{G}_{ij}$  predstavlja standardnu  
odstupanje uzorka u grupi  $S_j$  uzduž  
odgovarajuće koordinatne ose.

9. KORAK

U svakom vektoru  $\tilde{G}_j ; j=1, 2, \dots, N_c$  potražimo najveću komponentu i označimo je sa  $G_{j\max}$ .

10. KORAK

Ako je za neki  $G_{j\max} ; j=1, 2, \dots, N_c$   $G_{j\max} > \Theta_s$  i ako je ispunjen uvjet

$$(a) \quad \bar{D}_j > \bar{D} \quad \text{i} \quad N_j > 2(\Theta_n + 1)$$

ili uvjet

$$(b) \quad N_c \leq K/2$$

tada možemo "rasprijiti" aritmetičku sredinu  $\vec{m}_j$  u dva nove aritmetičke sredine

$$\vec{m}_j^+ \quad \text{i} \quad \vec{m}_j^-$$

brišemo  $\vec{m}_j$  i povećavamo  $N_c$  za 1.

Sredinu nove grupe  $\vec{m}_j^+$  dobivamo pribrajanjem određene vrijednosti  $y_j$  onoj komponenti  $\vec{m}_j$  koja ima najveće standardno odstupanje.

Sredinu nove grupe  $\vec{m}_j^-$  tvorimo tako da istoj komponenti  $\vec{m}_j$  oduzmemos  $y_j$ .

Obično uzimamo:  $y_j = k G_{j\max}$ ,  $0 < k \leq 1$

Ako u ovom koraku nastane dijeljenje

$\vec{m}_i$  nastavljamo sa 3. korakom, a suprotnom nastavljamo sa 11. korakom.

### 11. KORAK

Računamo udaljenost između parova aritmetičkih sredina grupa:

$$D_{ij} = \|\vec{m}_i - \vec{m}_j\| \quad i=1, 2, \dots, N_c-1; \\ j=i+1, i+2, \dots, N_c$$

### 12. KORAK

Udaljenosti  $D_{ij}$  uspoređujemo s parametrom  $\Theta_c$ .

L najmanjih vrijednosti koje su manje od  $\Theta_c$  razvrstavamo po padajućim vrijednostima:

$$D_{i_1 j_1} < D_{i_2 j_2} < \dots D_{i_L j_L}$$

pri čemu je L najveći broj parova sredista grupa koje možemo objединити.

### 13. KORAK

Objedinjavajuće aritmetičke sredine grupa započinjuju se kod najmanje udaljenosti među sredinama  $D_{i_l j_l}$ .

Ako za svaki  $l=1, 2, \dots, L$  sredine  $\vec{m}_{ie}$  i  $\vec{m}_{je}$  nisu bile objedinjene u toj iteraciji, objedinjavamo ih po sljedećem obrazcu:

14. KORAK  $\left\{ \begin{array}{l} \text{sred.} \\ \text{korak 2} \\ \text{i 3. korak} \end{array} \right.$

$$\vec{m}_l = \frac{1}{N_{ie} + N_{je}} [N_{ie} \vec{m}_{ie} + N_{je} \vec{m}_{je}]$$