1 Grupiranje (engl. Clustering)

-do sada baratali "nadgledanom" klasifikacijom ili učili (klasifikator) s učiteljem
-Grupiranje → "neoznačeni uzorci" / ne znamo pripadnosti takvih uzoraka niti znamo broj razreda iz kojih uzorci dolaze /
Zadatak: Otkriti organizaciju uzoraka i grupirati ih u "smislene" ("prirodne") grupe koje će nam omogućiti otkrivanje sličnosti i različitosti između uzoraka i zime dopustiti izvođenje zaključka o njima.
Ovakva se zamisao obilato rabi u :

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

-Grupiranje:

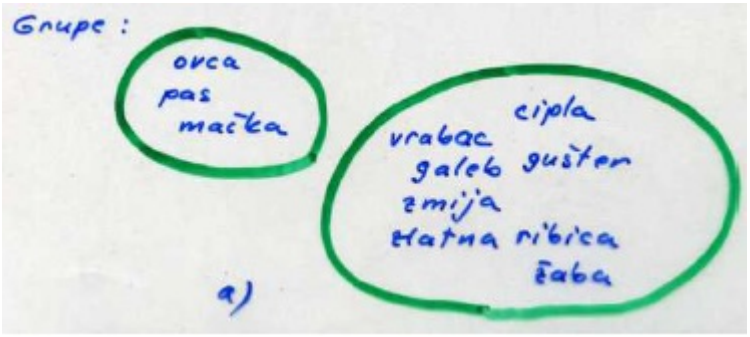
nenadgledano učenje (engl. unsupervised learning), učenje bez učitelja
PR

numerička taksonomija biologija i ekologija

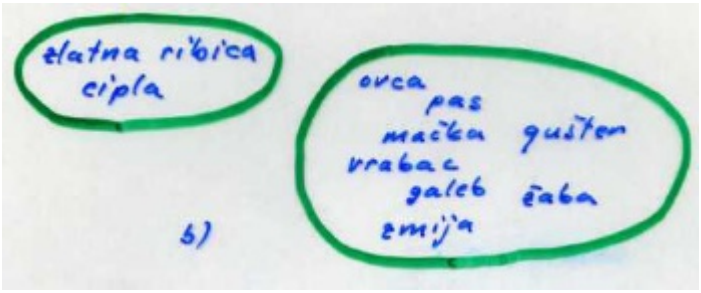
tipologija društvene znanosti

Primjer: Razmotrimo slijedeće životinje:
ovca, pas, mačka (sisavci)
vrabac, golub (ptice)
zmija, gušter (reptili)
zlatna ribica, cipla (ribe)
žaba (vodozemac)
Organizirajmo ih u grupe! -kriterij grupiranja?

a) Npr. da li ženke nose svoju (buduću) mladunčad?
b) da li imaju pluća?
c) okoliš u kojem žive?



Slika 1: Prikaz grupe a)



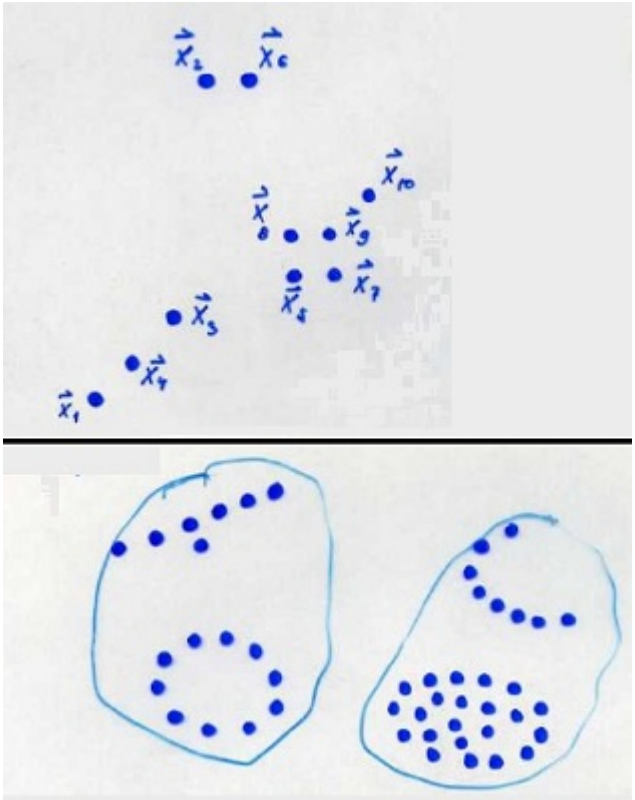
Slika 2: Prikaz grupe b)

Slika 3: Prikaz grupe c)

Osnovni koraci u postupku grupiranja:

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

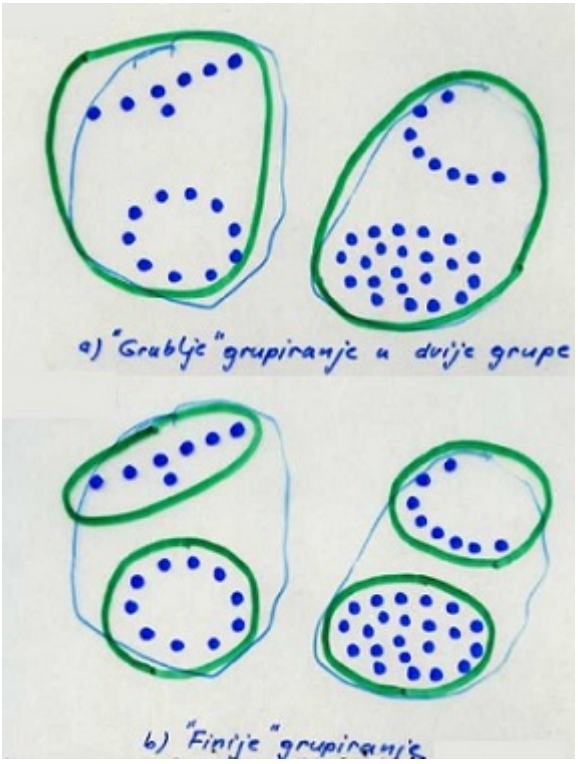
U brojnim slučajevima rabi se još dodatni korak:
težnja grupiranju - podrazumijeva različite testove koji pokazuju da li raspoloživi podaci imaju strukturu grupe (npr. skup podataka može biti potpuno slučajne prirode te pokušaj ostvarivanja "smislenih" grupa je besmislen)
Svi osnovni koraci su podložni subjektivnosti eksperta! (Subjectivity is a reality we have to live with from now on)



Slika 4: Primjeri uzoraka

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 → poticaj za postavljanje hipoteza
- Ispitivanje hipoteza
- Predviđanje na temelju grupa
  - Npr. analiza grupa je primjenjena na skupu podataka o pacijentima koji su oboljeli od iste bolesti
    rezultat - broj grupa pacijenata prema njihovoh reakciji na određene lijekove
  - novi pacijent - za njega identificiramo odgovarajuću grupu



Slika 5: Primjer grubljeg i finijeg grupiranja

1.1 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 (dichotomons)
Klasifikacija značajki na temelju relativnog značenja vrijednsoti 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°C; London 5°C
Smisleno je reći da je temperatura u Parizu za 5°C više od one u Londodnu
Besmisleno (ili skoro besmisleno) je reći da je Paris dvostruko topliji od Londona

(d) Omjer između značajki ima smisla
Npr. osoba koja ima 120kg je dvaput teža (i deblja) od osobe koja ima 60kg

1.2 Definicije grupiranja

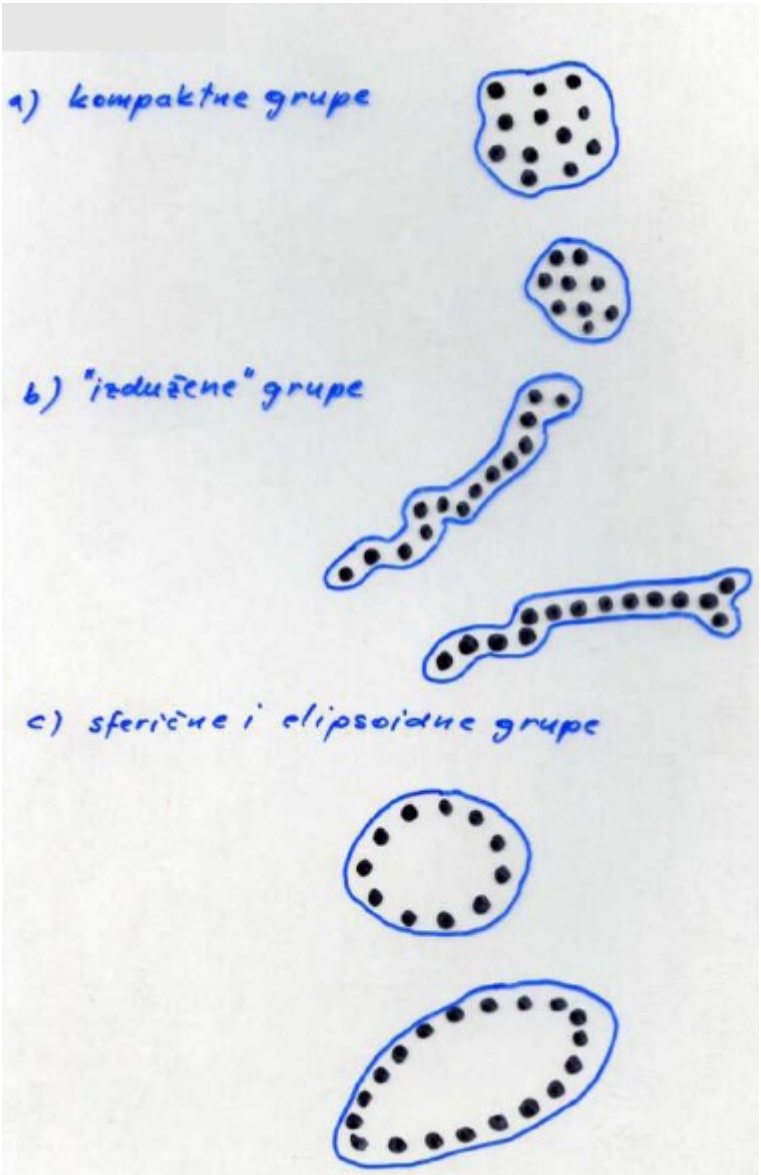
Definicije se temelje na "labavo" definiranim izrazima kao što su "slični", "različiti" (odnose se na uzorke u pojedinim razredima)
Everitt, 1981:
Vektori → točke u l-dimenzionalnom prostoru
Grupe → kontinuirana područja prostora koja imaju relativno visoku gustoću točaka i odvojena su od drugih kontinuiranih prostora visokih gustoća s područjima 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 $X = \left\{\vec{X_1}, \vec{X_2}, \ldots, \vec{X_N}\right\}$
m - grupiranje X-a odgovara dijeljenju X u m skupova (grupa) $C_1, C_2, \ldots, C_m$ (C-cluster) tako da su zadovoljena sljedeća tri uvjeta:

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

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

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

Važno:
Vektori sadržani u grupi C_i su "sličniji" jedan drugome i "manje sličniji" vektorima iz drugih grupa.
Kvantifikacija izraza "sličan" i "različit" zavisi od tipa grupe.



Slika 6: Tipovi grupa

U definiciji grupa zahtijevamo da svaki vektor pripada samo jednoj grupi → takva vrsta grupiranja naziva se Crisp grupiranje(izrazito, jasno)
Alternativni pristup grupiranju:neizrazito grupiranje(fuzzy clustering)
Neizraziro grupiranje skupa podataka X u m grupa određeno je s m funkcija u_j:
$u_j : X \to [0,1]$, $j = 1, 2, \ldots, m$ i
$\sum_{j=1}^{m} u_j(\vec{X_j}) = 1 \quad i = 1, 2, \ldots, N$
$0 < \sum_{i=1}^{N} u_j(\vec{X_i}) < N \quad j = 1, 2, \ldots, N$
u_j, $j = 1, 2, \ldots, m \to$ funkcije pripadnosti (engl. membership functions)
U slučaju neizrazitog grupiranja svaki vektor $\vec{X_i}$ pripada više od jednoj grupi istodobno s nekim stupnjem ili mjerom pripadnosti (iz intervala [0,1]).

1.3 Mjere bliskosti (sličnosti)(engl. proximity measures)

- Mjera različitosti (dissimilarity measure DM)
  DM je funkcija d od X, $d : X \to R$, gdje je R skup realnih brojeva takav da:
  $\exists d_0 \in R : -\infty < d_0 \leq d(\vec{X}, \vec{Y}) < +\infty$, $\forall \vec{X}, \vec{Y} \in X$ i
  $d(\vec{X}, \vec{Y}) = d(\vec{Y}, \vec{X})$, $\forall \vec{X}, \vec{Y} \in X$ i
  $d(\vec{X}, \vec{Y}) = d_0$ iff $\vec{X} = \vec{Y}$ i još dodatni uvjet:
  $d(\vec{X}, \vec{Z}) \leq$d(,)+$d(\vec{Y}, \vec{Z})$, $\forall \vec{X}, \vec{Y}, \vec{Z} \in X$
  d se naziva metrika DM
  d_0 - minimalna moguća različitost između bilo koja 2 vektora iz X (dobiva se kada su oni identični!)

- Mjera sličnosti (similarity measure SM)
  SM je funkcija s:
  $s : X \times X \to R$ tako da je:
  $\exists s_0 \in R : -\infty < s(\vec{X}, \vec{Y}) \leq s_0 < +\infty$, $\forall \vec{X}, \vec{Y} \in X$ i
  $s(\vec{X}, \vec{Y}) = s(\vec{Y}, \vec{X})$, $\forall \vec{X}, \vec{Y} \in X$ i
  $s(\vec{X}, \vec{X}) = s_0$, $\forall \vec{X} \in X$
  $s(\vec{X}, \vec{Y}) = s_0$ iff $\vec{X} = \vec{Y}$ i
  $s(\vec{X}, \vec{Y}) s(\vec{Y}, \vec{Z}) \leq [s(\vec{X}, \vec{Y}) + s(\vec{Y}, \vec{Z})] \cdot s(\vec{X}, \vec{Z})$, $\forall \vec{X}, \vec{Y}, \vec{Z} \in X$
  s se naziva metrika SM

Primjer:
Euklidska udaljenost $d_L(\vec{X}, \vec{Y})$
$d_L(\vec{X}, \vec{Y}) = \sqrt{\sum_{i=1}^{l} x_i - y_i{}^2}$, gdje je $\vec{X}, \vec{Y} \in X$
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: svi algoritmi grupiranja temelje se na mjeri bliskosti između vektora.
Neki (hijerarhijski algoritmi grupiranja) računaju udaljenosti IZMEĐU PAROVA SKUPOVA vektora iz X.
Bliskost između podskupova od X:
Neka je U skup koji sadrži podskupove od X.
D_i, $i = 1, 2, \ldots, k$ i
$U = \{D_1, D_2, \ldots, D_k\}$
Mjera bliskosti γ definirana nad U je:
$\gamma : U \times U \to R$, gdje je R skup realnih brojeva takav da
$\exists d_0 \in R : -\infty < d_0 \leq d(D_i, D_k) < +\infty$. $\forall D_i, D_k \in U$ i
$d(D_i, D_i) = d_0$, $\forall D_i \in U$ i
$d(D_i, D_k) = d(D_k, D_i)$, $\forall D_i, D_k \in U$ i
$d(D_i, D_k) = d_0$ iff $D_k = D_i$ i
$d(D_i, D_l) \leq d(D_i, D_k) + d(D_k, D_l)$, $\forall D_i, D_k, D_l \in U$
Primjer:
Neka je $X = \left\{\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_6\right\}$ i $U = \left\{\left\{\vec{X}_1, \vec{X}_2\right\}, \left\{\vec{X}_1, \vec{X}_4\right\}, \left\{\vec{X}_3, \vec{X}_4, \vec{X}_5\right\}, \left\{\vec{X}_1, \vec{X}_2, \vec{X}_3, \vec{X}_4, \vec{X}_5\right\}\right\}$
Definirajmo funkciju različitosti
$d^{ss}_{min}(D_i, D_j) = \min_{\vec{X} \in D_i, \vec{Y} \in D_j} d_L(\vec{X}, \vec{Y})$, gdje je d_L Euklidska udaljenost između dvaju vektora i $D_i, D_j \in U$.
Međutim d^{ss}_{min} nije metrika $D_1 = \left\{\vec{X}_1, \vec{X}_2\right\}$ i $D_2 = \left\{\vec{X}_1, \vec{X}_4\right\}$
-oni su različiti $D_1 \neq D_2$
$d^{ss}_{min}(D_1, D_2) = 0$ / zato što oba podskupa sadrže $\vec{X}_1$ /
Ne drži: $d(D_1, D_k) = d_0$ iff $D_k = D_i$

1.4 Mjere bliskosti između dviju točaka

A MJERE RAZLIČITOSTI

1) Utežnosna L_p metrika DM
   $d_p(\vec{X}, \vec{Y}) = (\sum_{i=1}^{l} w_i |x_i - y_i|^p)^{1/p}$, gdje su x_i i y_i i-te koordinate od $\vec{X}$ i $\vec{Y}$.
   $w_i \geq 0$ i-ti težnosni koeficijent
   Ako je $w_i = 1$, $i = 1, 2, \ldots, l$
   l_p metrika DM (unweighted metric)
   Euklidska udaljenost p=2; $w_i = 1$
   $d_2(\vec{X}, \vec{Y}) = (\sum_{i=1}^{l} w_i |x_i - y_i|^2)^{1/2}$
   Utežnosna l_2 metrika DM može se dalje generalizirati kao:
   $d(\vec{X}, \vec{Y}) = \sqrt{(\vec{X} - \vec{Y})^T B (\vec{X} - \vec{Y})}$, gdje je B pozitivno definitna matrica
   $l \times l$ realna matrica A se naziva pozitivno definitna ako za svaki vektor različit od 0 $\vec{X}$ vrijedi: $\vec{X}^T A \vec{X} > 0$

2) Mahalanobisova udaljenost - poseban slučaj metrike DM
$D = (\vec{X} - \vec{m})^T C^{-1} (\vec{X} - \vec{m})$
C - kovarijantna matrica
$\vec{m}$ - srednja vrijednost

3) Poseban slučaj L_p metrike je i (utežnosna) Manhattan norma
p=1
$d_1(\vec{X}, \vec{Y}) = \sum_{i=1}^{l} w_i |x_i - y_i|$

4) Utežnosna l_∞ norma
$d_\infty(\vec{X}, \vec{Y}) =_{1 \leq i \leq l} w_i |x_i - y_i|$

Primjer:
$\vec{X} = [0, 1, 2]^T$, $\vec{Y} = [3, 4, 2]^T$
$w_i = 1$ za $i = 1, 2, 3$
$d_1(\vec{X}, \vec{Y}) = \sum_{i=1}^{3} |x_i - y_i| = 6$
$d_2(\vec{X}, \vec{Y}) = (\sum_{i=1}^{3} |x_i - y_i|^2)^{1/2} = 2\sqrt{5}$
$d_\infty(\vec{X}, \vec{Y}) = 4$
Vrijedi $d_\infty(\vec{X}, \vec{Y}) < d_2(\vec{X}, \vec{Y}) < d_1(\vec{X}, \vec{Y})$

B MJERE SLIČNOSTI

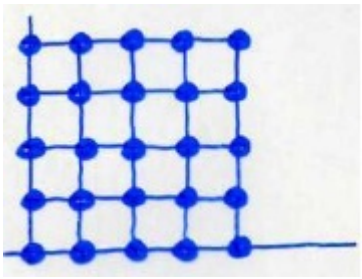
1) Skalarni produkt
$S_{inner}(\vec{X}, \vec{Y}) = \vec{X}^T \vec{Y} = \sum_{i=1}^{l} x_i y_i$
U većini slučajeva skalarni produkt se koristi kada su vektori $\vec{X}$ i $\vec{Y}$ normalizirani

2) Tanimoto mjera (SM) (Tanimoto distance)
/ za vektore realne i vektore s diskretnim vrijednostima /
$S_T(\vec{X}, \vec{Y}) = \frac{\vec{X}^T \vec{Y}}{\|\vec{X}\|^2 + \|\vec{Y}\|^2 - \vec{X}^T \vec{Y}}$ Zbrajanjem i oduzimanjem iznosa $\vec{X}^T \vec{Y}$ u nazivniku i malim preuređenjem dobivamo:
$S_T(\vec{X}, \vec{Y}) = \frac{1}{1 + \frac{(\vec{X} - \vec{Y})^T (\vec{X} - \vec{Y})}{\vec{X}^T \vec{Y}}}$
$\vec{X}^T \vec{Y} \rightarrow$ korelacija
$(\vec{X} - \vec{Y})^T (\vec{X} - \vec{Y}) \rightarrow$ Euklidska udaljenost^2

3) Fu-ova mjera sličnosti
$S_c(\vec{X}, \vec{Y}) = 1 - \frac{d_2(\vec{X}, \vec{Y})}{\|\vec{X}\| + \|\vec{Y}\|}$
$S_c(\vec{X}, \vec{Y})$ ima maksimum (=1) kada je $\vec{X} = \vec{Y}$ i minimum (=0) $\vec{X} = -\vec{Y}$

1.5 Vektori s diskretnim vrijednostima

Vektor $\vec{X}$ s koordinatama iz konačnog skupa $F = \{0, 1, \ldots, k-1\}$, k pozitivan cijeli broj.
Imamo k^l vektora $\vec{X} \in F^l$
vektori čvorovi u l-dimenzionalnoj rešetki



Slika 7: Primjer rešetke za l=2 i k=5

Razmotrimo $\vec{X}, \vec{Y} \in F^l$ i neka je $A(\vec{X}, \vec{Y}) = [a_(ij)]$, $i, j = 0, 1, \ldots, k-1$
$k \times k$ matrica u kojoj element a_{ij} predstavlja broj mjesta u kojem prvi vektor i simbol te odgovarajući element drugog vektora ima j simbol, $i, j \in F$.
A - matrica slučaja
Primjer:
l=6, k=3
$F = \{0, 1, 2\}$
$\vec{X} = [0, 1, 2, 1, 2, 1]^T$
$\vec{Y} = [1, 0, 2, 1, 0, 1]^T$

$$A = (\vec{X}, \vec{Y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}_{3\times 3}$$

Vrijedi $\sum_{i=0}^{k-1} \sum_{j=0}^{j-1} a_{ij} = l$

1.6 Mjere različitosti(za diskretne vektore

Hammingova udaljenost
Definirana je kao broj mjesta u kojima se dva vektora razlkuju
$d_H(\vec{X}, \vec{Y}) = \sum_{i=0}^{k-1} \sum_{j=0}^{j-1} a_{ij}$, $j \neq i$
Suma svih izvan dijagonalnih elemenata u matrici A
Specijalan slučaj k=2, vektori $\vec{X} \in F^l$ su binarni vektori, Hammingova udaljenost je
$d_H(\vec{X}, \vec{Y}) = \sum_{i=1}^{l} (x_i + y_i - 2x_i y_i) = \sum_{i=1}^{l} (x_i - y_i)^2$
U slučaju $F = \{-1, +1\}$ riječ je o bipolarnom vektoru
Hammingova udaljenost $d_H(\vec{X}, \vec{Y}) = 0.5(l - \sum_{i=1}^{l} x_i y_i)$

Tanimoto mjera između dvaju skupova X i Y

$\frac{n_{X\cap Y}}{n_X+n_Y-n_{X\cap Y}} = \frac{n_{X\cap Y}}{n_{X\cup Y}}$

$n_X, n_Y, n_{X\cap Y}, n_{X\cup Y}$ su kardinalni brojevi (broj elemenata) od $X, Y, X\cap Y, X\cup Y$.

1.7 Točkasti prikaz grupa

i) srednji vektor (srednja točka)
$\vec{m} = \frac{1}{n_c} \sum_{\vec{Y}\in C} \vec{Y}$
n_c kardinalni broj skupa C

ii) srednji centar $\vec{m_c} \in C$
$\sum_{\vec{Y}\in C} d(\vec{m_c}, \vec{Y}) \leq \sum_{\vec{Y}\in C} d(\vec{Z}, \vec{Y}), \forall \vec{Z} \in C$
d - mjera različitosti između dviju točaka

iii) median centar $\vec{m_{med}} \in C$
$med(d(\vec{m_{med}}, \vec{Y}) | \vec{Y} \in C) \leq med(d(\vec{Z}, \vec{Y}) | \vec{Y} \in C), \forall \vec{Z} \in C$
med (T), gdje je T skup od q skalara je minimalni broj u T koji je veći ili jednak točno [(q+1)/2] brojevima u T.
Algoritam: izlistati elemente iz T u rastućem redu i izabrati [(q+1)/2] element u listi.

1.8 KRITERIJI GRUPIRANJA

Postupci:

i) Heuristički - vođeni intuicijom i iskustvom

ii) Oni koji se temelje na minimizaciji (ili maksimizaciji) neke kriterijske funkcije ili (performance-index) indexa
Najčešći kriterij $J = \sum_{j=1}^{N_c} \sum_{\vec{X}\in S_j} \left\| \vec{X} - \vec{m_j} \right\|^2$
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 uzoraka
$C = \left\{\vec{X_1}, \vec{X_2}, \ldots, \vec{X_N}\right\}$

1) Korak:
   Izaberi nenegativan broj T (prag)

2) Korak:
   Izaberi bilo koji uzorak iz $C = \left\{\vec{X_1}, \vec{X_2}, \ldots, \vec{X_N}\right\}$ i proglasi ga središtem grupe $\vec{Z_1}$
   Pretpostavimo da smo izabrali $\vec{X_1} = \vec{Z_1}$

3) Korak:
   Izračunamo 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 najbliži

Postupak se provodi dok se ne obradi svih N uzoraka
Prednost algoritma:

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

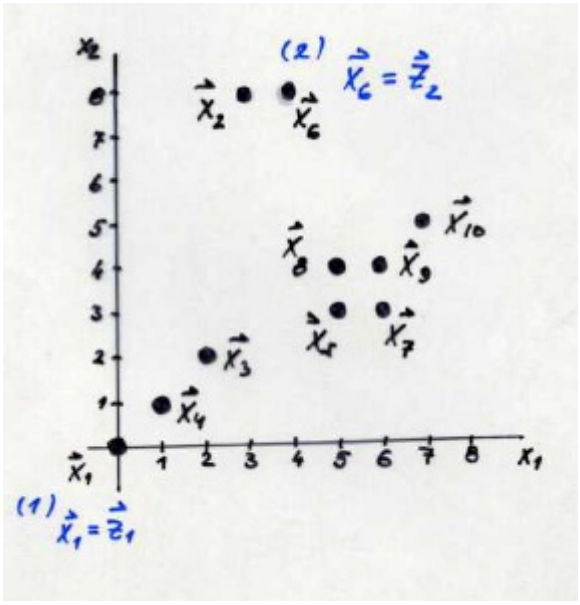
Nedostatak:

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

Heuristički algoritam grupiranja MAXMIN udaljenosti
(engl. maximum-minimum distance)

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

Primjer:



Slika 8: 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 proglasimo ga središtem $\vec{Z_2}$!

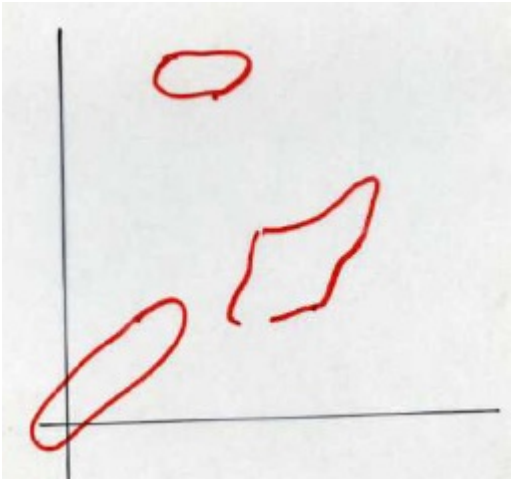
3. korak $\vec{Z_1} = \vec{X_1} : \vec{Z_2} = \vec{X_6}$
Izraččunamo udaljenosti između preostalih uzoraka i uzoraka$\vec{Z_1}$ i $\vec{Z_2}$
$(D_{12}, D_{13}, D_{14}, D_{15}, D_{17}, D_{18}, D_{19}, D_{110})$
$(D_{22}, D_{23}, D_{24}, D_{25}, D_{27}, D_{28}, D_{29}, D_{210})$
D_{ij} - i označava od kojeg središta, j do kojeg uzorka
Za svaki par izaberemo i pohranimo MINIMALNU UDALJENOST:
$D_{22}, D_{13}, D_{14}, D_{25}, D_{27}, D_{28}, D_{29}, D_{210}$

4. korak Izaberemo MAKSIMUM od tih minimalnih vrijednosti! (D_{27})

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 udaljenosti proglašava se središtem NOVE GRUPE $\vec{Z_3}$, u drugim slučajevima algoritam završava. $\vec{X_7} = \vec{Z_3}$

6. korak Izračunavamo udaljenosti uzoraka od $\vec{Z_1}$, $\vec{Z_2}$ i $\vec{Z_3}$
$(D_{12}, D_{13}, \ldots$
$(D_{22}, D_{23}, \ldots$
$(D_{32}, D_{33}, \ldots$
Postupak se ponavlja - traži se minimum trojki udaljenosti, pohranjuje se
- bira se maksimum i uspoređuje sa 1/2 udaljenosti $\vec{Z_1}$, $\vec{Z_2}$

7. korak Preostali uzorci se dodjeljuju majbližim središtima grupa!



Slika 9: Rezultat grupiranja MAXMIN

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

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

- kriterijska funkcija: $J = \sum_{j=1}^{N_c} J_j$, gdje je:

$J_j = \sum_{\vec{X} \in S_j} \left\| \vec{X} - \vec{Z_j} \right\|^2$

N_c - broj grupa K

1. korak Izaberimo $K \leq N$ središta grupa
$\vec{Z_1}(1), \vec{Z_2}(1), \ldots, \vec{Z_K}(1)$ N - je broj uzoraka

2. korak U k-tom koraku (iteraciji) razdijelimo uzorke $\vec{X_1}, \vec{X_2}, \ldots, \vec{X_N}$ u K grupa pomoću relacije:
$\vec{X} \in S_j(k)$ ako je $\left\| \vec{X} - \vec{Z_j(k)} \right\| < \left\| \vec{X} - \vec{Z_i(k)} \right\|$. $i = 1, 2, \ldots, K$ i $i \neq j$
$S_j(k)$ - označava skup uzoraka čiji je centar $\vec{Z_j}(k)$

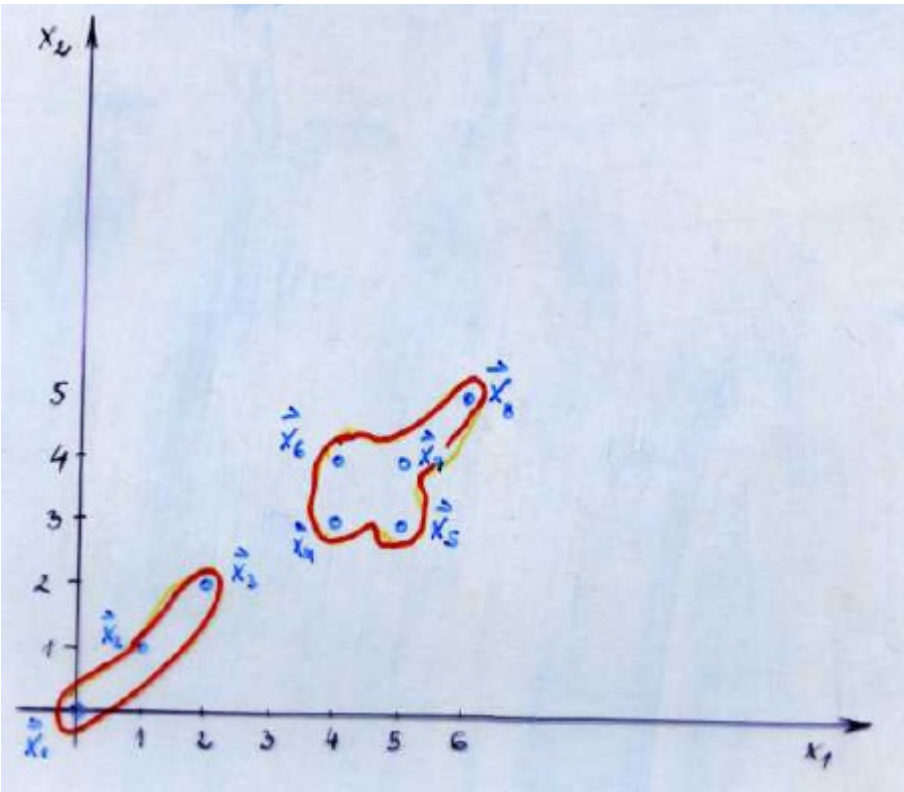
3. korak Izračunavamo nova središta grupa
$\vec{Z_j(k+1)}$, $j = 1, 2, \ldots, K$, tako da je kriterijska funkcija
$J = \sum_{j=1}^{K} \sum_{\vec{X} \in S_j(k)} \left\| \vec{X} - \vec{Z_j}(k+1) \right\|^2$, $j = 1, 2, \ldots, K$ minimalna
Središta grupa koja minimiziraju kriterijsku funkciju u k-toj iteraciji su ARITMETIČKE SREDNJE VRIJEDNOSTI UZORAKA POJEDINIH GRUPA
$\vec{Z_j(k+1)} = \frac{1}{N_j} \sum_{\vec{X} \in S_j(k)} \vec{X}$, za $j = 1, 2, \ldots, K$
N_j - broj uzoraka u grup

4. korak Ako je $\vec{Z_j(k+1)} = \vec{Z_j(k)}$ za sve $j = 1, 2, \ldots, K$ postupak završava.
Ukoliko nije, ponavljamo postupak od 2. koraka

Na rezultat grupiranja pomoću algoritma K-srednjih vrijednosti utječe:

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

Problem konvergencije?NEMA OPĆENITOG DOKAZA O KONVERGENCIJI ALGORITMA
Algoritam zahtjeva eksperimentiranje sa različirim vrijednostima K i različitim početnim konfiguracijama!
Primjer:
$\vec{X_1} = (0,0)'$, $\vec{X_2} = (1,0)'$, $\vec{X_3} = (0,1)'$, $\vec{X_4} = (1,1)'$, $\vec{X_5} = (2,1)'$, $\vec{X_6} = (1,2)'$, $\vec{X_7} = (2,2)'$, $\vec{X_8} = (3,2)'$, $\vec{X_9} = (6,5)'$, $\vec{X_{10}} = (7,6)'$, $\vec{X_{11}} = (8,6)'$, $\vec{X_{12}} = (6,7)'$, $\vec{X_{13}} = (7,7)'$, $\vec{X_{14}} = (8,7)'$, $\vec{X_{15}} = (9,7)'$, $\vec{X_{16}} = (7,8)'$, $\vec{X_{17}} = (8,8)'$, $\vec{X_{18}} = (9,88)'$, $\vec{X_{19}} = (8,9)'$, $\vec{X_{20}} = (9,9)'$,



Slika 10: Rezultat grupiranja K sr. vr

1. korak K=2; $\vec{Z}_1(1) = \vec{X}_1 = (0,0)'$ i $\vec{Z}_2(1) = \vec{X}_1 = (1,0)'$

2. korak Budući da je $\left\|\vec{X}_1 - \vec{Z}_1(1)\right\| < \left\|\vec{X}_1 - \vec{Z}_i(1)\right\|$ i $\left\|\vec{X}_3 - \vec{Z}_1(1)\right\| < \left\|\vec{X}_3 - \vec{Z}_i(1)\right\|$ i=2 imamo:
$S_1(1) = \left\{\vec{X}_1, \vec{X}_3\right\}$ $S_2(1) = \left\{\vec{X}_2, \vec{X}_4, \ldots, \vec{X_2}0\right\}$;

3. korak Računamo nova središta grupa:

$$\vec{Z}_1(2) = \tfrac{1}{N_1} \sum_{\vec{X} \in S_1(1)} \vec{X} = \tfrac{1}{2}(\vec{X}_1 + \vec{X}_3) = \begin{pmatrix} 0.0 \\ 0.5 \end{pmatrix}$$

$$\vec{Z}_2(2) = \tfrac{1}{N_2} \sum_{\vec{X} \in S_2(1)} \vec{X} = \tfrac{1}{18}(\vec{X}_2 + \vec{X}_4 + \ldots + \vec{X_{20}}) = \begin{pmatrix} 5.67 \\ 6.33 \end{pmatrix}$$

4. korak Budući da je $\vec{Z}_j(2) \neq \vec{Z}_j(1)$: j=1,2; vraćamo se na 2. korak

2'. korak $\left\|\vec{X}_l - \vec{Z}_1(2)\right\| < \left\|\vec{X}_l - \vec{Z}_2(2)\right\|$ za $l = 1, 2, \ldots, 8$ i
$\left\|\vec{X}_l - \vec{Z}_2(2)\right\| < \left\|\vec{X}_l - \vec{Z}_1(2)\right\|$ za $l = 9, 10, \ldots, 20$
$S_1(2) = \left\{\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_8\right\}$
$S_2(2) = \left\{\vec{X}_9, \vec{X_{10}}, \ldots, \vec{X_{20}}\right\}$

3'. korak Obnovimo vrijednosti centara:

$$\vec{Z_1}(3) = \frac{1}{N_1} \sum_{\vec{X} \in S_1(2)} \vec{X} = \frac{1}{8}(\vec{X_1} + \vec{X_2} + \ldots, \vec{X_8}) = \begin{pmatrix} 1.25 \\ 1.13 \end{pmatrix}$$

$$\vec{Z_2}(3) = \frac{1}{N_2} \sum_{\vec{X} \in S_2(2)} \vec{X} = \frac{1}{12}(\vec{X_9} + \vec{X_{10}} + \ldots + \vec{X_{20}}) = \begin{pmatrix} 7.67 \\ 7.33 \end{pmatrix}$$

4. korak Budući da $\vec{Z_j}(3) \neq \vec{Z_j}(2)$: za j=1,2; vraćamo se na 2. korak

2". korak i 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 :

$$\vec{Z_1} = \begin{pmatrix} 1.25 \\ 1.13 \end{pmatrix}$$

$$\vec{Z_2} = \begin{pmatrix} 7.67 \\ 7.33 \end{pmatrix}$$

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 $m_i(0)$
- $zat = 0, 1, 2, \ldots$ izračunajmo:

(1) centar mjehurića (c):
$\|X(t) - m_c(t)\| = \min_i \{\|X(t) - m_i(t)\|\}$

(2) "popravimo" vrijednost težinskih vektora $m_i(t+1) = m_i(t) + \alpha(t)(X(t) - m_i(t))$, za $i \in N_c$
N_c - broj elemenata u radijusu c "mjehurić"
$m_i(t+1) = m_i(t)$ za sve ostale i
$\alpha = \alpha(t)$ - monotono padajuća funkcija vremena; i $N_c = N_c(t)$ su empirijske funkcije vremena

t_1, t_2, t_3 - threshold values; t_1-splitting, t_2-merging, t_3-discarding
ISODATA (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 spit 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