

Grupiranje (engl. Clustering)

Uvod u raspoznavanje uzoraka

ak. god. 2019/2020

Prof. dr. sc. Slobodan Ribarić

Grupiranje (engl. Clustering)

Do sada radili s „nadgledanom klasifikacijom” ili učili klasifikator „učiteljem” na temelju označenih uzoraka (uzoraka za vježbanje)

Grupiranje → barata neoznačenim uzorcima za koje ne znamo pripadnost razredu niti broj razreda iz kojih uzorci „dolaze” (učenje bez učitelja!)

Zadatak grupiranja: Otkriti organizaciju uzoraka i grupirati uzorke u smislene („prirodne”) grupe koje će nam omogućiti otkrivanje sličnosti i različitosti između uzoraka (i grupa) i time dopustiti izvođenje korisnih zaključaka o njima.

Grupiranje (engl. Clustering)

Ovakav pristup se koristi u sljedećim područjima:

- biologiji i zoologiji,
- psihijatriji i patologiji,
- sociologiji,
- arheologiji,
- geologiji,
- tehnici (računalnom vidu, strojnom učenju, umjetnoj inteligenciji, ...)

Grupiranje (engl. Clustering)

Grupiranje – nenadgledano učenje (engl. Unsupervised learning); učenje bez učitelja

Primjer:

Razmotrimo sljedeće životinje:

ovca, pas, mačka;

vrabac, galeb;

zmija, gušter;

zlatna ribica, skuša;

žaba;

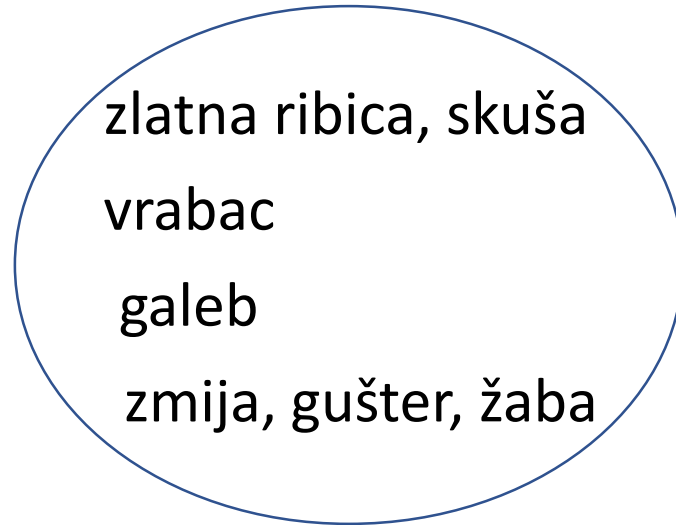
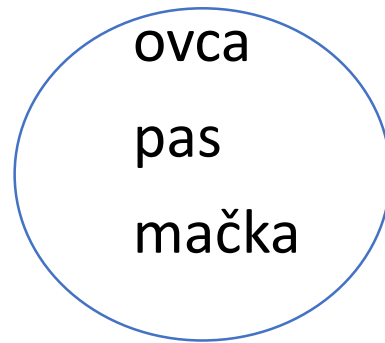
- Organizirajmo ih u grupe (ili tzv. "prirodne razrede")!

VAŽNO: kriterij grupiranja???

Grupiranje (engl. Clustering)

Kriterij grupiranja: da li ženke nose svoju (buduću) mladunčad?

Grupe:



Grupiranje (engl. Clustering)

Da li imaju pluća?

zlatna ribica, skuša

ovca, pas, mačka, gušter, galeb,
žaba, zmija

Okoliš u kojem žive:

ovca, pas, mačka, zmija, galeb, vrabac
gušter

žaba

zlatna ribica
skuša

Grupiranje (engl. Clustering)

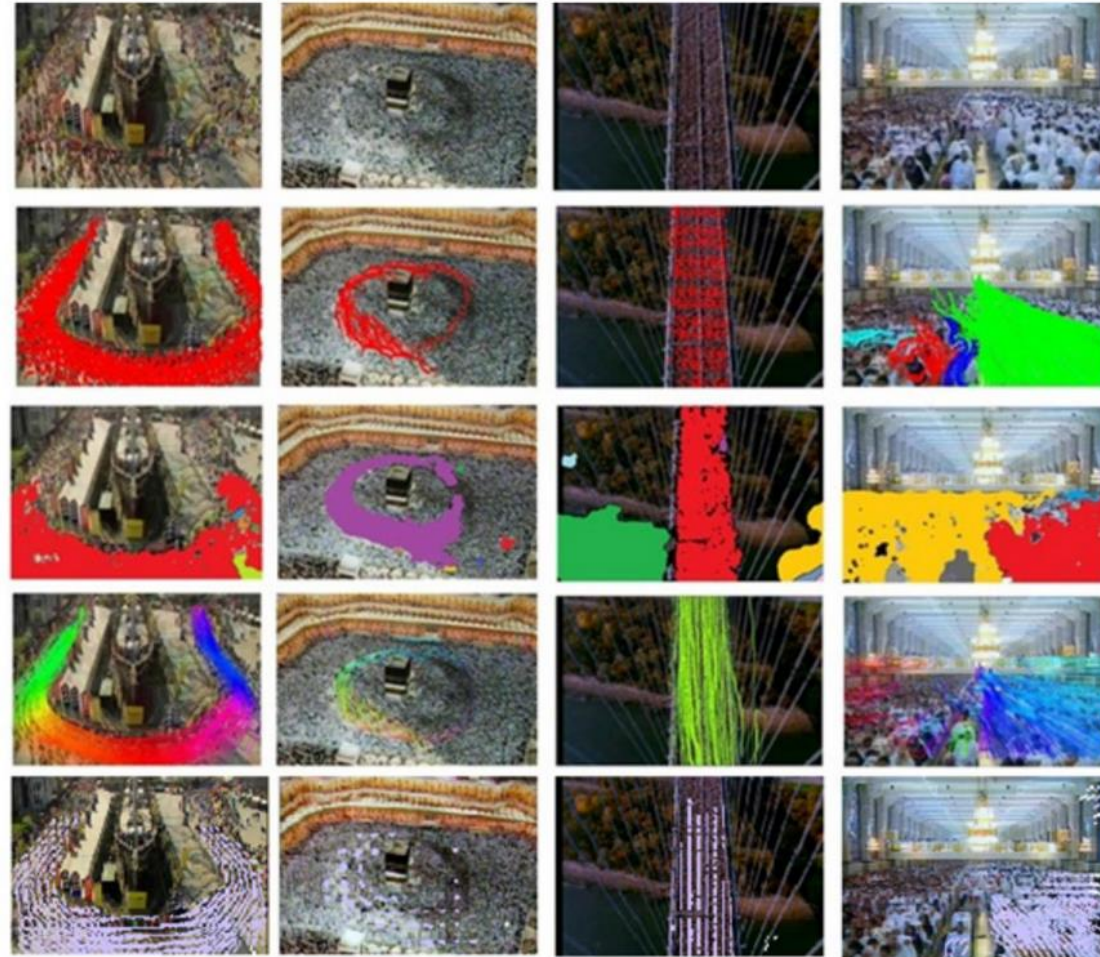
Osnovni koraci u postupku grupiranja

- i) izbor značajki
- ii) izbor mjere sličnosti ili različitosti
- iii) kriterij grupiranja (zavisi od tumačenja eksperta čemu daje naglasak na „smislenom” razvrstavanju neoznačenih uzoraka)
- iv) algoritam grupiranja
- v) validacija/vrednovanje rezultata
- vi) tumačenje dobivenih rezultata

Grupiranje (engl. Clustering)

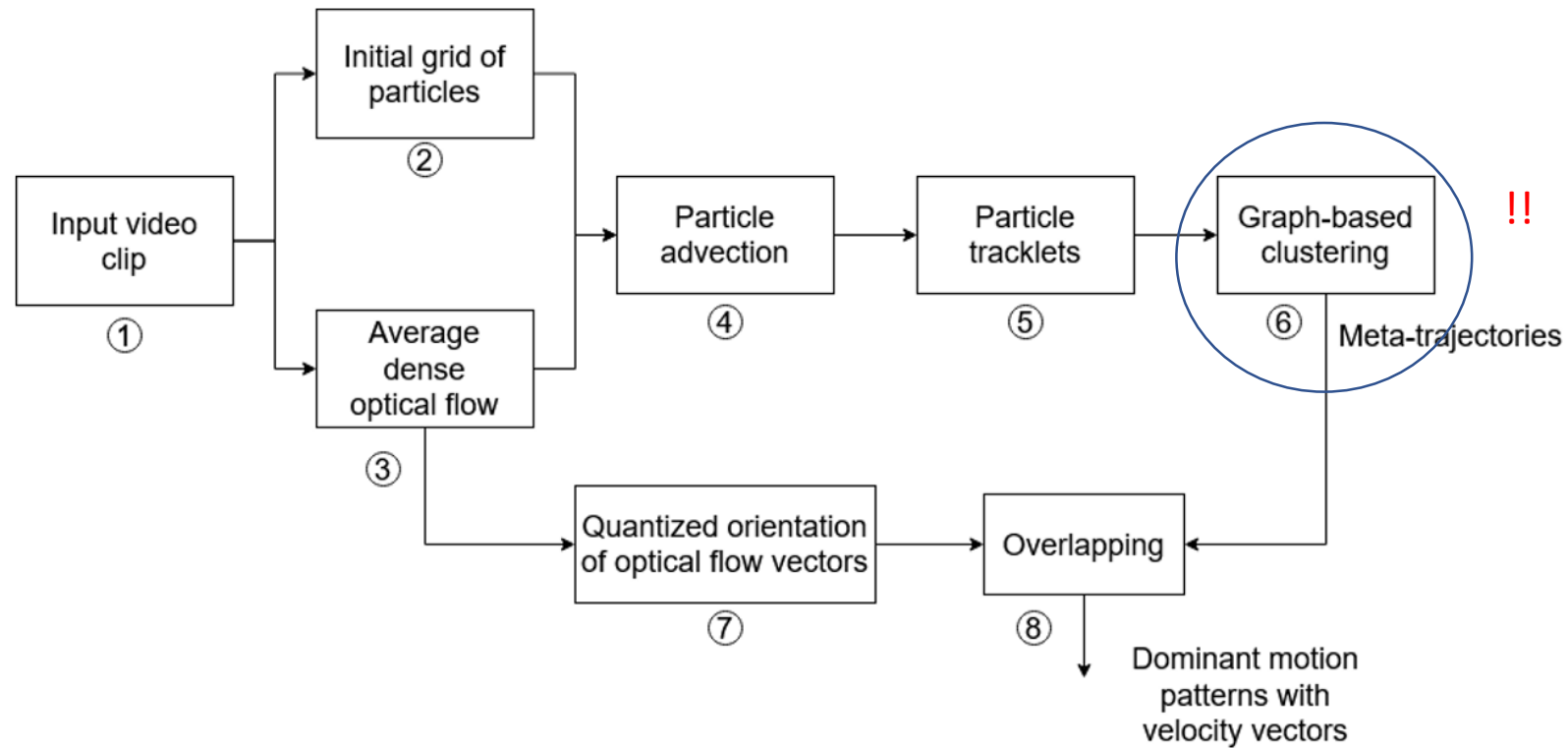
Primjer: Računalni vid

Grupiranje mnoštva ljudi
na temelju značajki kretanja
/ulaz video sekvence/



Grupiranje (engl. Clustering)

Primjer (nastavak):



Grupiranje (engl. Clustering)

Pored koraka i) – vi) često se koristi još jedan dodatni korak:

vii) uporaba testova koji trebaju pokazati da li raspoloživi podaci imaju „strukturu” grupe (npr. skup podataka može biti potpuno slučajne prirode te pokušaj otkrivanja „smislenih” grupa je besmislen

Pozor: koraci i) – vi) su podložni subjektivnosti eksperta

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

Grupiranje (engl. Clustering)

Koliko je grupa?

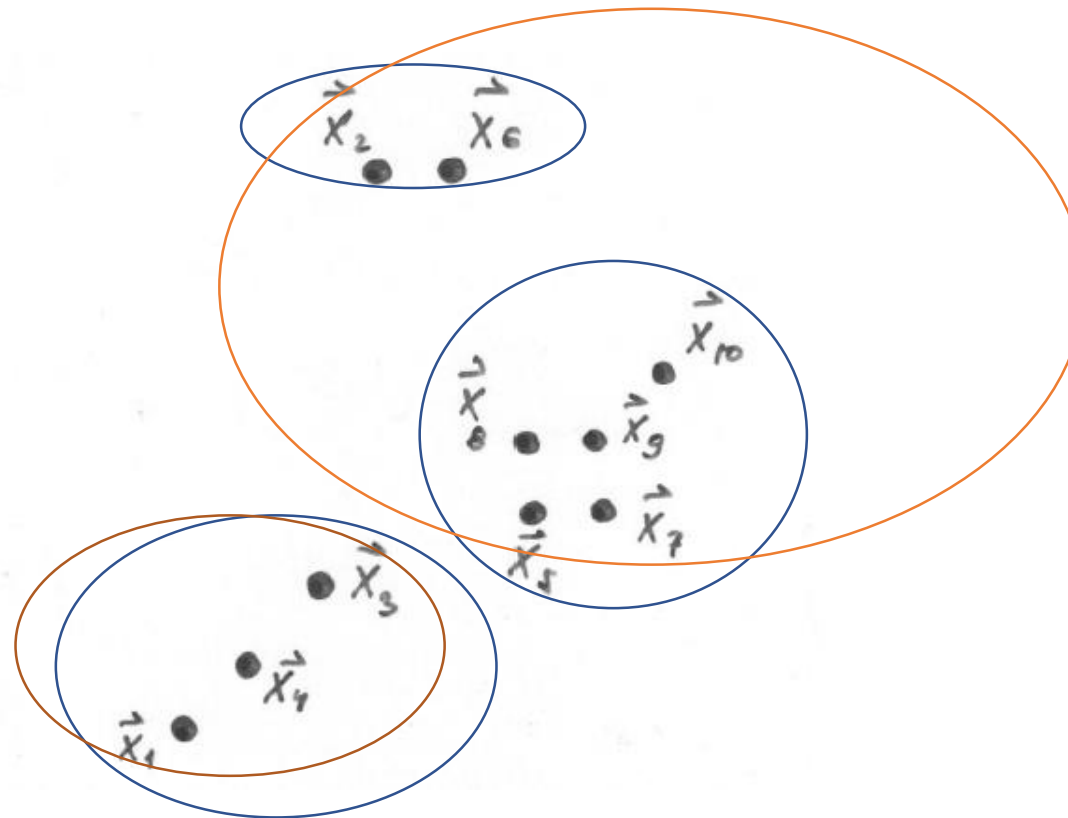


Grupiranje (engl. Clustering)

Primjer: koliko je grupa?

Tri?

Dvije?



Grupiranje (engl. Clustering)

Primjeri uporabe grupiranja

- **redukcija raspoloživih podataka** $N \gg 1$ podataka, dobivamo $m \ll N$ grupa
- **generiranje hipoteza**: koristimo analizu grupa (grupiranje) u cilju utvrđivanja i zaključivanja u vezi prirode podataka, grupiranje – poticaj za postavljanje hipoteze
- ispitivanje hipoteza
- predviđanje na temelju rezultata grupiranja (npr. analiza grupa primijenjena na skupu podataka o pacijentima koji su oboljeli od iste bolesti)

rezultat – broj grupa pacijenata prema njihovoj reakciji na određene lijekove; novi pacijent – za njega možemo identificirati odgovarajuću grupu!

Grupiranje (engl. Clustering)

Vrste značajki:

- mogu zauzimati vrijednosti iz nekog kontinuiranog opsega (podskup od R) ili iz nekog konačnog diskretnog skupa; (ako je konačni skup diskretan i ima samo dva elementa tada se značajka naziva **binarna** ili **dihotomna**)

Klasifikacija značajki na temelju relativnog značaja vrijednosti koje mogu zauzimati:

- i) nominalne
- ii) uređene
- iii) intervalno skalirane
- iv) skalirane omjerom

Grupiranje (engl. Clustering)

Primjer:

- i) nominalne: spol osoba; npr. 1 za muškarce; 0 za žene (ili obratno) – kvantitativno uspoređivanje između tih (nominalnih) vrijednosti nema smisla
- ii) uređene: karakterizacija sposobnosti: 5, 4, 3, 3, 2, 1
odličan, vrlo dobar, dobar dovoljan, nedovoljan
- iii) intervalno skalirane: npr. mjerenje temperature u stupnjevima Celzijusa: Pariz, London; smisleno je reći da je temperatura u Parizu za viša od one u Londonu
- iv) skalirane omjerom: omjer između značajki ima smisla, npr. osoba teška 100 kg je dvaput teža od osobe koja teži 50 kg.

Grupiranje (engl. Clustering)

Definicija grupiranja (Everitt, 1981.)

- Ako su značajke predložene vektorom (točka u n- dimenzionalnom prostoru) onda su **grupe** kontinuirana područja koja imaju veliku gustoću točaka i **odvojene su** od drugih kontinuiranih prostora velikih gustoća s područjem relativno **malih gustoća** točaka
- takve grupe se nazivaju i „prirodne grupe” (engl. Natural clusters)

Grupiranje (engl. Clustering)

Formalna definicija grupiranja:

- neka je X skup podataka: $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$
- m - grupiranje X odgovara dijeljenju X u m skupova (grupa): G_1, G_2, \dots, G_m
- Pri čemu su zadovoljena sljedeća tri svojstva:

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

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

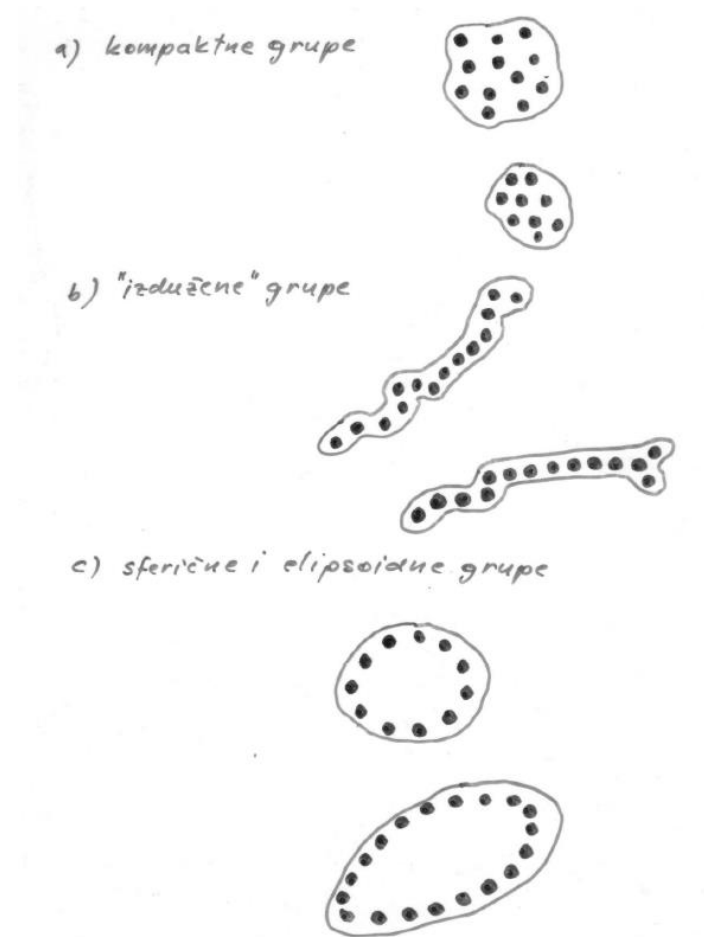
Izrazito, jasno grupiranje (engl. Crisp)

iii) $G_i \cap G_j = \emptyset$ za $i \neq j, i, j = 1, 2, \dots, m$

VAŽNO: vektori sadržani u grupi G_i su „sličniji” jedan drugome i „manje slični” vektorima iz drugih grupa!

Grupiranje (engl. Clustering)

Kvantifikacija izraza „sličan” i „različit” zavisi od tipa grupe:



Grupiranje (engl. Clustering)

Drugačiji pristup grupiranju

Skup podataka X se grupira u m grupa a grupiranje je određeno s m funkcija u_j :

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

i

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

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

Neizrazito, nejasno grupiranje (engl. Fuzzy)
/svaki vektor iz \mathbf{x}_i pripada istodobno više od
jednoj grupi s nekom mjerom pripadnosti iz
intervala $[0, 1]$ /

$u_j, \quad j = 1, 2, \dots, m$ funkcija pripadnosti (engl. membership function)

Grupiranje (engl. Clustering)

Mjere bliskosti (sličnosti) /engl. Proximity measures/

Mjere različitosti /engl. Dissimilarity measures DM/

Metrika - metričke funkcije /njihova svojstva već poznajemo/:

l_p – metrika

$$d_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^n w_i |x_i - y_i|^p \right)^{1/p}$$

$$w_i \geq 0$$

ako je $w_i = 1$ za $i = 1, 2, \dots, n$ - "netežinska" metrika

Grupiranje (engl. Clustering)

Euklidska udaljenost

l_2 – metrika

$$d_2(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^n (x_i - y_i)^2 \right)^{1/2}$$

Manhattan norma:

l_1 – metrika

$$d_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n w_i |x_i - y_i|$$

Grupiranje (engl. Clustering)

l_∞ – norma

$$d_\infty(\mathbf{x}, \mathbf{y}) = \max_{1 \leq i \leq n} w_i |x_i - y_i|$$

Mahalanobisova udaljenost

$$mdi^2 = (\mathbf{x} - \boldsymbol{\mu}_i)^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)$$

Σ – kovarijacijska matrica

$\boldsymbol{\mu}$ – srednja vrijednost

Grupiranje (engl. Clustering)

Mjere sličnosti

Skalarni produkt

$$S_{inner}(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y} = \sum_{i=1}^n x_i y_i$$

koristi se kada su vektori normalizirani

Tanimoto mjera (udaljenost)

$$S_T(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x}^T \mathbf{y}}$$

vrijednostima

za vektore realne i vektore s diskretnim

Grupiranje (engl. Clustering)

Fuova mjera sličnosti

$$S_c(\mathbf{x}, \mathbf{y}) = 1 - \frac{d_2(\mathbf{x}, \mathbf{y})}{\|\mathbf{x}\| + \|\mathbf{y}\|}$$

- kada S_c ima maksimalnu vrijednost a kada minimalnu?

Grupiranje (engl. Clustering)

Postupci grupiranja

- i) heuristički – vođeni intuicijom i iskustvom
- ii) oni koji se temelje na optimizaciji neke kriterijske funkcije ili indeksa (engl. Performance-indeks)
- iii) kombinacija pristupa i) i ii)

Grupiranje (engl. Clustering)

ii) oni koji se temelje na optimizaciji neke kriterijske funkcije (engl. Performance-index)

Najčešći kriterij:

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

N_c – broj grupa

S_j – skup uzoraka koji pripadaju j – toj grupi

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

N_j – broj uzoraka u grupi S_j

Grupiranje (engl. Clustering)

i) Heuristički postupci grupiranja

- jednostavan algoritam grupiranja

Imamo N neoznačenih uzoraka: $C = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$

1. korak

Izaberi nenegativan broj (prag) T

2. korak

Izaberi bilo koji uzorak iz $C = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ i proglasi ga središtem grupe \mathbf{z}_1
- pretpostavimo da smo izabrali $\mathbf{x}_1 = \mathbf{z}_1$

3. korak

Izračunamo udaljenost $d_2(\mathbf{z}_1, \mathbf{x}_2)$ i uspoređujemo je s pragom T

Grupiranje (engl. Clustering)

3. korak

Izračunamo udaljenost $d_2(\mathbf{z}_1, \mathbf{x}_2)$ i uspoređujemo je s pragom T

a) ako je $d_2(\mathbf{z}_1, \mathbf{x}_2) > T$, proglašavamo \mathbf{x}_2 središtem nove grupe:

$$\mathbf{z}_2 = \mathbf{x}_2$$

b) ako je $d_2(\mathbf{z}_1, \mathbf{x}_2) \leq T$ onda \mathbf{x}_2 dodjeljujemo grupi sa središtem \mathbf{z}_1

- pretpostavimo da je $d_2(\mathbf{z}_1, \mathbf{x}_2) > T$, $\mathbf{z}_2 = \mathbf{x}_2$

Grupiranje (engl. Clustering)

4. korak

Računamo udaljenosti $d_2(\mathbf{z}_1, \mathbf{x}_3)$ i $d_2(\mathbf{z}_2, \mathbf{x}_3)$

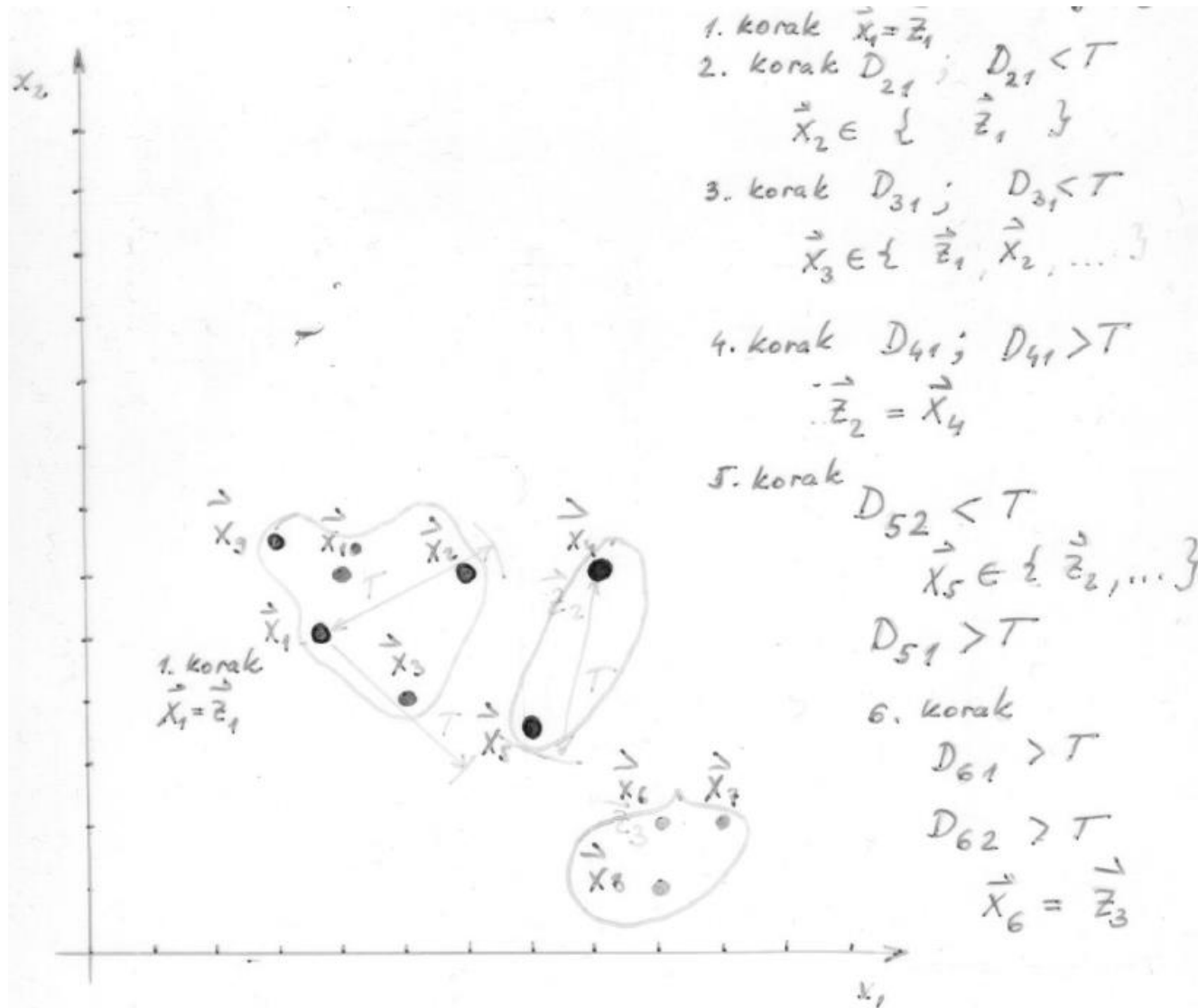
a) ako su $d_2(\mathbf{z}_1, \mathbf{x}_3)$ i $d_2(\mathbf{z}_2, \mathbf{x}_3) > T$ formiramo središte nove grupe $\mathbf{z}_3 = \mathbf{x}_3$

b) ako nije a) onda se \mathbf{x}_3 dodjeljuje grupi čijem je središtu najbliži

Postupak se provodi dok se ne obradi svih N uzoraka;

Grupiranje (engl. Clustering)

Primjer:



T vrijednost praga T

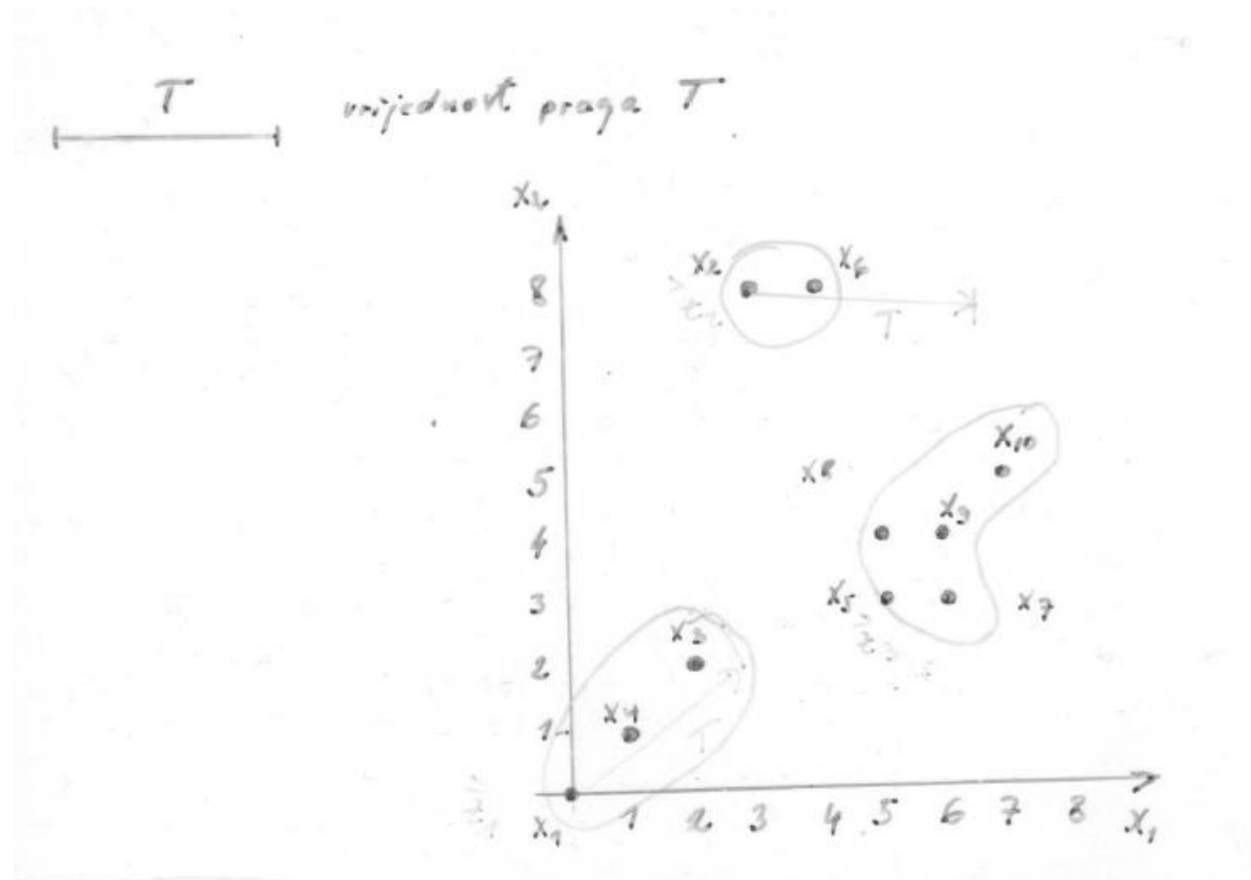
Grupiranje (engl. Clustering)

Primjer:

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

$$\mathbf{z}_2 = \mathbf{x}_2$$

$$\mathbf{z}_3 = \mathbf{x}_5$$



Grupiranje (engl. Clustering)

Prednost algoritma:

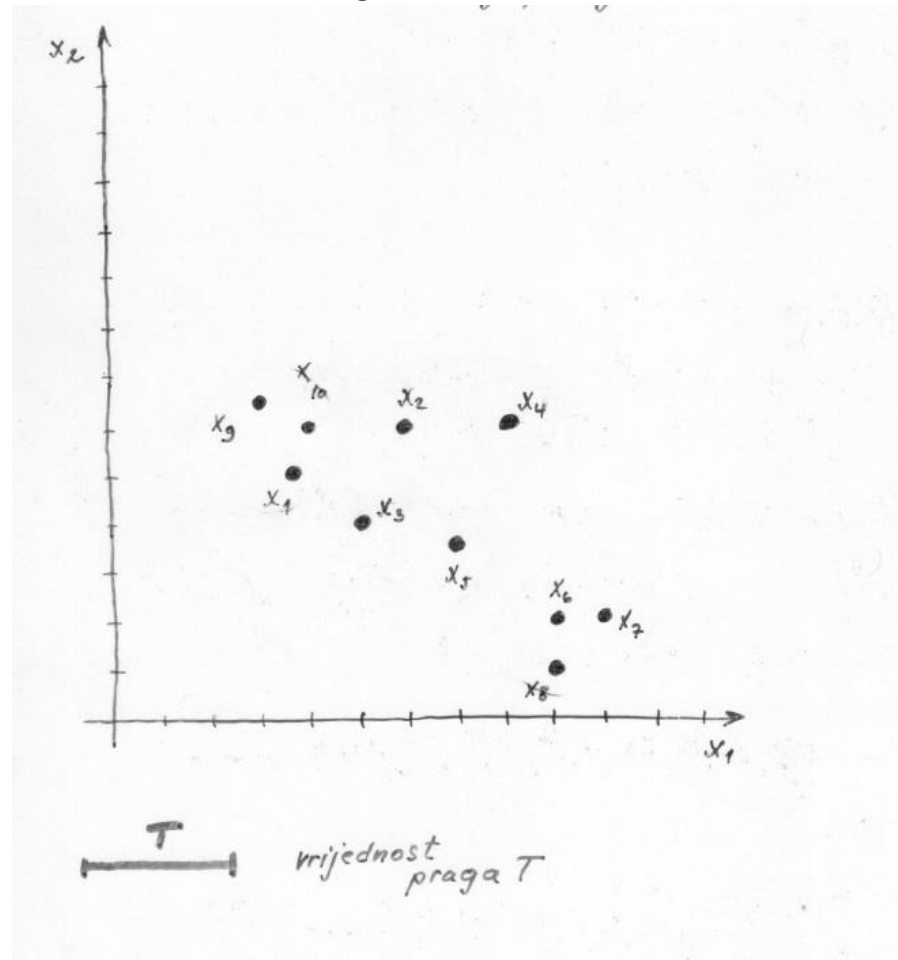
- jednostavnost – rezultat se dobiva jednim prolazom kroz skup neoznačenih uzoraka C

Nedostatak:

- rezultat zavisi od izbora vrijednosti praga T
- zavisi i od izbora prvog središta grupe
- zavisi i od redoslijeda uzimanja uzoraka iz C

Grupiranje (engl. Clustering)

Za vježbu grupirajte uzorke (x_8 neka je središte prve grupe)

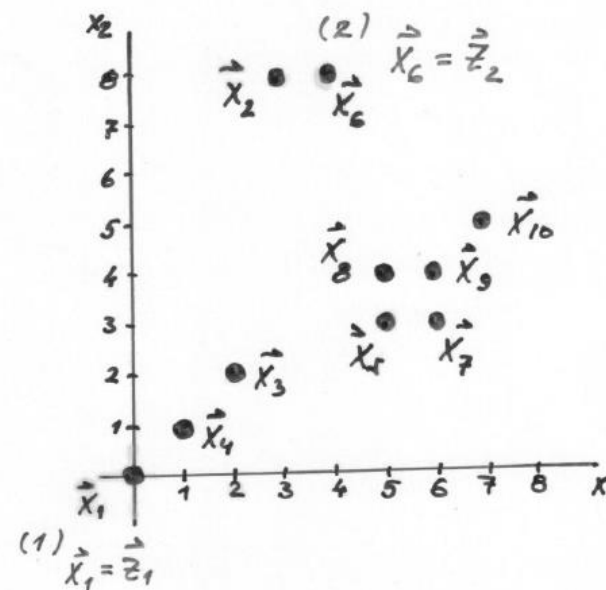


Grupiranje (engl. Clustering)

Heuristički algoritam grupiranja: **MAXMIN** (engl. Maximum-minimum distance clustering)

- sličan prethodnom algoritmu ali s tom razlikom da se prvo identificiraju područja grupa koja su najudaljenija
- temelji se na Euklidskoj udaljenosti

Algoritam MAXMIN ćemo „pratiti” na primjeru:



Grupiranje (engl. Clustering)

1. korak

Izaberemo $\mathbf{z}_1 = \mathbf{x}_1$ središte prve grupe

2. korak

Odredimo najudaljeniji uzorak od $\mathbf{z}_1 = \mathbf{x}_1$

(to je u našem slučaju \mathbf{x}_6)

i proglašavamo ga središtem \mathbf{z}_2

3. korak

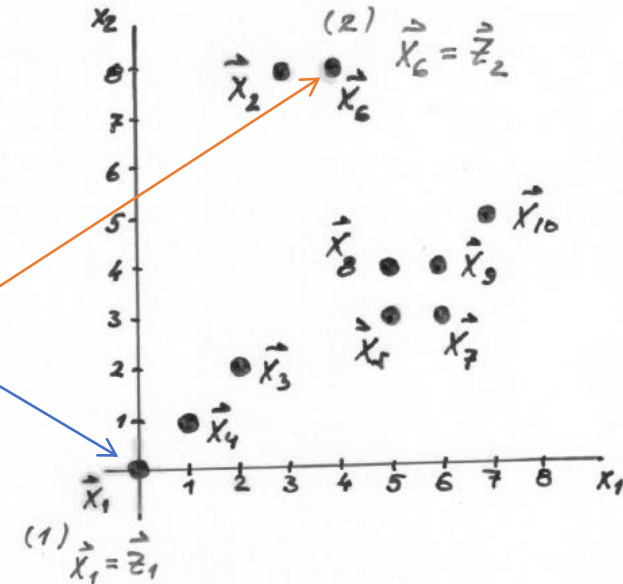
Računamo udaljenosti između preostalih uzoraka i uzoraka \mathbf{z}_1 i \mathbf{z}_2

označava prvo središte

$D_{1,2} D_{1,3} D_{1,4} D_{1,5} D_{1,7} D_{1,8} D_{1,9} D_{1,10}$

označava drugo središte

$D_{2,2} D_{2,3} D_{2,4} D_{2,5} D_{2,7} D_{2,8} D_{2,9} D_{2,10}$



Grupiranje (engl. Clustering)

- za svaki par izaberemo i pohranimo **minimalnu** vrijednost

označava prvo središte

$D_{1,2}$ $D_{1,3}$ $D_{1,4}$ $D_{1,5}$ $D_{1,7}$ $D_{1,8}$ $D_{1,9}$ $D_{1,10}$
 $D_{2,2}$ $D_{2,3}$ $D_{2,4}$ $D_{2,5}$ $D_{2,7}$ $D_{2,8}$ $D_{2,9}$ $D_{2,10}$

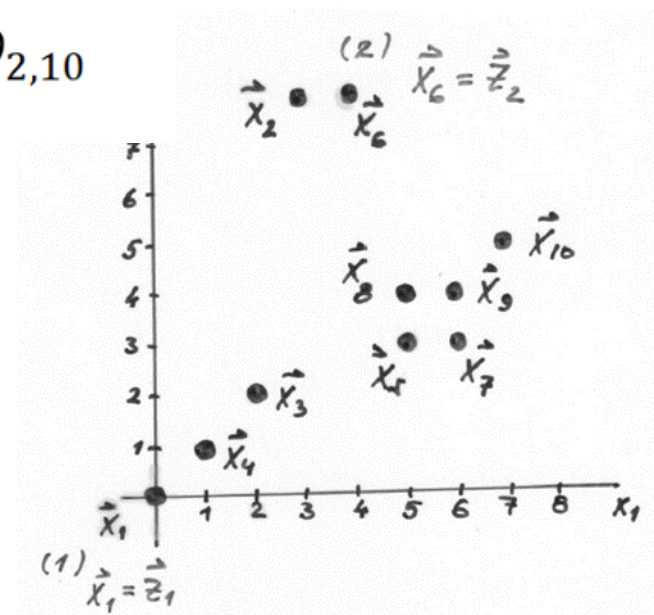
označava drugo središte

- to su u našem primjeru

$D_{1,2}$ $D_{1,3}$ $D_{1,4}$ $D_{1,5}$ $D_{1,7}$ $D_{1,8}$ $D_{1,9}$ $D_{1,10}$

$D_{2,2}$ $D_{2,3}$ $D_{2,4}$ $D_{2,5}$ $D_{2,7}$ $D_{2,8}$ $D_{2,9}$ $D_{2,10}$

$D_{2,2}$ $D_{1,3}$ $D_{1,4}$ $D_{2,5}$ $D_{2,7}$ $D_{2,8}$ $D_{2,9}$ $D_{2,10}$



Grupiranje (engl. Clustering)

4. korak

Izaberemo **MAKSIMUM** od tih minimalnih udaljenosti

- to je u našem primjeru $D_{2,7}$.

5. korak

Ako je ta udaljenost signifikantna u odnosu na udaljenost između z_1 i z_2

(npr. najmanje $\frac{1}{2}$ udaljenosti) uzorak koji odgovara toj udaljenosti proglašava se središtem **nove grupe – u drugim slučajevima algoritam završava**

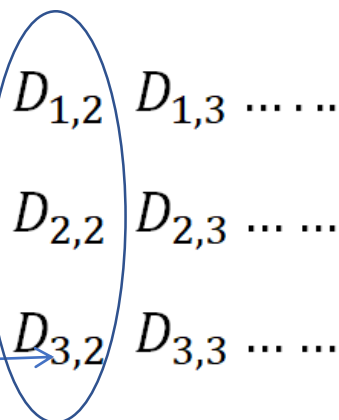
$$x_7 = z_2$$

Grupiranje (engl. Clustering)

6. korak

Sada računamo udaljenosti:

označava središte
treće grupe $z_3 = x_7$



The diagram shows a matrix of distances $D_{i,j}$ where i and j range from 1 to 3. The elements are arranged in three rows and three columns. A blue oval encircles the entire third column, which contains the elements $D_{1,2}$, $D_{2,2}$, and $D_{3,2}$. A blue arrow points from the text 'označava središte treće grupe $z_3 = x_7$ ' to the element $D_{3,2}$ within this oval.

$$\begin{matrix} D_{1,2} & D_{1,3} & \dots & \dots \\ D_{2,2} & D_{2,3} & \dots & \dots \\ D_{3,2} & D_{3,3} & \dots & \dots \end{matrix}$$

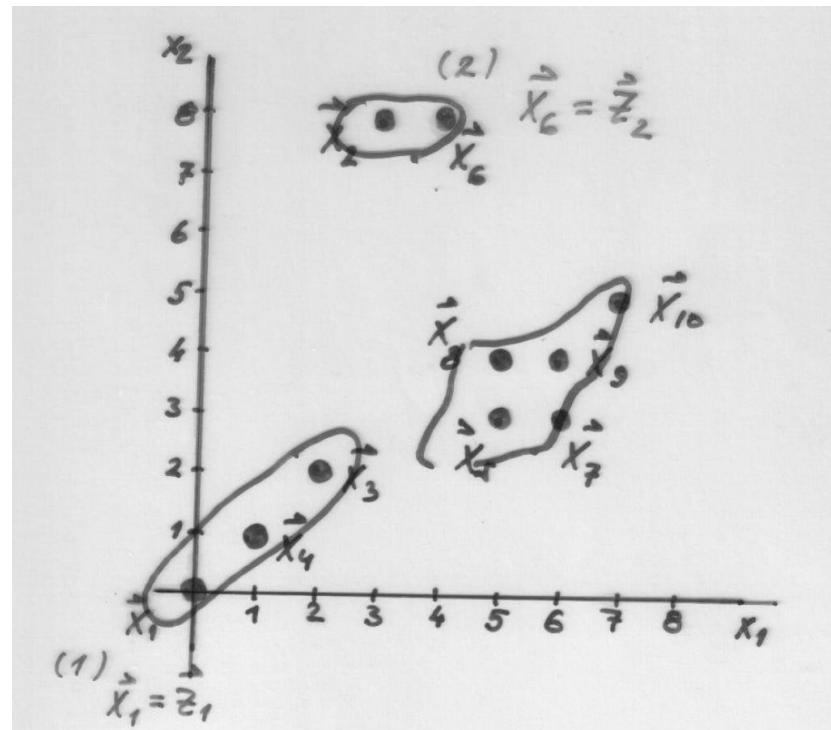
Postupak se ponavlja – traži se **minimum** trojki udaljenosti i pohranjuje se – bira se maksimum od ovih minimuma i uspoređuje se s $1/2$ udaljenosti z_1 i z_2

7. Korak

- ako maksimum od minimalnih udaljenosti nije signifikantan, preostali uzorci se dodjeljuju najbližim središtima grupa, inače postupak se ponavlja

Grupiranje (engl. Clustering)

Rezultat grupiranja **maxmin** algoritmom



Grupiranje (engl. Clustering)

Postupak grupiranja na temelju minimizacije kriterijske funkcije

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

- kriterijska funkcija

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

gdje je

$$J_j = \sum_{\mathbf{x} \in S_j} \|\mathbf{x} - \mathbf{z}_j\|^2$$

$$N_c = K$$

Grupiranje (engl. Clustering)

1. korak

Izaberimo $K \leq N$, gdje je N broj neoznačenih uzoraka

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

2. Korak

U k -tom koraku (iteraciji) razdijelimo uzorke

$$\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N$$

u K grupa pomoću relacije:

$$\mathbf{x} \in S_j(k) \text{ ako je } \|\mathbf{x} - \mathbf{z}_j(k)\| < \|\mathbf{x} - \mathbf{z}_i(k)\|$$

$$i = 1, 2, \dots, K \text{ i } i \neq j$$

$S_j(k)$ – označava skup uzoraka čiji je centar $\mathbf{z}_j(k)$

Grupiranje (engl. Clustering)

3. korak

Izračunamo nova središta grupa:

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

tako da je kriterijska funkcija

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

minimalna;

Središta grupa koja minimiziraju kriterijsku funkciju u k -toj iteraciji su aritmetičke srednje vrijednosti uzoraka pojedinih grupa:

Grupiranje (engl. Clustering)

Središta grupa koja minimiziraju kriterijsku funkciju u k -toj iteraciji su aritmetičke srednje vrijednosti uzoraka pojedinih grupa:

$$\mathbf{z}_j(k + 1) = \frac{1}{N_j} \sum_{\mathbf{x} \in S_j(k)} \mathbf{x}$$

N_j je broj uzoraka u grupi

4. korak

Ako je $\mathbf{z}_j(k + 1) = \mathbf{z}_j(k)$ za sve $j = 1, 2, \dots, K$ postupak završava, ako to nije

zadovoljeno ponavljamo postupak od 2. koraka

Grupiranje (engl. Clustering)

Na rezultat grupiranja K -srednjih vrijednosti utječe:

- broj grupa
- izbor početnih središta grupa
- redoslijed kojim se neoznačeni uzorci uzimaju
- geometrijska svojstva podataka

Još jedna značajka:

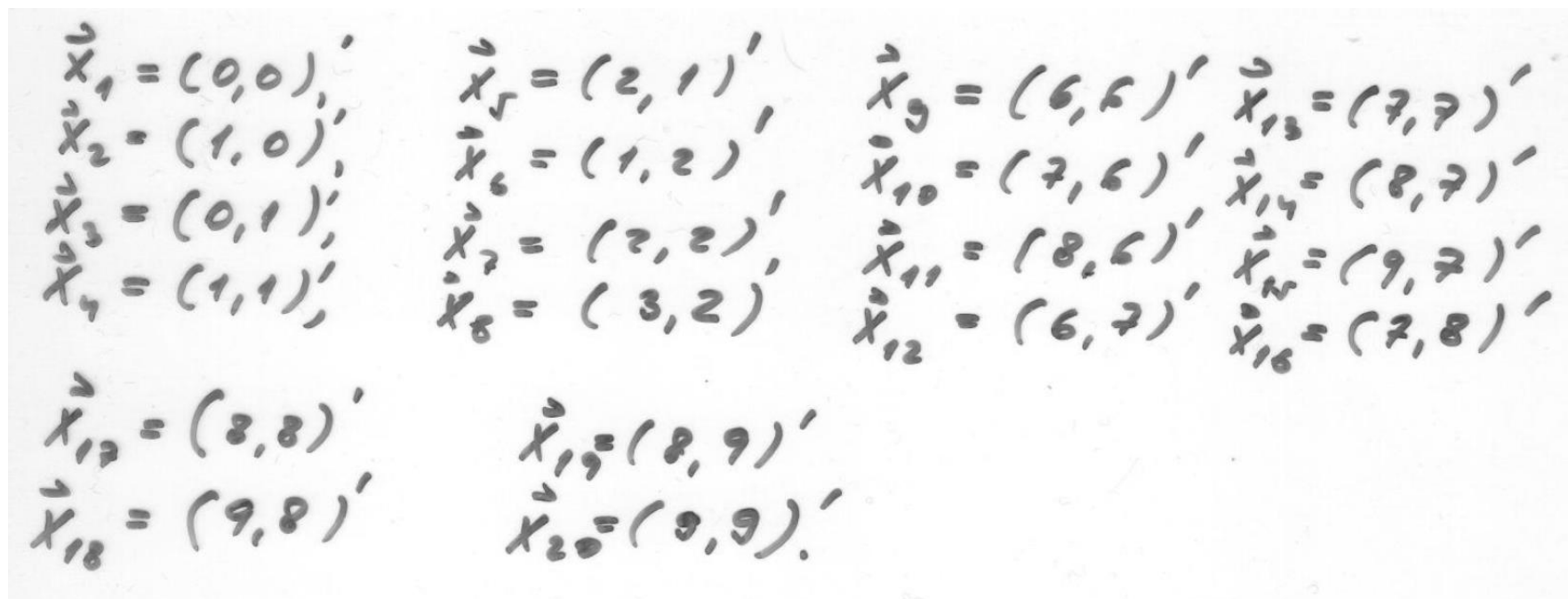
- nema općenitog dokaza o konvergenciji algoritma

Algoritam zahtijeva eksperimentiranje s različitim vrijednostima K i različitim početnim konfiguracijama

Grupiranje (engl. Clustering)

Primjer algoritma *K-srednjih* vrijednosti

Zadan je skup neoznačenih 2-D uzoraka



Handwritten list of 20 2D vectors \vec{x}_1 through \vec{x}_{20} :

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

Grupiranje (engl. Clustering)

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

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

$i=2$ imamo :

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

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

Grupiranje (engl. Clustering)

3. KORAK

Računamo nova središta grupa:

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

$$\vec{x}_1 = (0, 0);$$

$$\vec{x}_3 = (0, 1);$$

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

Grupiranje (engl. Clustering)

4. KORAK

Budući da je $\vec{z}_j(2) \neq \vec{z}_j(1)$; $j=1, 2$

vraćamo se na KORAK 2.

2. KORAK

$$\|\vec{x}_\ell - \vec{z}_1(2)\| < \|\vec{x}_\ell - \vec{z}_2(2)\| \quad \Leftrightarrow$$

$$\ell = 1, 2, \dots, 8 \quad ;$$

$$\|\vec{x}_\ell - \vec{z}_2(2)\| < \|\vec{x}_\ell - \vec{z}_1(2)\| \quad \Leftrightarrow$$

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

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

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

Grupiranje (engl. Clustering)

3'. KORAK

Obnovimo vrijednosti centara :

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

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

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

$$= \begin{pmatrix} 7.67 \\ 7.33 \end{pmatrix}$$

Grupiranje (engl. Clustering)

4. KORAK

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

2. KORAK ; 3. KORAK

Daje isti rezultat kao u prethodnoj
iteraciji : $\bar{z}_1(4) = \bar{z}_1(3)$
 $\bar{z}_2(4) = \bar{z}_2(3)$

Grupiranje (engl. Clustering)

4^o KORAK $\vec{z}_j(4) = \vec{z}_j(3)$ za $j=1, 2$
algoritam je konvergirao; kao sljedeće
centre grupa: $\vec{z}_1 = \begin{pmatrix} 1.25 \\ 1.13 \end{pmatrix}$; $\vec{z}_2 = \begin{pmatrix} 7.62 \\ 7.33 \end{pmatrix}$

Grupe za $K = 2$ su:

$$\{X_1, X_2, X_3, \dots, X_8\} \text{ i } \{X_9, X_{10}, X_{11}, \dots, X_{20}\}$$

Grupiranje (engl. Clustering)

iii) kombinacija pristupa i) i ii)

heuristički + optimizacija kriterijske funkcije

- ISODATA algoritam (Iterative-Self-Organizing Data A) /B. H. Ball, 1965/
(sličan algoritmu *K- srednjih vrijednosti* ali omogućava da se u fazi izvođenja algoritma mijenja broj grupa)

Grupiranje (engl. Clustering)

Pojednostavljeni ISODATA algoritam (za područje AI; Patterson, 1990)

- vrijednosti praga:

t_1 – raspršenje

t_2 – stapanje

t_3 – zanemarivanje

1. korak

Izaberi k uzoraka kao „sjeme” za inicijalna središta grupa. To se može napraviti tako da se uzme prvih k neoznačenih uzoraka (ili slučajnim izborom k neoznačenih uzoraka ili uzimanjem k neoznačenih uzoraka koji su međusobno udaljeni nekom izabranom udaljenosti d).

Grupiranje (engl. Clustering)

2. korak

Grupirajte preostale uzorke na temelju najmanje udaljenosti do pojedinih središta.

3. korak

Nakon što su svi uzorci grupirani, izračunajte nova središta za svaku grupu.

(Središte može biti definirano kao srednja vrijednost svih vektora koji pripadaju pojedinoj grupi.)

4. korak

Ako je raspršenje (izraženo pomoću udaljenosti pojedinih uzoraka u grupi od središta) takvo da prelazi prag t_1 , onda rasprši tu grupu na dvije grupe i izračunaj središta novih grupa.

Grupiranje (engl. Clustering)

5. korak

Ako je udaljenost između dva središta manja od t_2 stopite te dvije grupe u jednu i izračunaj novo središte.

6. korak

Ako grupa ima manje od t_3 članova (neoznačena uzorka), zanemarite tu grupu.

U daljnjem postupku je se „grupa” ignorira.

7. korak

Ponavljajte korake od 3 do 6 sve dok se dogodi da nema više promjene središta ili dok nije dostignuta granica pretpisanih iteracijskih koraka.

Grupiranje (engl. Clustering)

Grupiranje na temelju teorije grafova

- postupci grupiranja na temelju teorije grafova imaju neka zajednička svojstva:
 - svakom se uzorku iz skupa neoznačenih uzoraka dodjeljuje vrh (čvor) **povezanog grafa** G
 - grupe se određuju traženjem **komponenata povezanosti** grafa G

Def. 1.

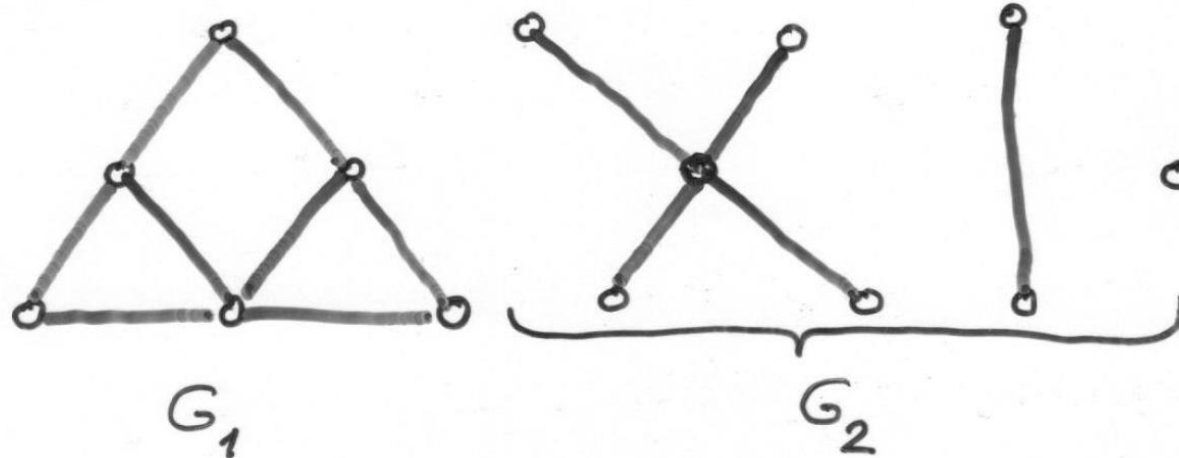
Neusmjereni /neorijentirani/ graf je **povezan** ako se njegova dva proizvoljna vrha mogu povezati putem koji čine bridovi (lukovi) grafa. /ako postoje vrhovi koji se ne mogu povezati putem – graf je **nepovezan**/

Grupiranje (engl. Clustering)

Def. 2.

Nepovezani graf se sastoji od dva ili više odvojenih dijelova. Odvojeni dijelovi grafa nazivaju se **komponente povezanosti grafa**.

Primjer



Grupiranje (engl. Clustering)

Većina postupaka grupiranja na temelju teorije grafova koristi **matricu sličnosti** koja se generira pomoću udaljenosti između neoznačenih uzoraka:

$$D_{kl} = \|\mathbf{x}_k - \mathbf{x}_l\| \text{ za } k = 1, 2, \dots, N \text{ i } l = 1, 2, \dots, N$$

-Umjesto Euklidske udaljenosti može se koristiti i neka druga mjera udaljenosti

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

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

Grupiranje (engl. Clustering)

Matrica sličnosti definira graf sličnosti u kojem vrhovi odgovaraju uzorcima iz skupa neoznačenih uzoraka a bridovi grafa povezuju vrhove i i j samo ako je $S_{ij} = 1$

Jednostavan algoritam grupiranja (R.O. Duda et al. , 2001) (Single-linkage Algorithm)

Dva uzorka x_i i x_j u istoj grupi ako i samo ako postoji niz

$$x_1, x_2, \dots, x_k$$

za koji vrijedi da je x_i sličan x_1 , x_1 sličan x_2 , ... , itd. ... x_{j-1} sličan x_j - tako da grupa odgovara komponentama povezanosti grafa sličnosti (engl. Single-linkage algorithm);

Grupiranje (engl. Clustering)

Grupiranje pomoću minimalnog razapinjućeg stabla grafa (minimal spanning tree)

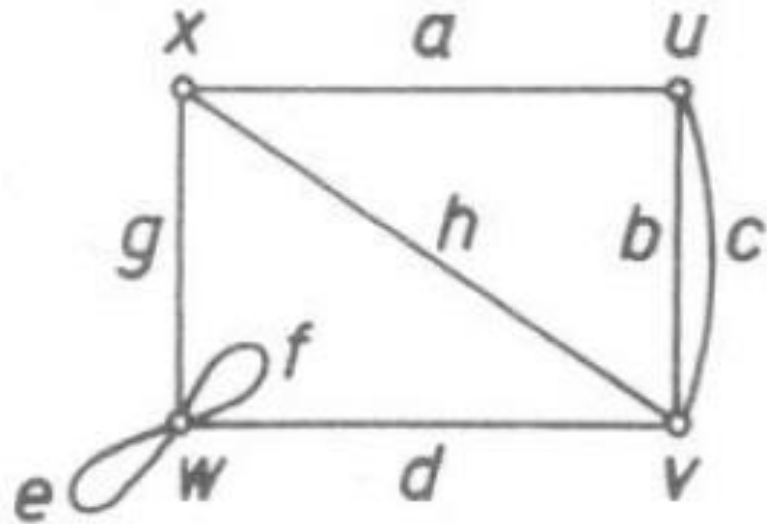
- Aciklički graf – je onaj graf koji ne sadrži cikluse
- Šetnja u grafu G je netrivialan konačni niz $W = v_0 e_1 v_1 e_2 \dots e_k v_k$ čiji su članovi naizmjenice vrhovi v_i i bridovi e_i , tako da su krajevi od e_i vrhovi v_{i-1} i v_i , za svako $1 \leq i \leq k$.

Vrhovi v_1 do v_{k-1} su unutarnji vrhovi

- Šetnja je **zatvorena** ako ima pozitivnu duljinu a početak i kraj se podudaraju.
- Zatvorena šetnja, kod koje su početak i unutrašnji vrhovi različiti, zove se **ciklus**

Grupiranje (engl. Clustering)

Primjer:



ciklus:

x a u b v h x

- Zatvorena šetnja, kod koje su početak i unutrašnji vrhovi različiti, zove se **ciklus**

Grupiranje (engl. Clustering)

Aciklički graf – je onaj graf koji ne sadrži cikluse;

- **Stablo** je povezani aciklički graf;
- Svaka dva vrha u stablu povezana su jednim jedinim putem;
- Razapinjuće stablo grafa G je razapinjući podgraf koji je stablo;

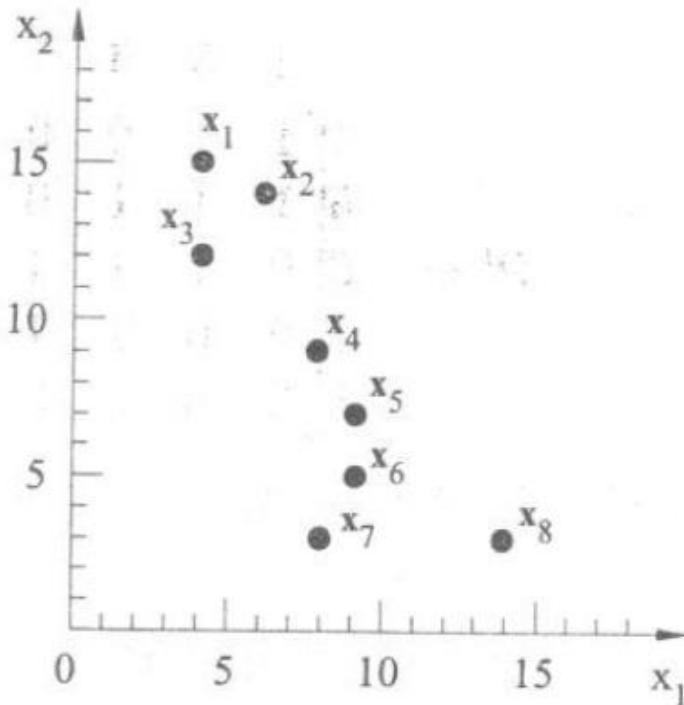
Grupiranje (engl. Clustering)

Algoritam

- Uzorke predočimo kao vrhove potpunog neusmjerenog grafa (svaki par vrha spojen je bridom);
- Svaki brid grafa „opteretimo” s udaljenosti između odgovarajućih vrhova
- Oblikujemo razapinjuće stablo grafa ali tako da je ukupna duljina svih njegovih **bridova najmanja – minimalno razapinjuće stablo grafa!**
- Pronaći suvišne grane u minimalno razapinjućem stablu i ukloniti ih iz stabla.
- Dijelovi (povezane komponente) tako „raspadnutog” razapinjućeg stabla određuju grupe.

Grupiranje (engl. Clustering)

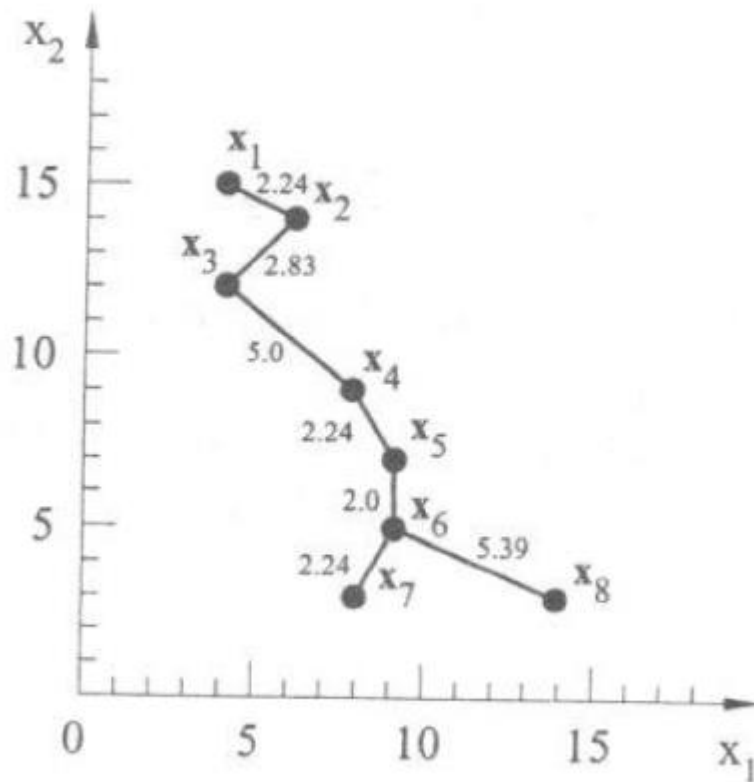
Primjer (N. Pavešić, 2000)



1	2	3	4	5	6	7	8	vzorec (roj)
0.00	2.24	3.00	7.21	9.43	11.18	12.65	15.62	1
	0.00	2.83	5.39	7.62	9.49	11.18	13.60	2
		0.00	5.00	7.07	8.60	9.85	13.45	3
			0.00	2.24	4.12	6.00	8.49	4
				0.00	2.00	4.12	6.40	5
					0.00	2.24	5.39	6
						0.00	6.00	7
							0.00	8

Grupiranje (engl. Clustering)

Minimalno razapinjuće stablo grafa (algoritam C. Zahn, 1971):



Grupiranje (engl. Clustering)

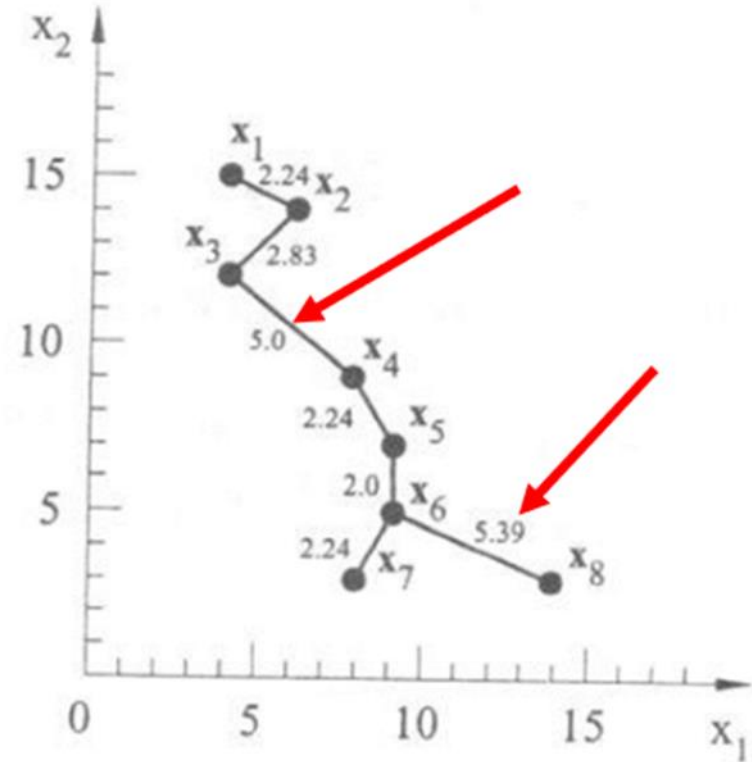
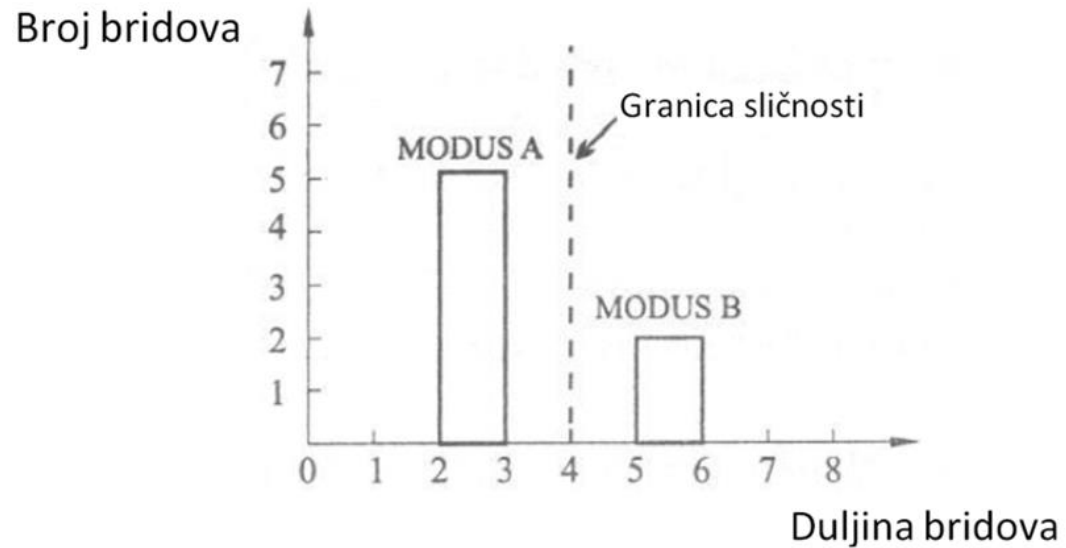
Kako otkriti suvišne grane?

- Odrediti distribuciju duljina (udaljenosti) u minimalnom razapinjućem stablu
- Histogram duljina – očekujemo da će histogram biti bimodalan (imati dva karakteristična vrha):
 - jedan vrh odgovara udaljenostima koje se javljaju unutar grupe
 - drugi vrh histograma odgovara udaljenostima između grupa

Prag koji određuje suvišne grane – granična duljina (udaljenost) nalazi se između ta dva vrha

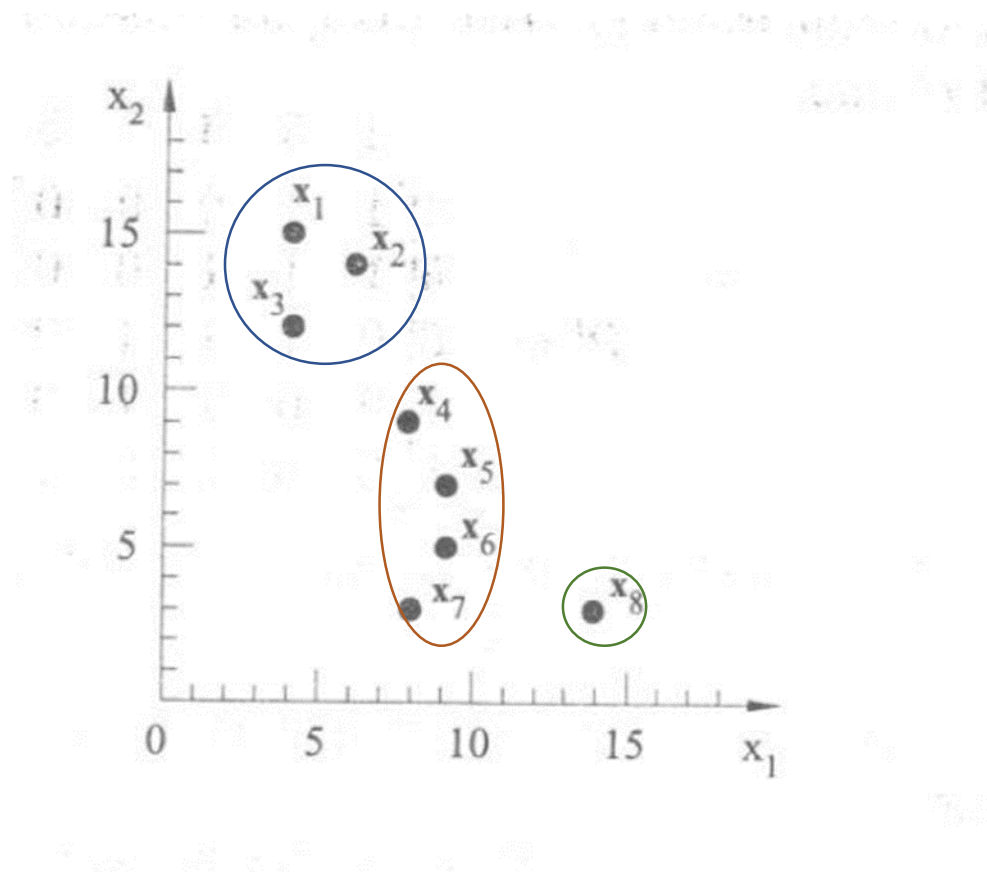
Grupiranje (engl. Clustering)

Primjer (nastavak):



Grupiranje (engl. Clustering)

Primjer (nastavak):



Grupiranje (engl. Clustering)

Hijerarhijski postupci traženja grupa

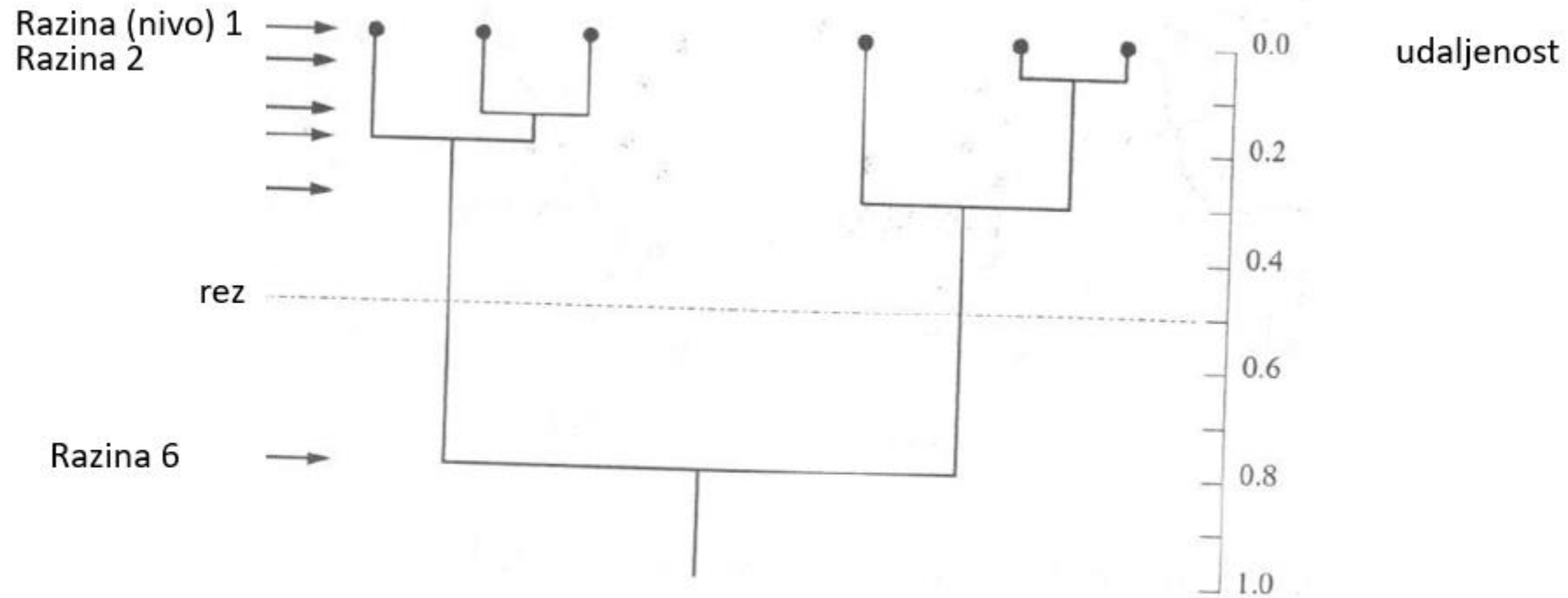
- ne dobivamo jedno rastavljanje skupa neoznačenih uzoraka već niz grupa počevši od grupe koja ima $N = N_c$ uzoraka u grupi (N_c je ukupan broj neoznačenih uzoraka), zatim $N - 1, N - 2, \dots, 1$ na temelju kriterija najboljeg rastavljanja na grupe.

Grupiranje se predstavlja razinama a u svakoj sljedećoj razini objedinjujemo najslićnije grupe u jednu.

Grupe koje su objedinjene na jednoj razini ne mogu se „razdružiti” na višim razinama. Slijed grupiranja možemo prikazati stablom koje se naziva **dendrogram**.

Grupiranje (engl. Clustering)

Primjer: dendrogram



Algoritam (Lance et al. 1967)

Grupiranje (engl. Clustering)

Hijerarhijski postupci grupiranja:

- objedinjavajući (engl. Agglomerative) – temelje na pristupu „s dna prema vrhu”
- razdvajajući (engl. Divisive) – temelje na pristupu „s vrha prema dnu”

(lit. S. Theodoridis, K. Koutroumbas, Pattern Recognition)