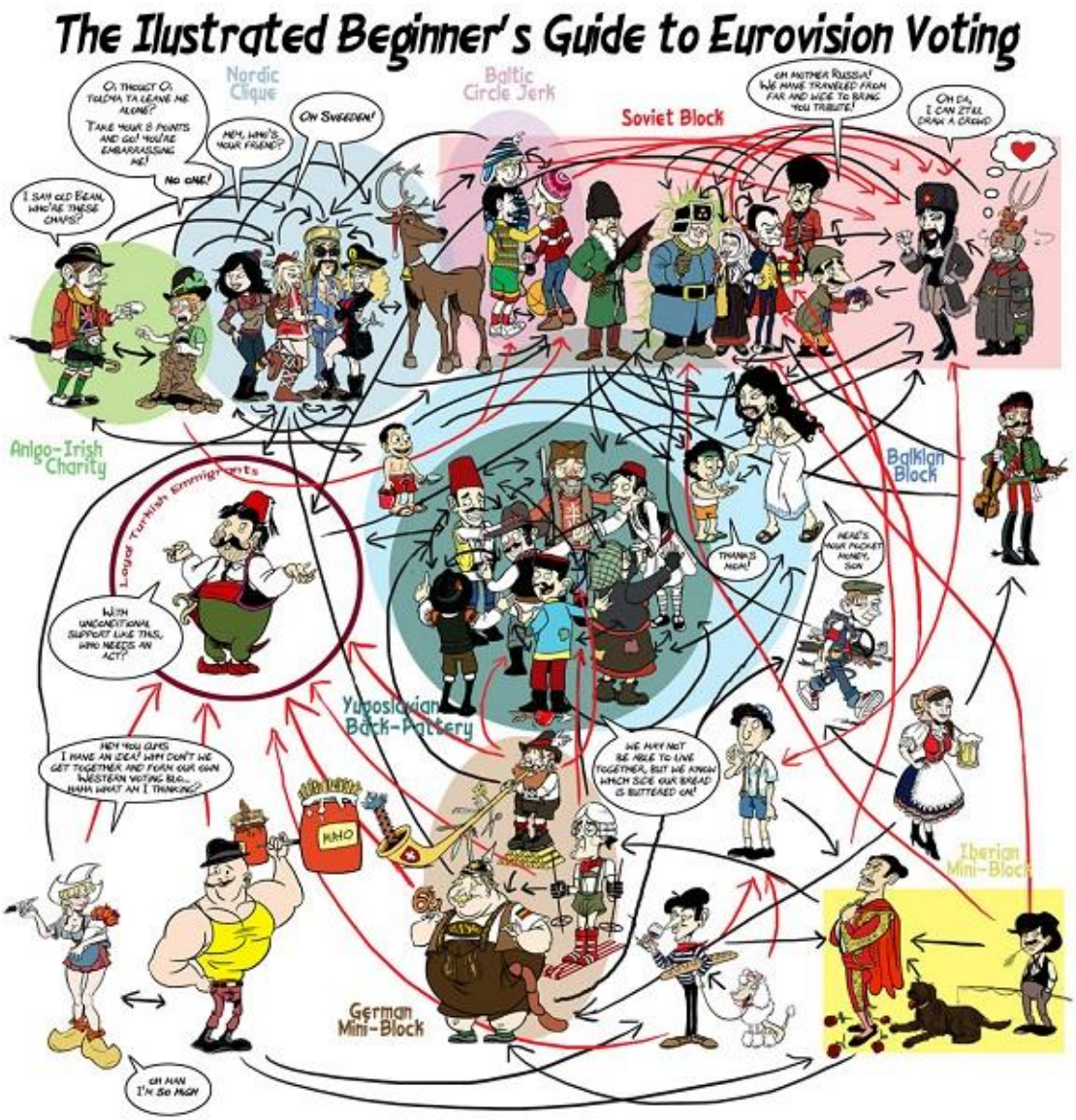
ODKRIVANJE SKUPIN

clustering

doc. dr. Matej Guid

Fakulteta za računalništvo in informatiko
Univerza v Ljubljani

januar 2022



Kaj je naravna razdelitev v skupine?

DOLOČANJE SKUPIN JE SUBJEKTIVNO!

ženske

moški

imajo brke

nimajo brkov

KAJ JE PODOBNOST?

„**podóbnost** -i ž (ọ́) *lastnost, stanje podobnega:* opaziti podobnost med starši in otroki“ (SSKJ)

podobnost je težko definirati… *dokler je ne opazimo* ☺

PRIMER: YIPPY (PREJ CLUSTY)

(meta)iskalnik

temelji na tehnologiji Clusty iz univerze Carnegie Mellon

prikaz kategorij

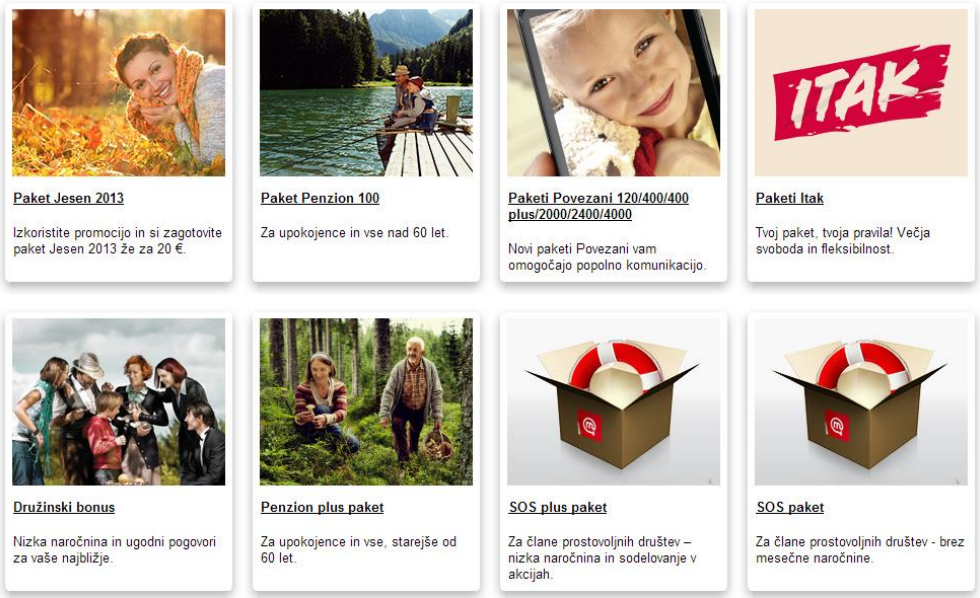
prikaz zadetkov izbrane kategorije

vrača skupine (*clusters*!) rezultatov… „iskalnik primeren za družine in otroke vseh starosti“

PRIMER: SEGMENTACIJA TRGA ZA NAMENE TRŽENJA

- segmentacija uporabnikov je eden izmed ključnih dejavnikov za uspešno trženje

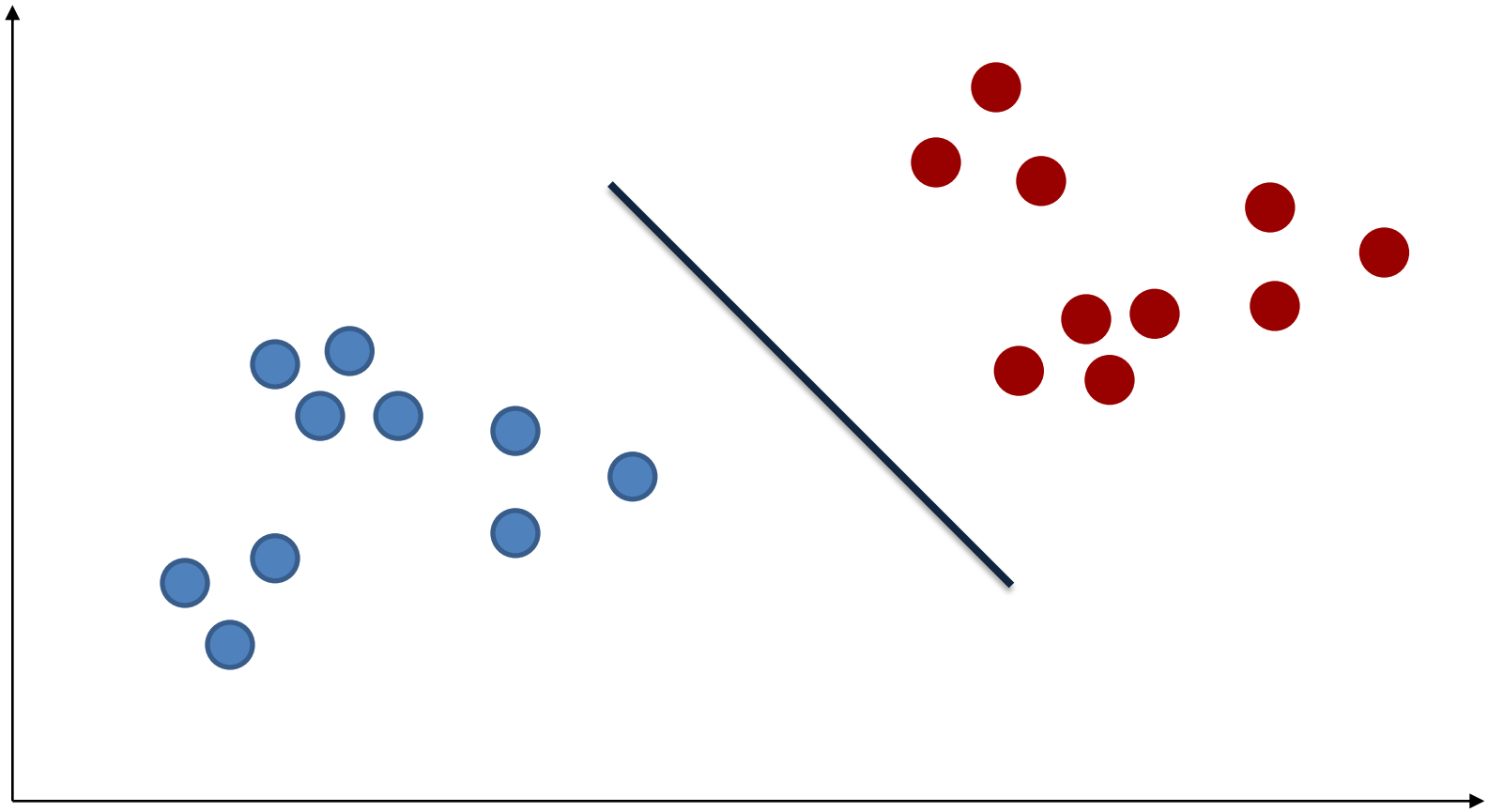
- s pomočjo tehnik **odkrivanja skupin** lahko potencialne kupce razdelimo v ustrezne skupine
- za vsako skupino lahko potem tržniki uporabijo drugačen pristop, način komunikacije itd.



NADZOROVANO UČENJE…

nadzorovano učenje

supervised learning

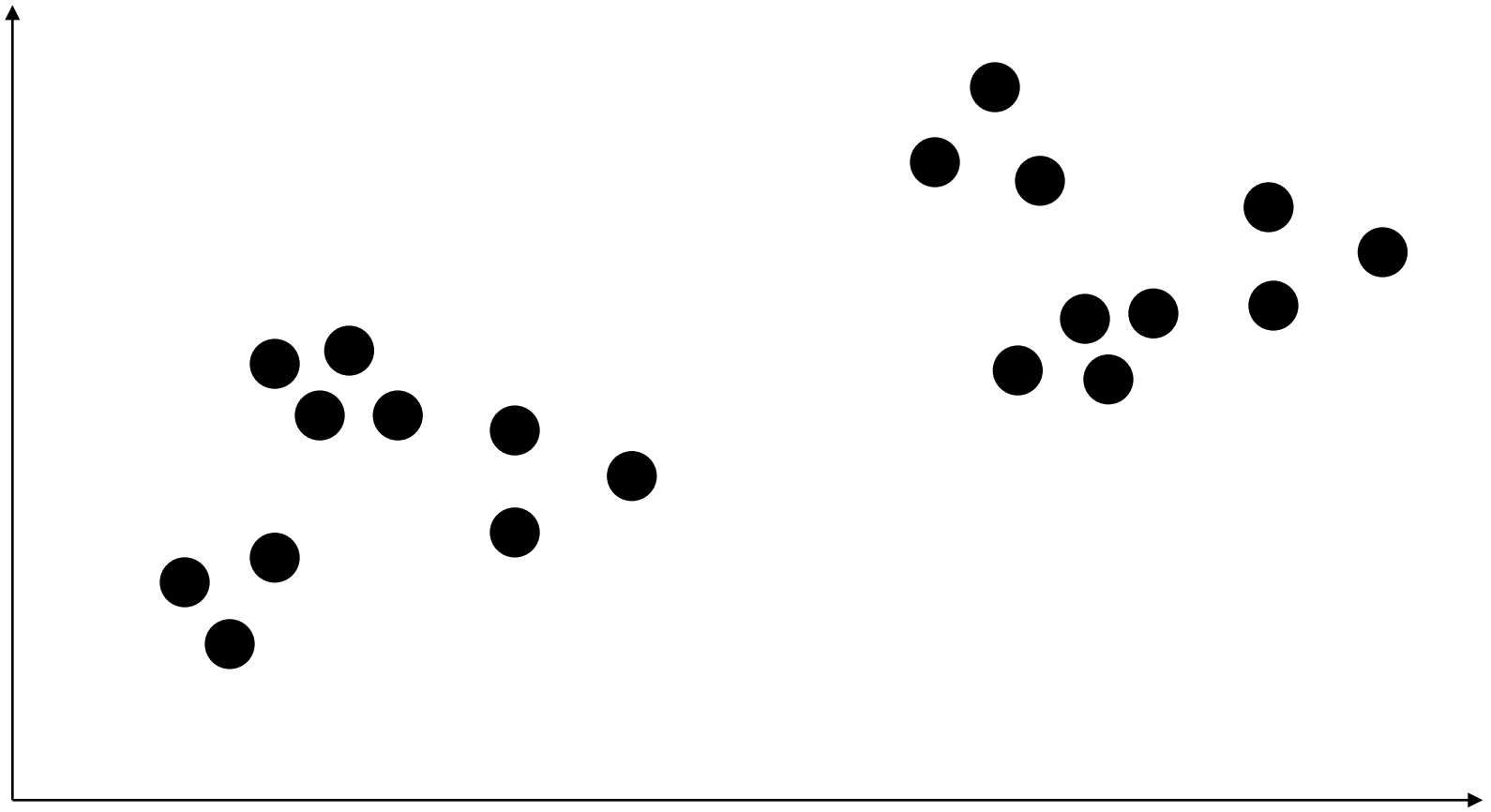


- strojno učenje iz učne množice primerov, ki imajo **pripadajoče izhodne vrednosti**
- rezultat: **napovedni model**, za poljuben vhodni primer napove pripadajočo izhodno vrednost

… IN NENADZOROVANO UČENJE…

nenadzorovano učenje

unsupervised learning

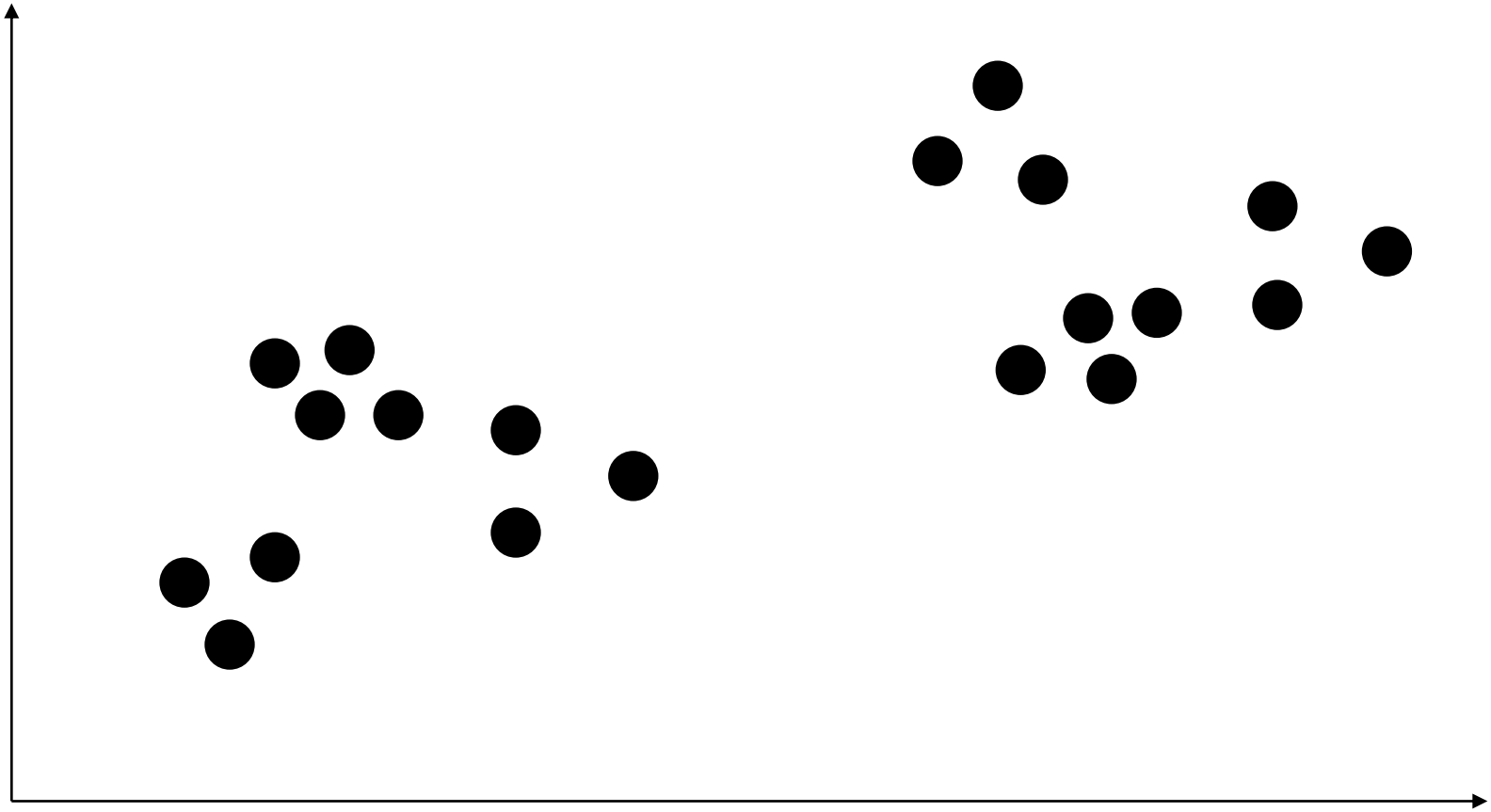


- strojno učenje iz učne množice primerov
- rezultat: **povzetek**, **pojasnjevanje** učne množice primerov

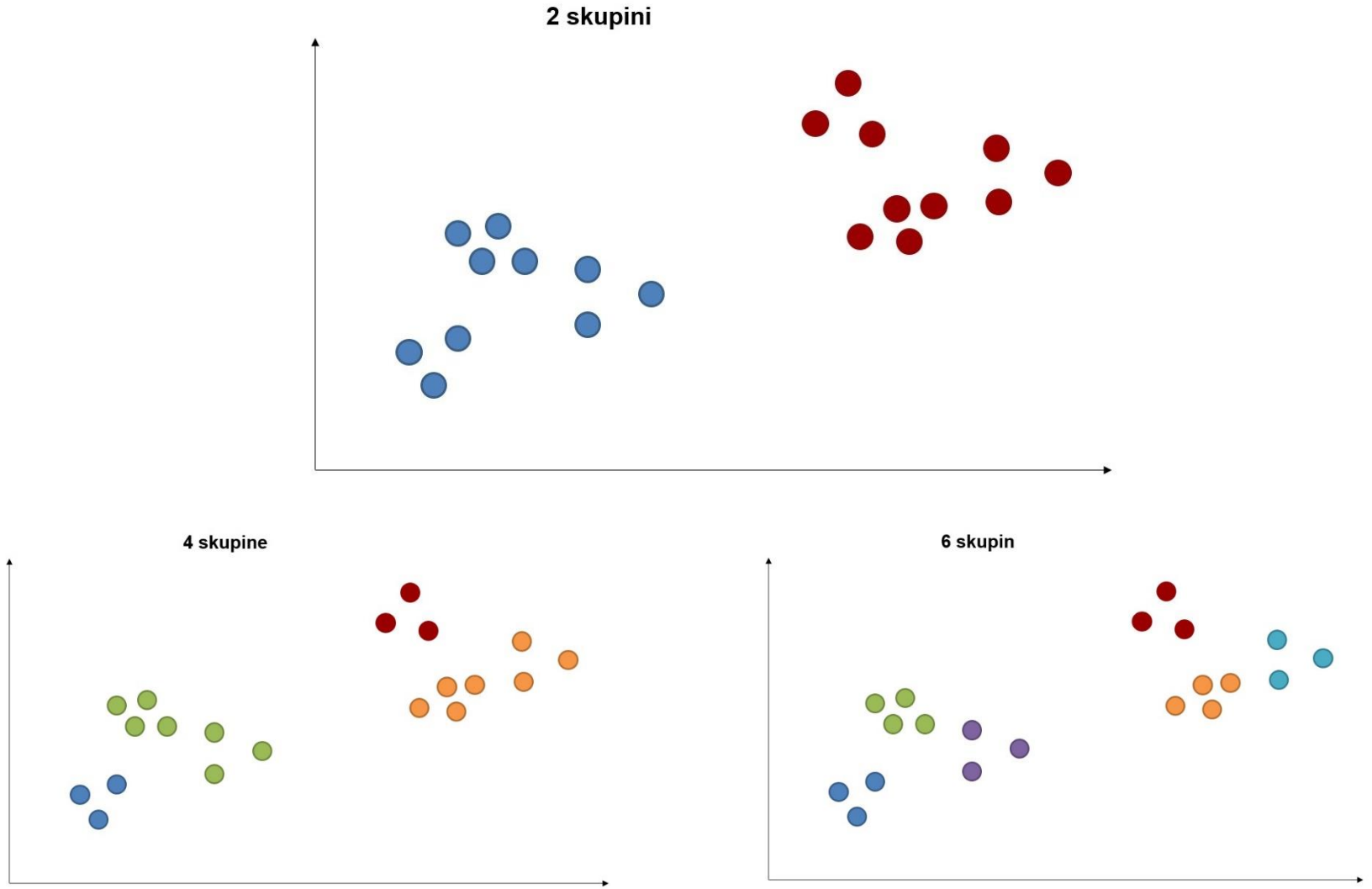
ODKRIVANJE SKUPIN

KOLIKO SKUPIN?

Na koliko skupin bi bilo smiselno razdeliti naslednje primere?



KOLIKO SKUPIN?



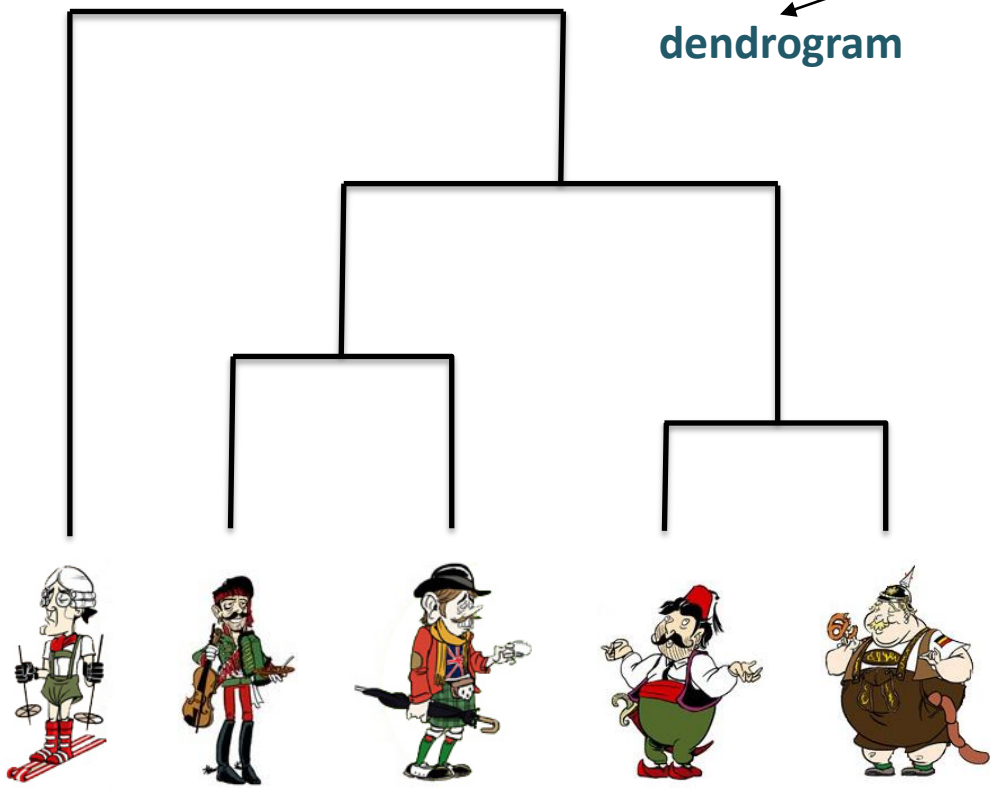
Število skupin je mnogokrat odvisno od potreb naročnika oz. končnega uporabnika!

Tipi odkrivanja skupin

razbitje razvrstitev primerov v skupine, ki se med seboj ne prekrivajo
vsak primer pripada natanko eni skupini

hierarhično razvrščanje v skupine množica gnezdenih skupin je organizirana v hierarhično drevo

dendrogram



hierarhija skupin

razbitje

Zaželene lastnosti algoritmov za odkrivanje skupin

- **skalabilnost** (tako glede časovne kot tudi prostorske zahtevnosti)
- zmožnost dela z različnimi tipi podatkov
- minimalno poznavanje znanja o domeni (npr. za določitev vhodnih parametrov)
- **neobčutljivost** na šum (ang. *noise*) in prisotnost osamelcev (ang. *outliers*)
- neodvisnost od podanega vrstnega reda učnih primerov
- možnost podajanja omejitev (temelječih na znanju o domeni)
- **enostavnost interpretacije** rezultatov, uporabnost

Učni podatki

Rezultati „ekstercev“ za 16 učencev razreda 8.A. Rezultati posameznih predmetov so podani v percentilih z ozirom na republiške rezultate.

ime	slo	ang	zgo	geo	mat	bio	fiz	kem	tel
Albert	22	81	32	39	21	37	46	36	99
Branka	91	95	65	96	89	39	11	22	29
Cene	51	89	21	39	100	59	100	89	27
Dea	9	80	18	34	61	100	90	92	8
Edo	93	99	39	100	12	47	17	12	63
Franci	49	83	17	33	92	30	98	91	73
Helena	91	99	97	89	49	96	81	94	69
Ivan	12	69	32	14	34	12	33	48	96
Jana	91	80	20	10	82	93	87	91	22
Leon	39	100	19	29	99	31	77	79	23
Metka	20	91	10	15	71	99	78	93	12
Nika	90	60	45	34	45	20	15	5	100
Polona	100	98	97	89	32	72	22	13	37
Rajko	14	4	15	27	61	42	51	52	39
Stane	9	22	8	7	100	11	92	96	29
Zala	85	90	100	99	45	38	92	67	21

Ali so v razredu kakšne značilne skupine učencev?

Kako se razlikujejo med sabo? Kateri učenci so si med seboj podobni?

MERE RAZLIČNOSTI

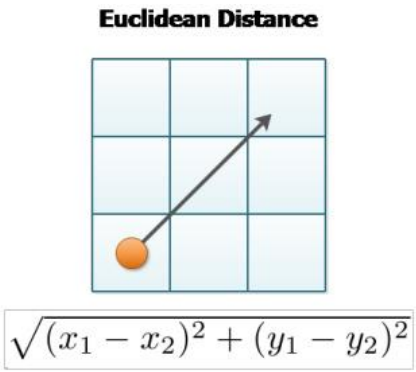
Vzemimo učna primera a in b:

$a = (a_1, a_2, \ldots, a_m)$
$b = (b_1, b_2, \ldots, b_m)$

Želimo poiskati take množice primerov, kjer so si le-ti med seboj čim bolj podobni!

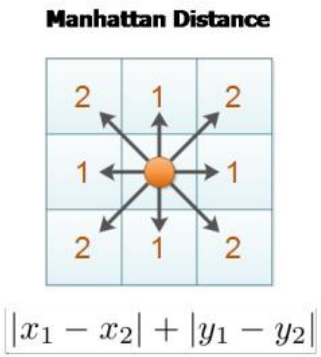
Evklidska razdalja

$$d(a, b) = \sqrt{\sum_{i=1}^{m}(a_i - bi)^2}$$



Manhattanska razdalja

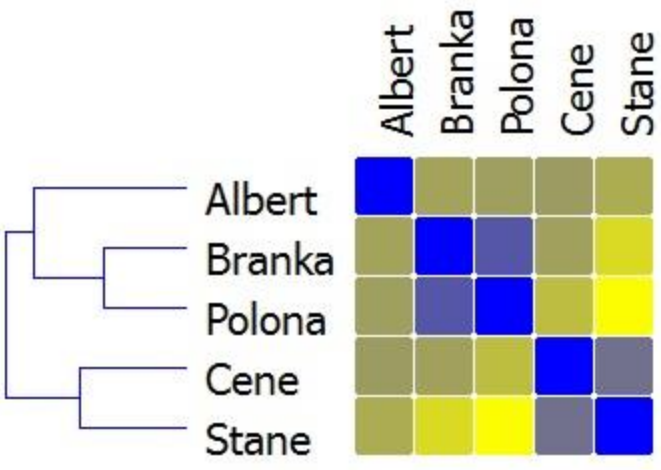
$$d(a, b) = \sum_{i=1}^{m} |ai - bi|$$



Kako pa bi merili razdalje v naslednjih primerih?

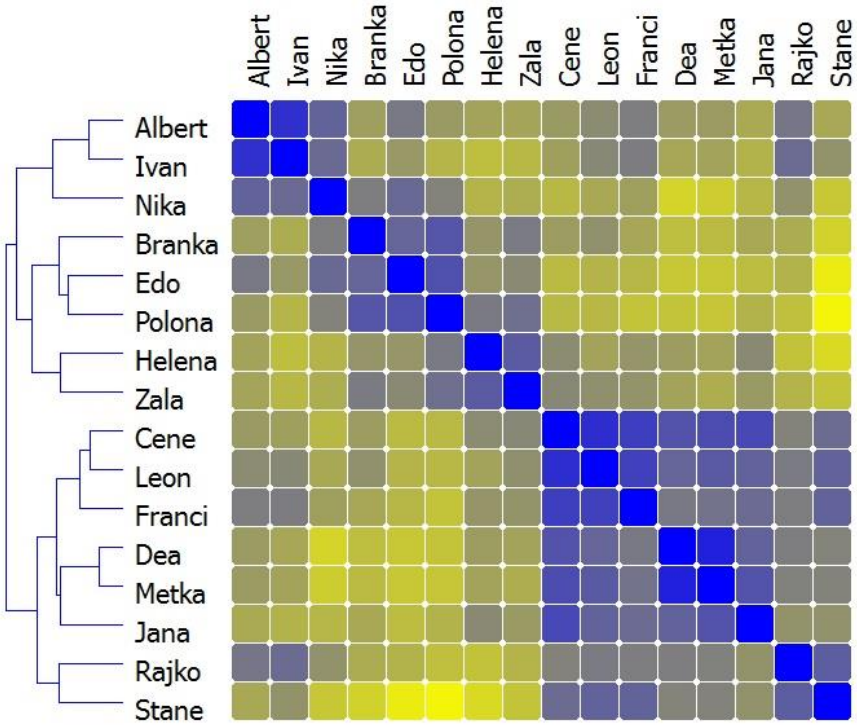
- atributi, ki zavzemajo diskretne vrednosti
- besedila, nizi znakov, dokumenti
- predmeti pri priporočilnih sistemih

Vizualizacija različnosti



mera: evklidska razdalja

kako so pari urejeni?
kakšne vzorce opazimo?



domenski atributni prostor Evklidski

primerna mera različnosti med primeri je evklidska razdalja

ODKRIVANJE SKUPIN: DEFINICIJE

učni primeri, učni podatki *training examples, training data*

$x^{(i)}$ = i-ti primer, $x \in U$, $i = 1 \ldots m$

$x_j^{(i)}$ = vrednost j-te značilke i-tega primera

$\mathbf{x} = (x_1, x_2, \ldots, x_m)^T$

$x_{slo}^{(Albert)} = 22$

število primerov $m = |U|$ m = moč učne množice

skupine *clusters*

$C \subset U$ skupina primerov

$|C_i| > 0$ zanimajo nas neprazne skupine

$\bigcup_{C_i \in C} C_i = U$ unija vseh skupin je enaka učni množici

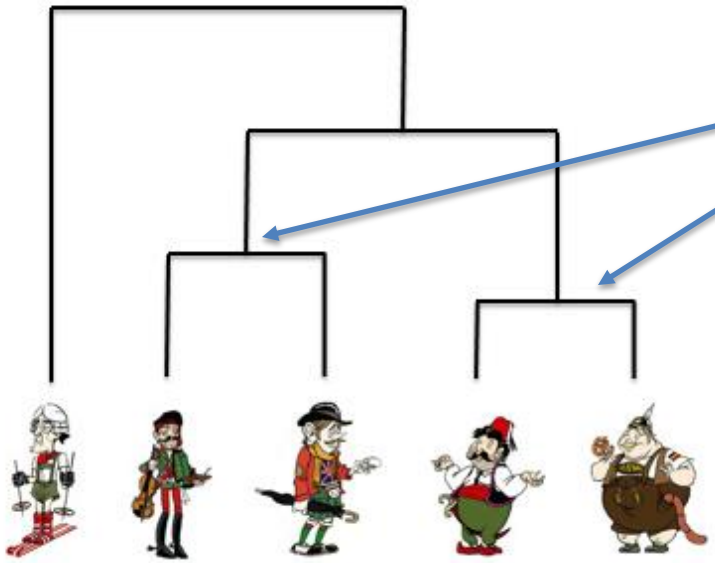
razbitje $C_i \cap C_j = \emptyset; \quad C_i, C_j \in C$ skupine si med seboj ne delijo nobenega primera

hierarhija skupin $C_i \cap C_j \in \{C_i, C_j, \emptyset\}$ presek dveh skupin je v celoti ena od skupin ali pa prazna množica

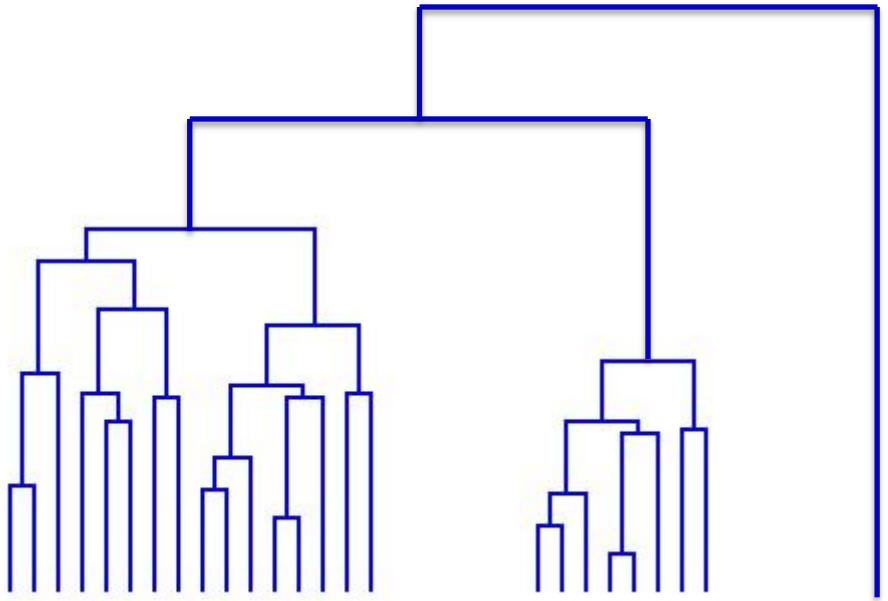
Dendrogram in hierarhično razvrščanje v skupine

dendrogram

- orodje za prikazovanje podobnosti (različnosti) med skupinami,
- ponazori postopek združevanja skupin



Mera podobnosti
višina najnižjega skupnega vozlišča



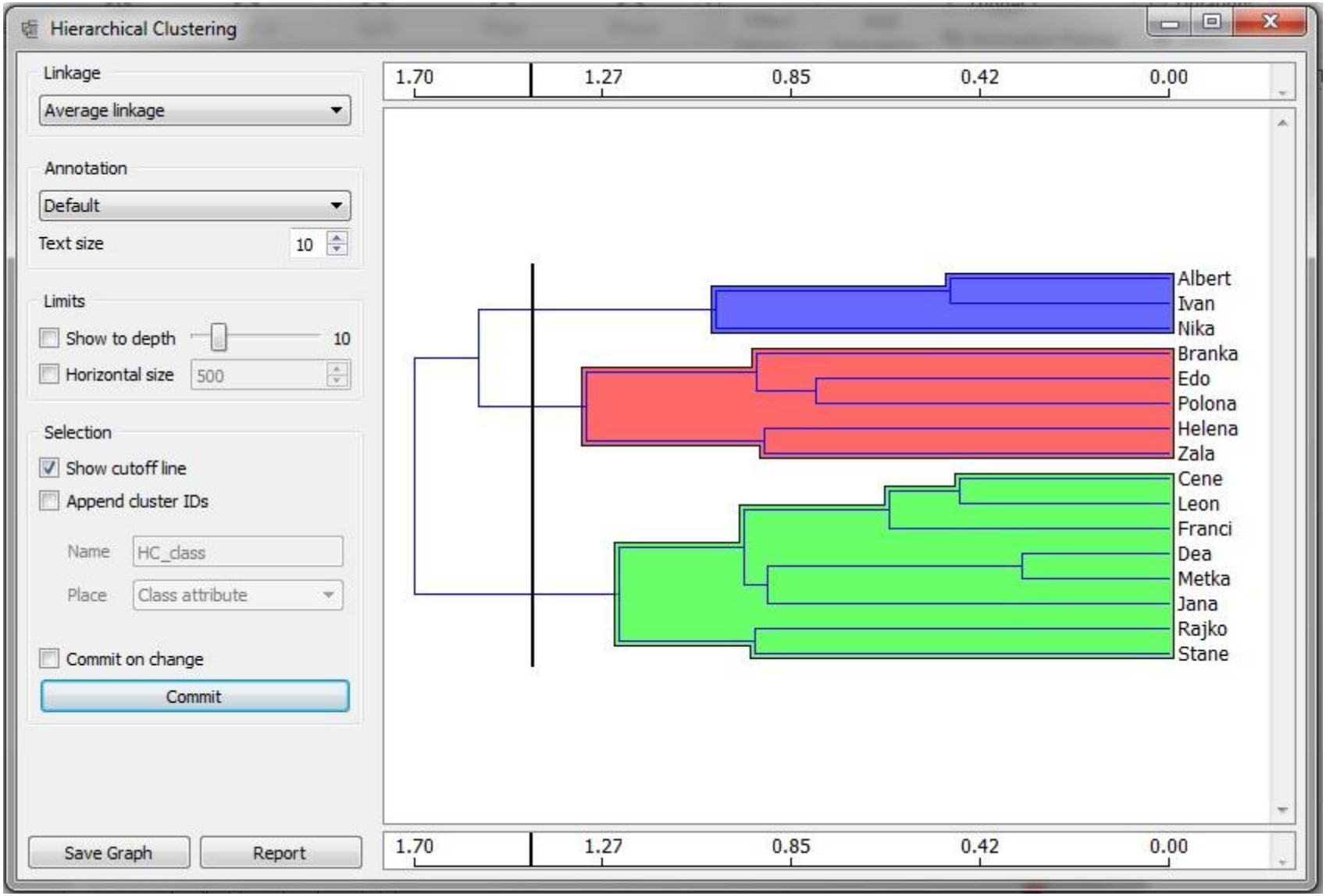
2 skupini

osamelec

včasih lahko s pomočjo dendrograma hitro ugotovimo „pravo“ število skupin...

...in morebitne osamelce (ang. *outliers*)

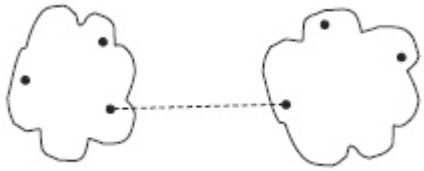
Dendrogrami v okolju Orange



Ocenjevanje razdalj med skupinami

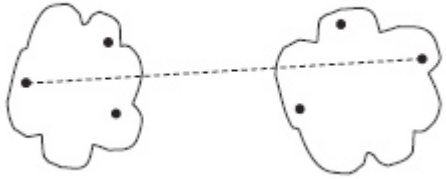
skupine podane z nizom primerov → **kako meriti razdalje med skupinami?**

razdalja med najbližjima primeroma



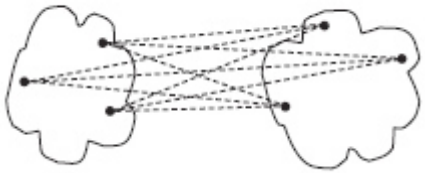
single link

razdalja med najbolj oddaljenima primeroma



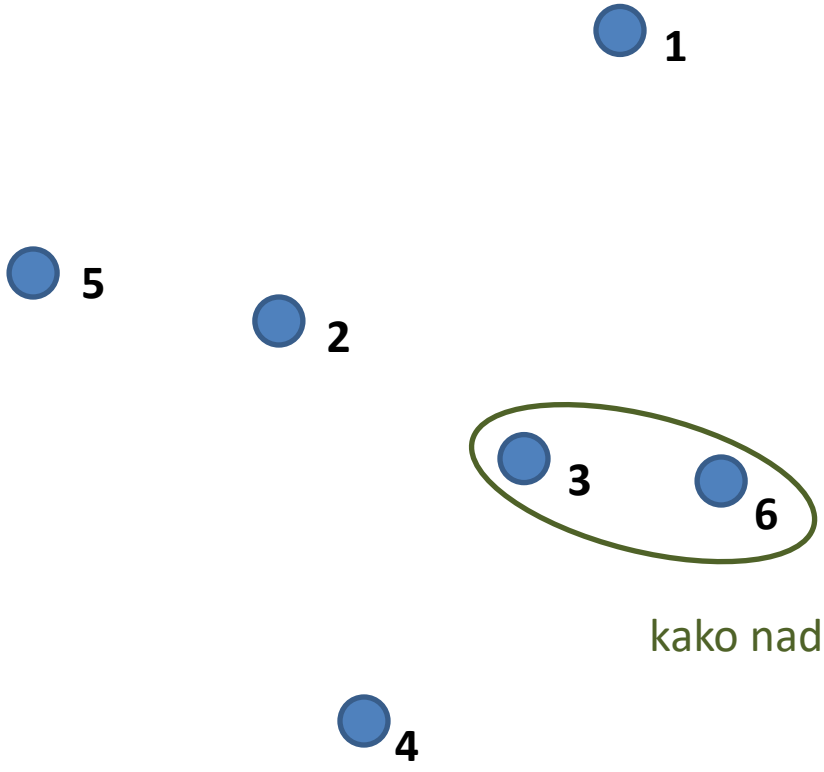
complete link

povprečna razdalja med vsemi pari primerov

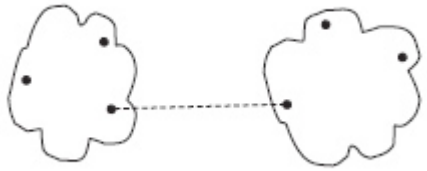


group average

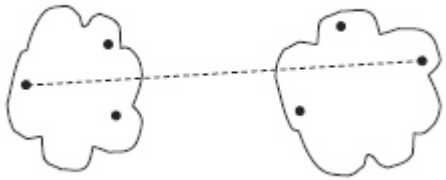
KAKO MERITI RAZDALJE MED SKUPINAMI?



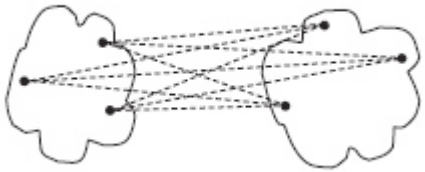
kako nadaljevati?



single link



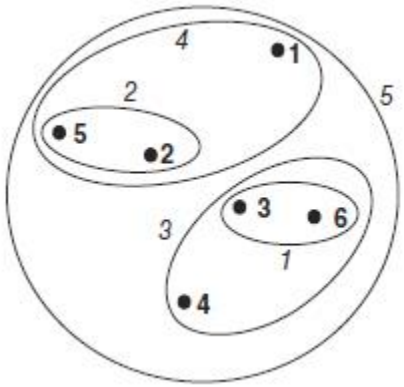
complete link

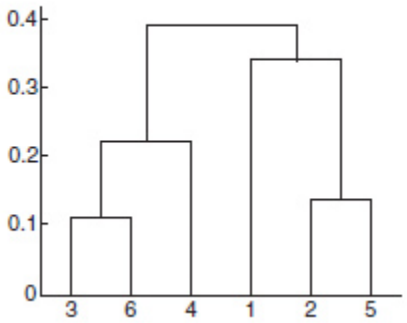


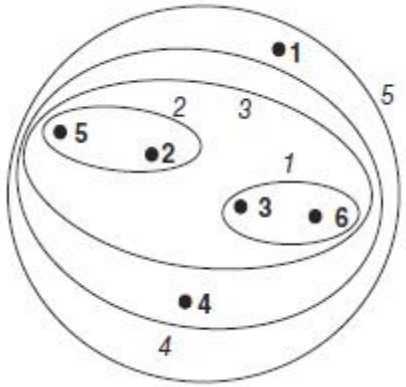
group average

RAZLIČNI PRISTOPI: RAZLIČNI REZULTATI

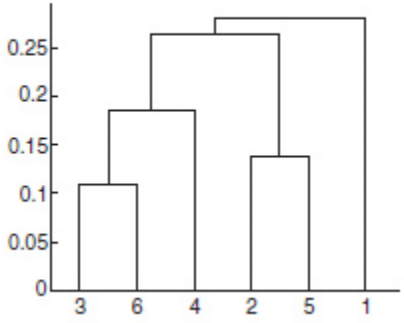
complete link

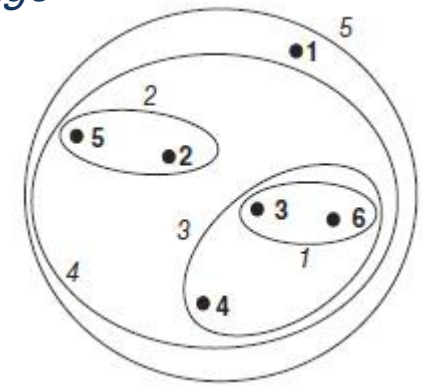




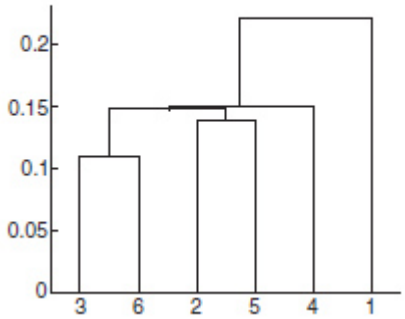


group average





single link



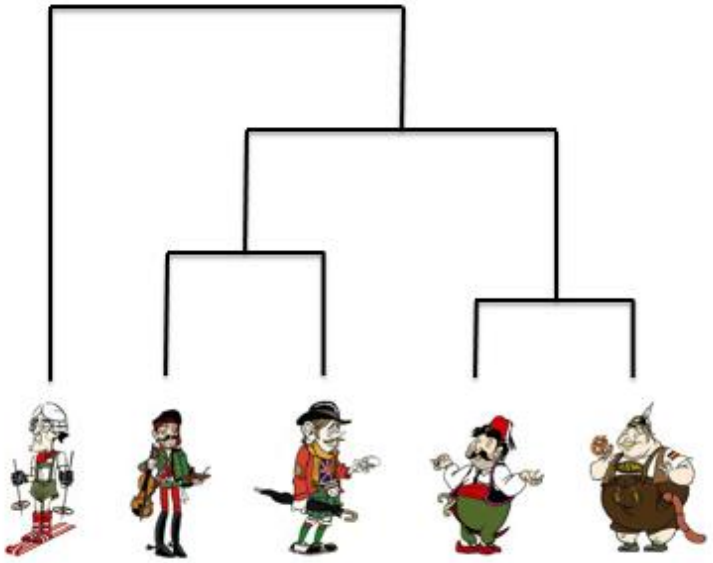
Alternativa je **Wardova razdalja**, ki uvaja podobno objektivno funkcijo kot metoda voditeljev…..

Hierarhično razvrščanje v skupine

število možnih dendrogramov z *n* listi:

$$\frac{(2n-3)!}{2^{(n-2)}(n-2)!}$$

število listov	število dendrogramov
3	3
4	15
5	105
…	…
10	34.459.425



časovna zahtevnost: O(m^2 log(m))

od spodaj navzgor *Bottom-Up: agglomerative*

Na začetku naj vsak primer tvori svojo skupino. Poišči najboljši par za združitev v novo skupino. Ponavljaj, dokler niso vse skupine združene v eno.

od zgoraj navzdol *Top-Down: divisive*

Na začetku naj vsi primeri tvorijo eno skupino. Poišči možnosti za razdelitev skupine na dve. Izberi najboljšo razdelitev in rekurzivno nadaljuj postopek na obeh straneh.

Matrika razdalj

d( , ) = ? **8**

d( , ) = ? **2**







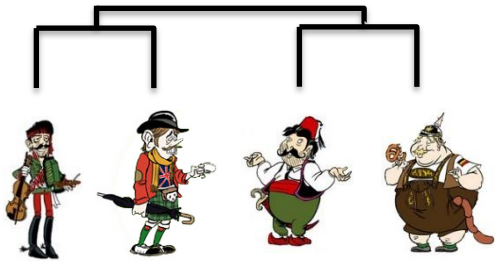


0				
6	0			
6	2	0		
7	5	3	0	
8	5	3	1	0

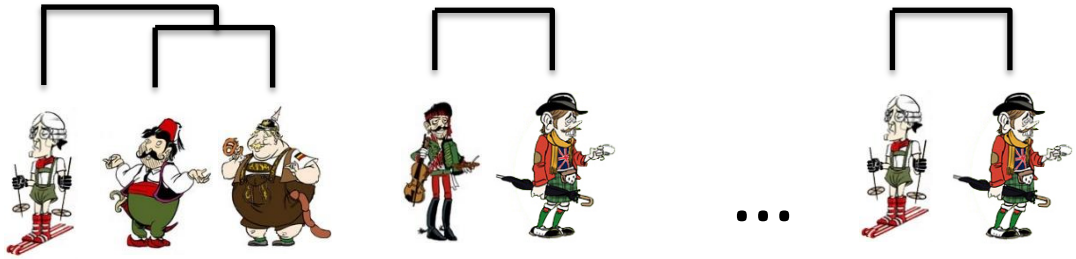
Hierarhično razvrščanje v skupine: postopek

Na začetku naj vsak primer tvori svojo skupino. Poišči najboljši par za združitev v novo skupino. Ponavljaj, dokler niso vse skupine združene v eno.

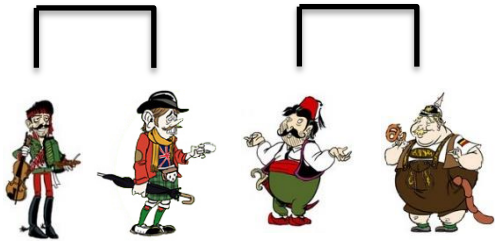
vzemi v obzir vse možne združitve…



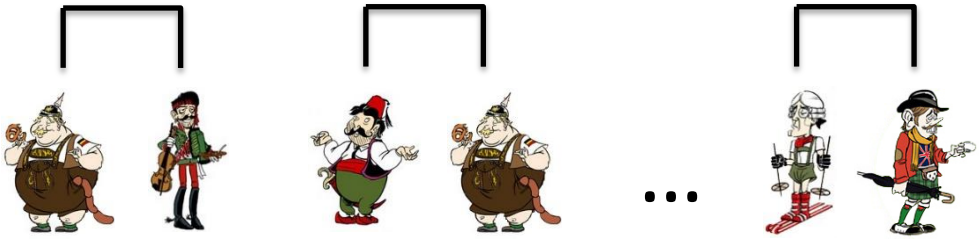
… in izberi najboljšo



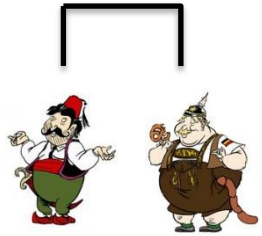
vzemi v obzir vse možne združitve…



… in izberi najboljšo

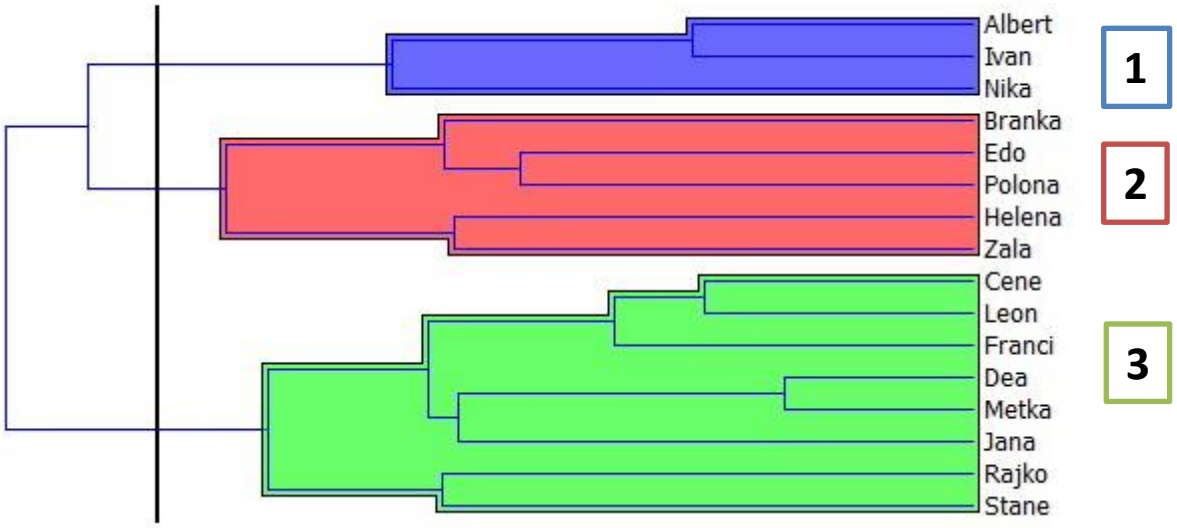


vzemi v obzir vse možne združitve…



… in izberi najboljšo

Razlaga



Kaj so lastnosti identificiranih skupin?

V čem so dobri pripadniki posameznih skupin?

	slo	ang	zgo	geo	mat	bio	fiz	kem	tel
U	54	78	40	47	62	52	62	61	47
skupina 1	41	70	36	29	33	23	31	30	98
skupina 2	92	96	80	95	45	58	45	42	44
skupina 3	35	69	16	24	83	58	84	85	29

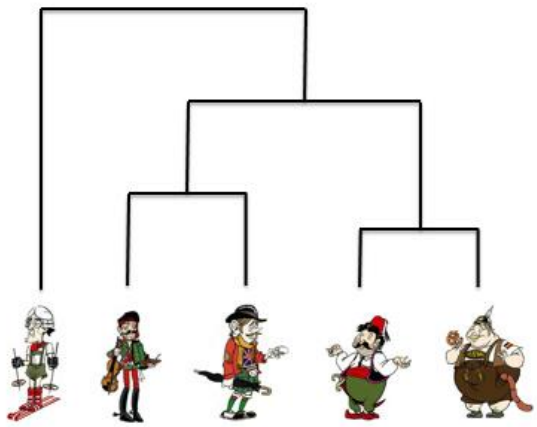
šport

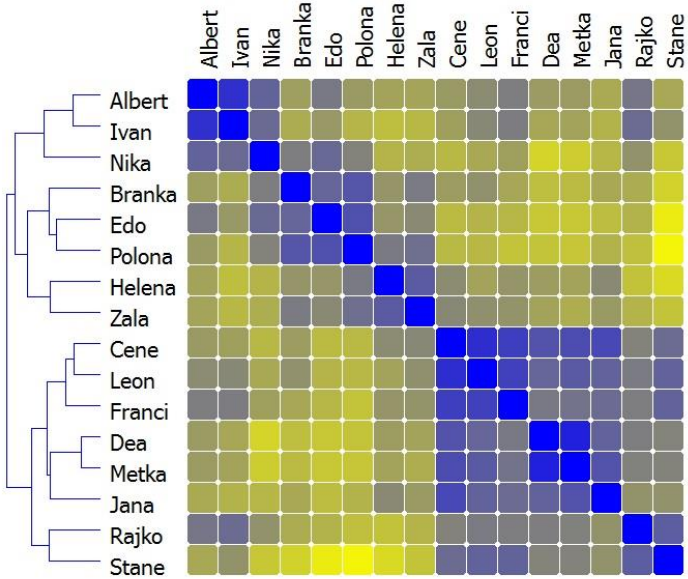
družboslovni predmeti

naravoslovni predmeti

HIERARHIČNO RAZVRŠČANJE V SKUPINE: POVZETEK

- ni potrebno vnaprej opredeliti števila skupin
- v nekaterih domenah se rezultati lepo skladajo s človeško intuicijo
- časovno zelo potratni pri velikem številu primerov (problem: skalabilnost!)
- ne vodijo nujno k „idealni“ rešitvi (problem: lokalni optimumi)
- interpretacija rezultatov je lahko (zelo) subjektivna

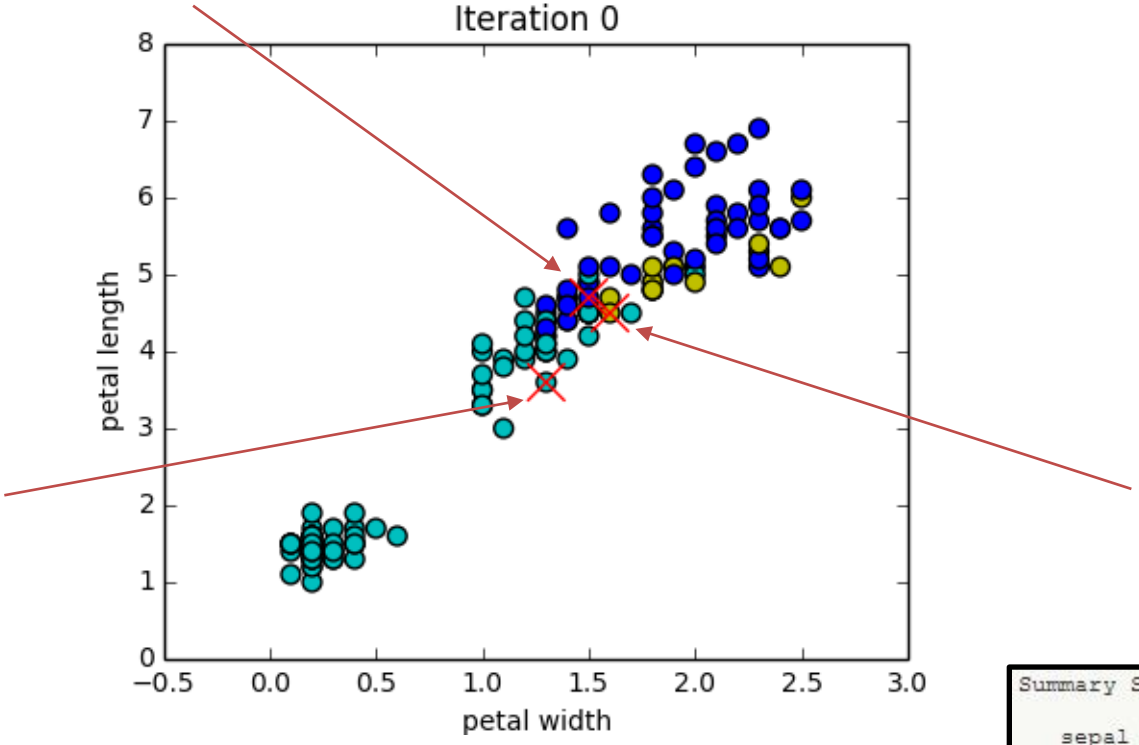




metoda voditeljev

K-means

- uporabnik vnaprej določi število skupin K
- voditelji določajo te skupine oziroma so središča skupin, njihova lega se spreminja



Iris Plants Database

Instances: 150 (3 classes)

Attributes: 4 (numeric)

Class Distribution: 33.3% for each class

```
Summary Statistics:
                Min  Max   Mean    SD   Class Correlation
   sepal length: 4.3  7.9   5.84  0.83    0.7826
    sepal width: 2.0  4.4   3.05  0.43   -0.4194
   petal length: 1.0  6.9   3.76  1.76    0.9490  (high!)
    petal width: 0.1  2.5   1.20  0.76    0.9565  (high!)
```

METODA VODITELJEV: POSTOPEK

metoda voditeljev *K-means*

1. prični s *k* naključno izbranimi voditelji, **M** = $\{m^{(i)}; i \in 1 \ldots K\}$
2. <u>ponavljaj</u>
   - določi razvrstitev C tako, da vsak primer priredíš najbližjemu voditelju
   - novi voditelji naj bodo centroidi $R(C_i)$ skupin $C_i \in \mathbf{C}$, $m^{(i)} \leftarrow R(C_i)$
3. <u>dokler</u> se lega voditeljev spreminja

centroidi so navadno kar geometrijska središča primerov skupine: *centroids*

$$R(C_i) = \frac{1}{|C_i|} \sum_{x \in Ci} x$$

O(IKm) I = število iteracij

Za srednje velike probleme (npr. nekaj 1000 primerov) je lahko potrebnih manj kot 100 iteracij.

Začetni izbor voditeljev

Kvaliteta razbitja in število iteracij sta zelo odvisna od začetnega izbora voditeljev!

naključni izbor voditeljev

lahko vodi do neoptimalnega razbitja

izbor razpršenih voditeljev

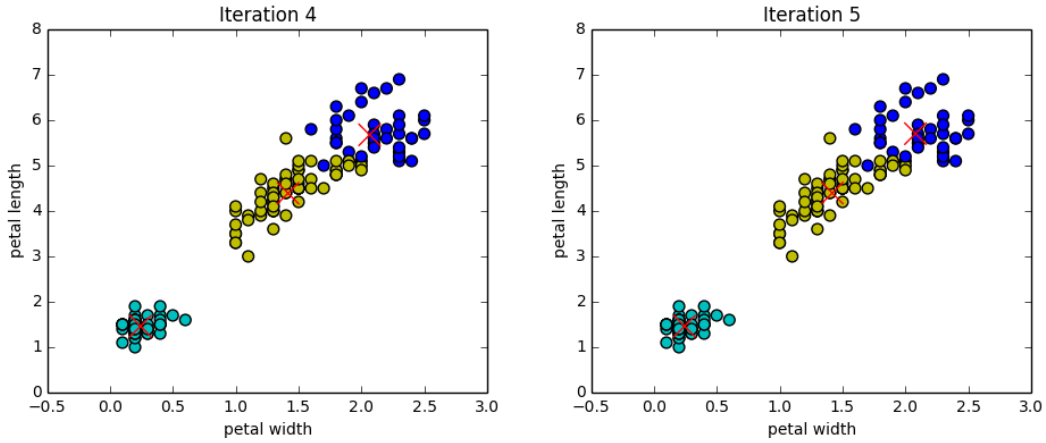
- prvi voditelj: najbolj oddaljen od drugih
- vsi nadaljnji voditelji naj bodo najbolj oddaljeni od voditeljev, ki smo jih že določili

uporaba hierarhičnega razvrščanja

- s hierarhičnim razvrščanjem v skupine poiščemo K skupin
- njihova središča uporabimo kot začetne voditelje

Ustavitveni kriterij

- postopek tipično ustavimo takrat, ko noben primer ne zamenja svojega centroida



- pri velikih množicah primerov lahko določimo število dovoljenih zamenjav (npr. 10)

Iris Plants Database

Instances: 150 (3 classes)

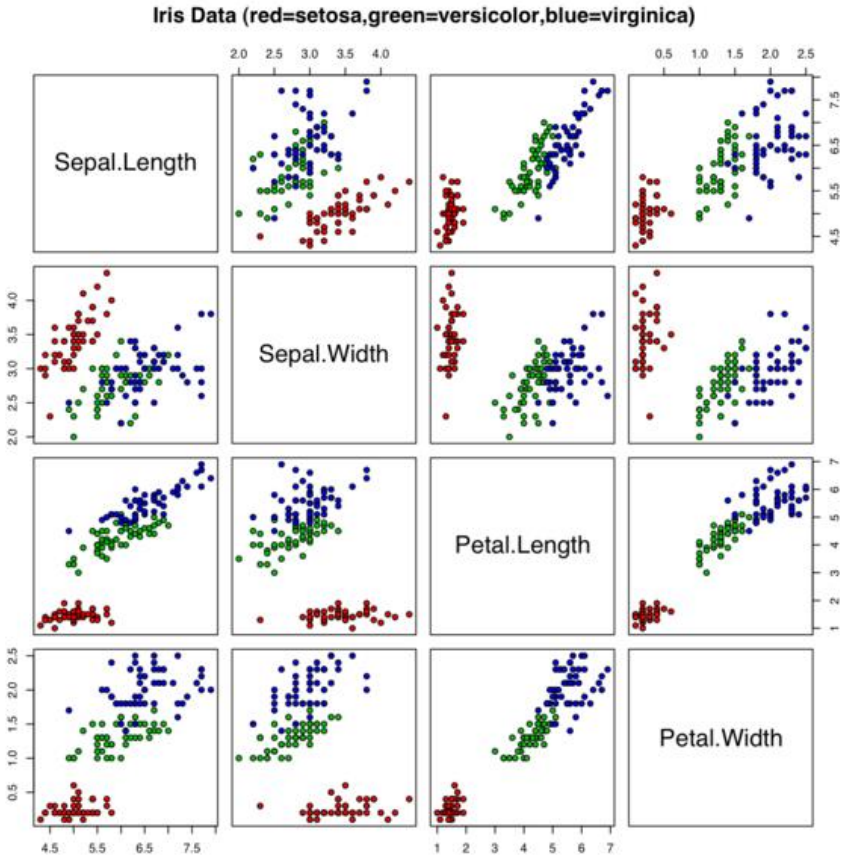
Attributes: 4 (numeric)

Class Distribution: 33.3% for each class

```
Summary Statistics:
                Min  Max   Mean    SD   Class Correlation
   sepal length: 4.3  7.9   5.84  0.83    0.7826
    sepal width: 2.0  4.4   3.05  0.43   -0.4194
   petal length: 1.0  6.9   3.76  1.76    0.9490  (high!)
    petal width: 0.1  2.5   1.20  0.76    0.9565  (high!)
```

VEČDIMENZIONALNOST

Ne pozabimo: opravka imamo z več kot le dvema značilkama (atributoma)!



podrobnosti: *Iris* flower data set (Wikipedia)

Iris Plants Database

Instances: 150 (3 classes)

Attributes: 4 (numeric)

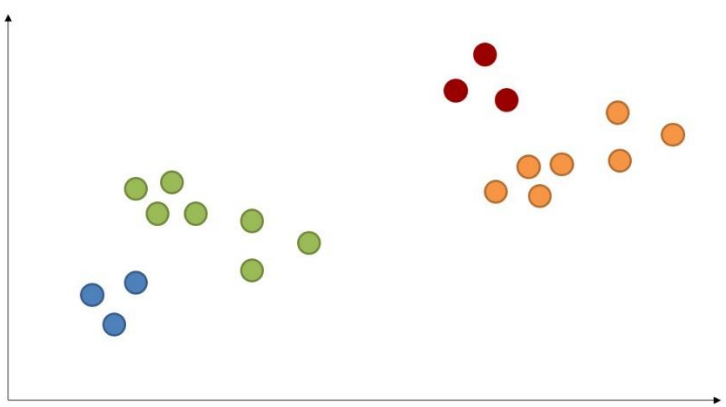
Class Distribution: 33.3% for each class

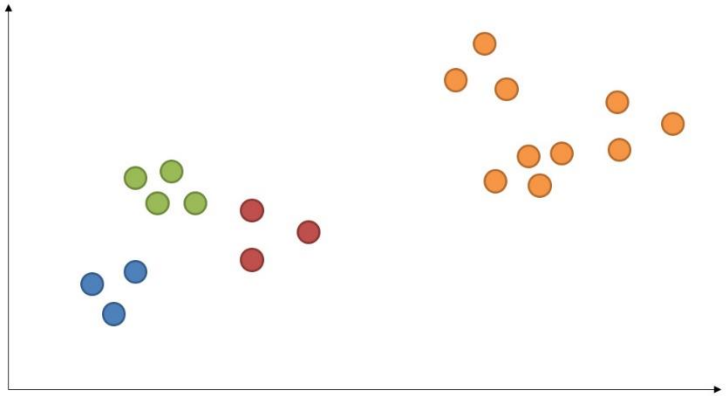
Summary Statistics:

	Min	Max	Mean	SD	Class Correlation	
sepal length:	4.3	7.9	5.84	0.83	0.7826	
sepal width:	2.0	4.4	3.05	0.43	-0.4194	
petal length:	1.0	6.9	3.76	1.76	0.9490	(high!)
petal width:	0.1	2.5	1.20	0.76	0.9565	(high!)

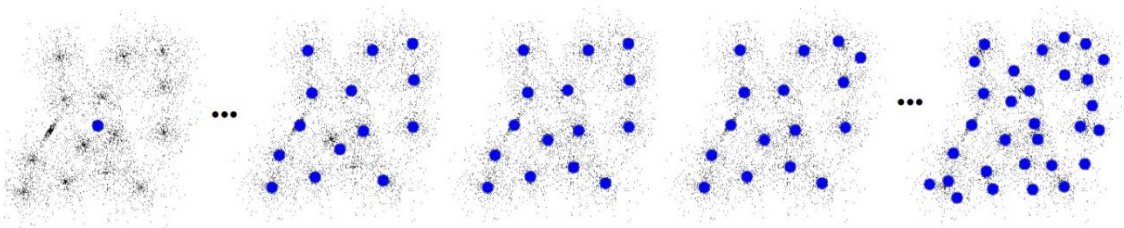
„PRAVILNA“ RAZDELITEV NA SKUPINE…

nesrečna postavitev začetnih voditeljev tipično vodi do **lokalnega optimuma**…





Kako pa vemo, katera razdelitev na skupine je „bolj pravilna”?

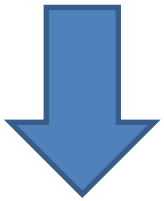


Ne pozabimo: tudi število skupin je tipično neznano, pogosto je celo tako, da je naš cilj ugotoviti to število!

VSOTA KVADRATOV NAPAK

Naj bo vsota oddaljenosti primerov od pripadajočih voditeljev (centroidov) $m^{(i)} = R(C_i)$ čim manjša:

$$\sum_{i=1}^{K} \sum_{x \in Ci} d(m^{(i)} - x)$$

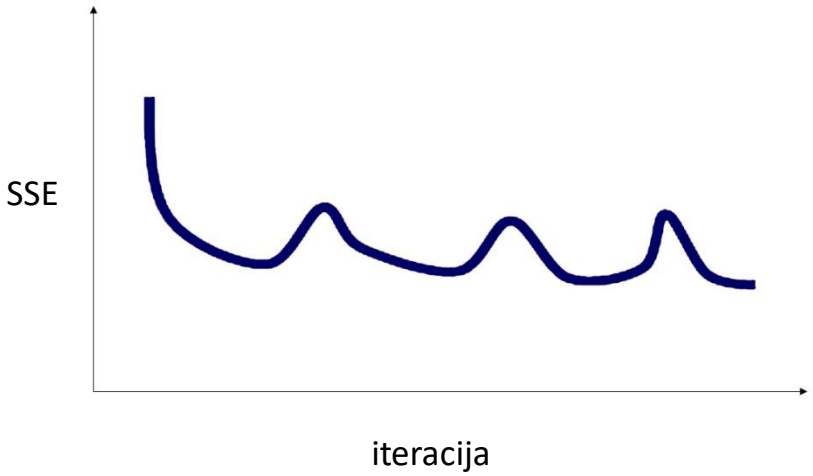


če upoštevamo evklidsko razdaljo:

$$SSE = \sum_{i=1}^{K} \sum_{x \in Ci} (m^{(i)} - x)2$$

iščemo razbitje s čim manjšim SSE

KAJ JE NAROBE PRI TEJ SLIKI?



$$SSE = \sum_{i=1}^{K} \sum_{x \in Ci} (m^{(i)} - x)2$$

POZOR: ta mera ima rada **čim večje število skupin!**

- ○ izbrali smo preveliko število skupin (*K*)
- ○ začetni voditelji so bili „nesrečno“ postavljeni
- ✓ mora biti napaka v programski kodi, SSE se nikoli ne bi smel povečati

ISKANJE OPTIMALNEGA RAZBITJA...

Da se izognemo lokalnim optimumom, lahko celoten postopek večkrat ponovimo:

```
For i = 1 to 100 {
```

tudi število *K* je lahko predmet optimizacije

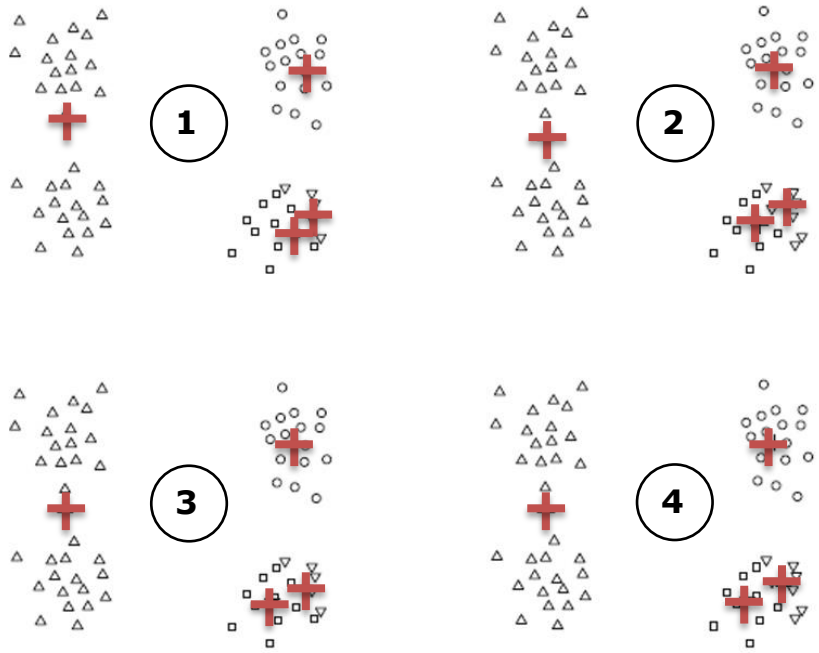
1. prični s *k* naključno izbranimi voditelji, **M** = $\{m^{(i)}; i \in 1 \dots K\}$
2. ponavljaj
   - določi razvrstitev C tako, da vsak primer priredíš najbližjemu voditelju
   - novi voditelji naj bodo centroidi $R(C_i)$ skupin $C_i \in \mathbf{C}$, $m^{(i)} \leftarrow R(C_i)$
3. dokler se lega voditeljev spreminja

```
        izračunaj in shrani vrednost optimizacijskega kriterija (npr. SSE)
}

izberi razbitje z najmanjšo vrednostjo
```

Slabost metode ponovitev z naključno izbranimi voditelji

… prej opisan pristop ima določene omejitve oz. slabosti (odvisno od podatkov in števila iskanih skupin)

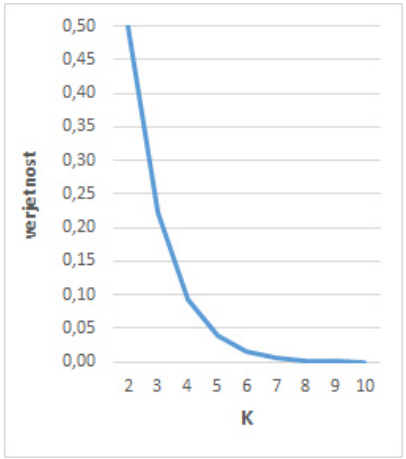


Vzemimo, da imamo *K* skupin in da so vse skupine enako velike. Verjetnost, da bo vsaka skupina imela natanko enega začetnega voditelja, je (v tem primeru) zelo nizka:

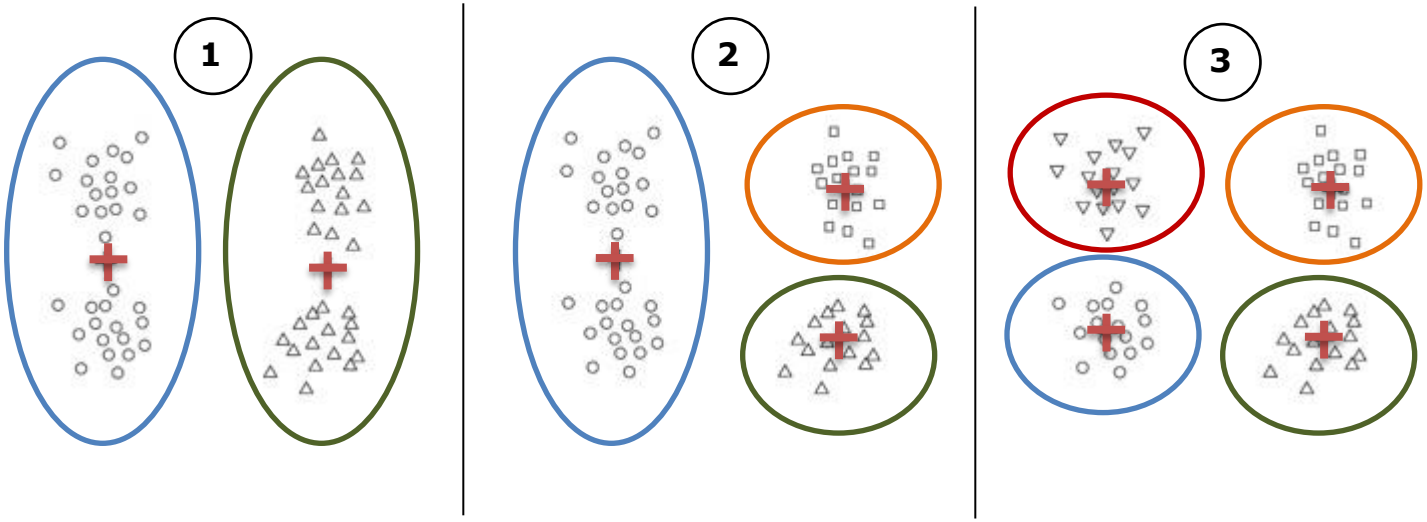
$$\frac{K!nK}{(Kn)^K} = \frac{K!}{K^K}$$

K = 4 ➡ p = 0,09375

K = 10 ➡ p = 0,00036



METODA VODITELJEV IN UPORABA BISEKCIJE



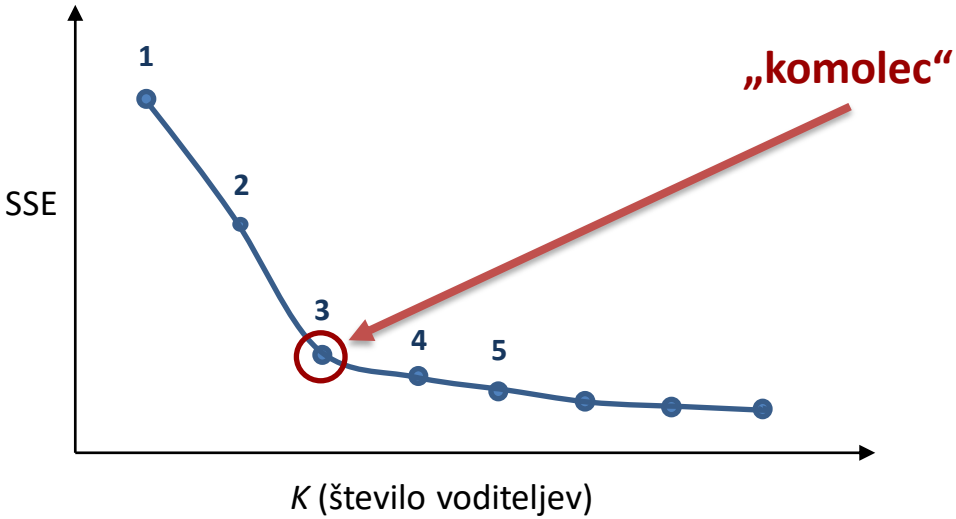
1. prični s celotno množico primerov, ki naj predstavlja eno samo skupino
2. <u>ponavljaj</u>
   - izberi najprimernejšo skupino za razdelitev (npr. z največjim SSE)

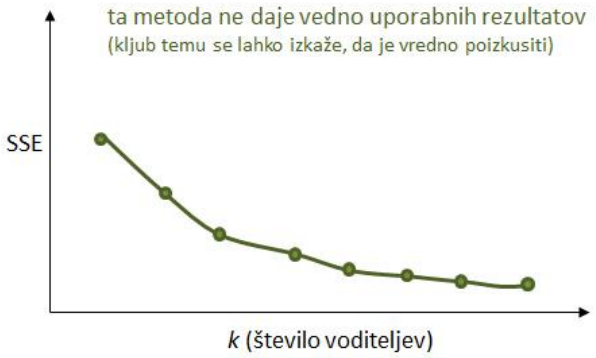
   <u>ponovi večkrat</u>
   - izvedi bisekcijo s pomočjo osnovne metode voditeljev

   kako pametno določiti začetne voditelje?
   - upoštevaj tisto razdelitev na dve skupini, ki vodi do najnižjega SSE
3. <u>dokler</u> ni doseženo vnaprej določeno število skupin

Izbiranje števila voditeljev

- metoda „komolec“:





- ocena „čez palec“: $K \approx \sqrt{\frac{m}{2}}$

- s pomočjo metod za ocenjevanje kvalitete razbitja

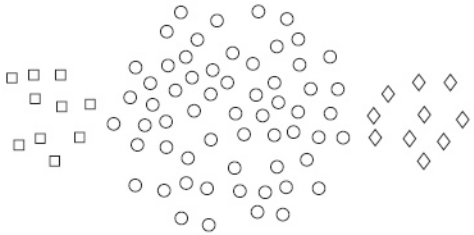
Prednosti in slabosti metode voditeljev

prednosti

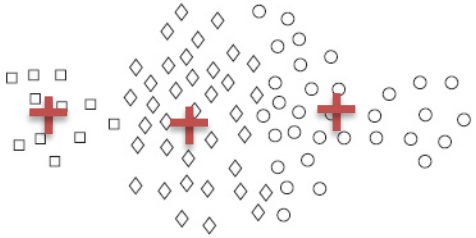
enostavnost, hitrost, zmožnost uporabe na različnih tipih podatkov…

slabosti

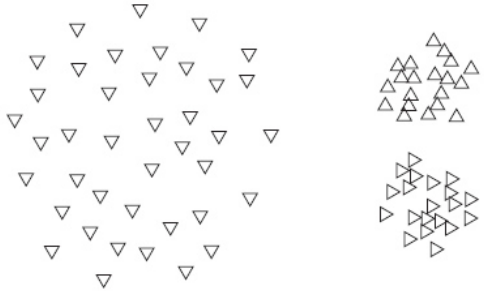
delo s podatki, ki vsebujejo skupine različnih velikosti in različnih gostot, nenavadnih oblik, z osamelci…



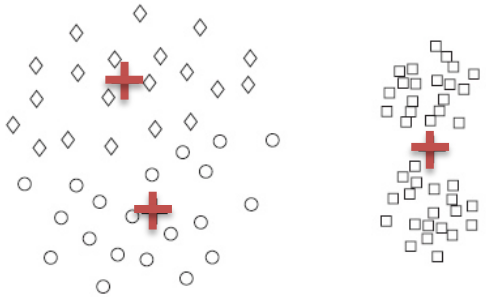
originalne točke



metoda voditeljev, K = 3



originalne točke



metoda voditeljev, K = 3

Naknadna obdelava rezultatov odkrivanja skupin

pri metodi voditeljev si pogosto želimo izboljšati SSE z naknadnim procesiranjem rezultatov… (ne da bi povečali *K*)

razdelitev skupine

običajno razdelimo skupino z največjo vrednostjo SSE

vpeljava novega voditelja

merimo lahko prispevek SSE vsake točke posebej
izberemo najbolj oddaljeno točko vsake skupine

povečevanje števila skupin

možno je tudi kombiniranje!

razpršitev skupine

izbrišemo voditelja in priredimo točke najbližjim voditeljem
želimo si čim manjšega povečanja SSE

združevanje skupin

združimo dve skupini tako, da je povečanje SSE čim manjše
(podobno kot pri Wardovi metodi pri hierarhičnem razvrščanju)

zmanjševanje števila skupin

Ocenjevanje kvalitete razbitja

Kakšna razbitja na skupine iščemo?

čim večja podobnost primerov znotraj skupine

kohezija

čim večja različnost med primeri iz različnih skupin

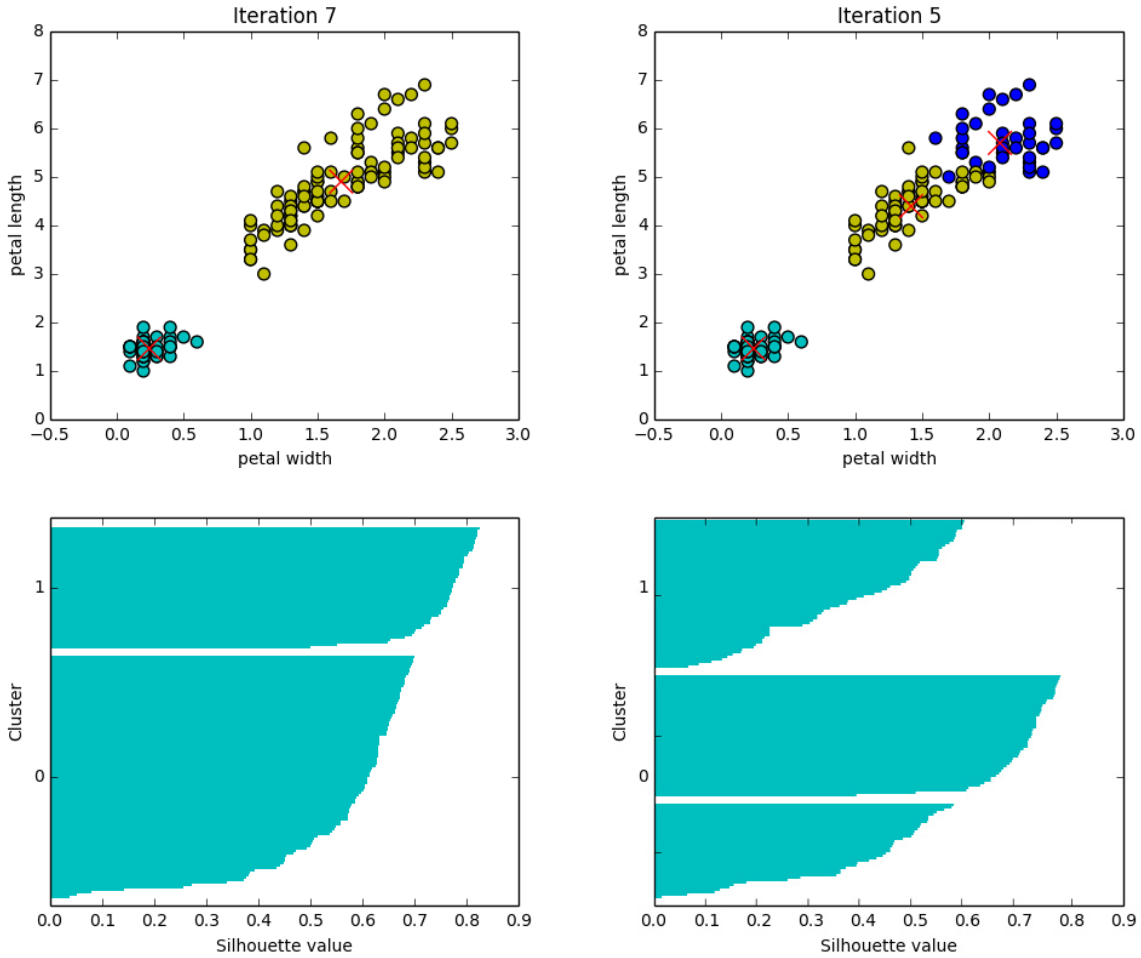
ločljivost

METODE

- vsota kvadratov napak
- silhuetni koeficient oz. silhueta razbitja
- prečno preverjanje
- …

Silhuetni koeficient

Mera kvalitete razbitja, ki združuje tako kohezijo kot ločljivost: **silhuetni koeficient** (oz. silhueta razbitja).



Iris Plants Database

Instances: 150 (3 classes)

Attributes: 4 (numeric)

Class Distribution: 33.3% for each class

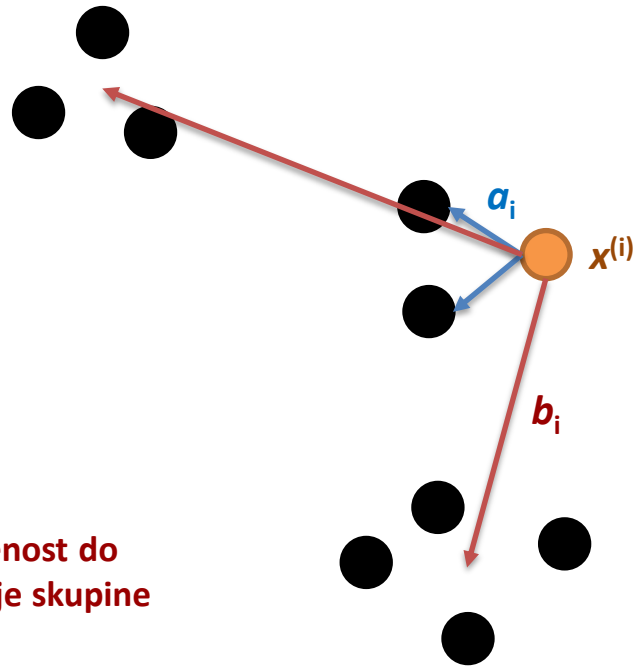
Kateri primeri so tisti, ki imajo kratko silhueto?

SILHUETNI KOEFICIENT: POSTOPEK

Silhuetni koeficient izračunamo z naslednjim postopkom:

- Naj bo a_i povprečna razdalja primera $x^{(i)}$ do vseh ostalih primerov v njegovi skupini.

notranja razpršenost



- Za primer $x^{(i)}$ in neko skupino C_j; $x^{(i)} \notin C_j$, ki je različna od te, ki vsebuje $x^{(i)}$, izračunaj povprečno razdaljo med $x^{(i)}$ in primeri v tej skupini. Poišči skupino C_j, kjer je ta razdalja najmanjša. Imenujmo to razdaljo b_i.

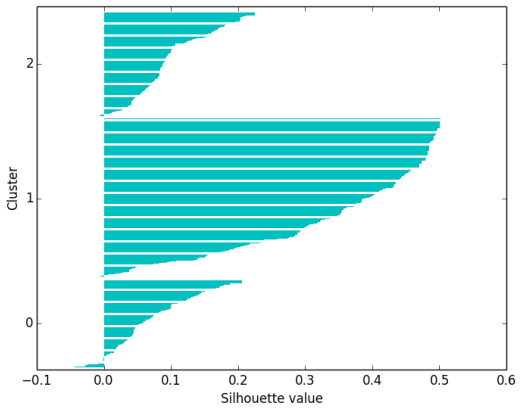
oddaljenost do sosednje skupine

- Za primer $x^{(i)}$ je njegova silhueta enaka: $s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}$

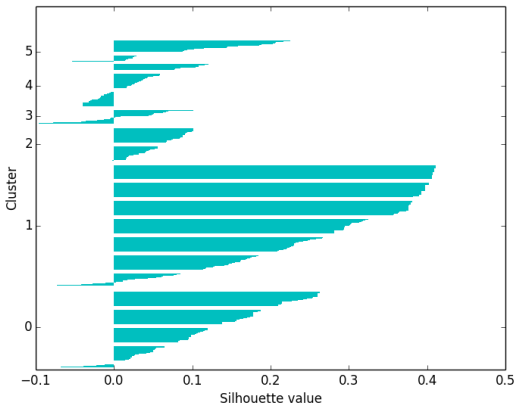
- Silhueta razbitja je enaka povprečni silhueti primerov v učni množici: $\frac{1}{|\mathbf{U}|} \sum_{i=1}^{|\mathbf{U}|} s_i$

SILHUETNI KOEFICIENT: INTERPRETACIJA

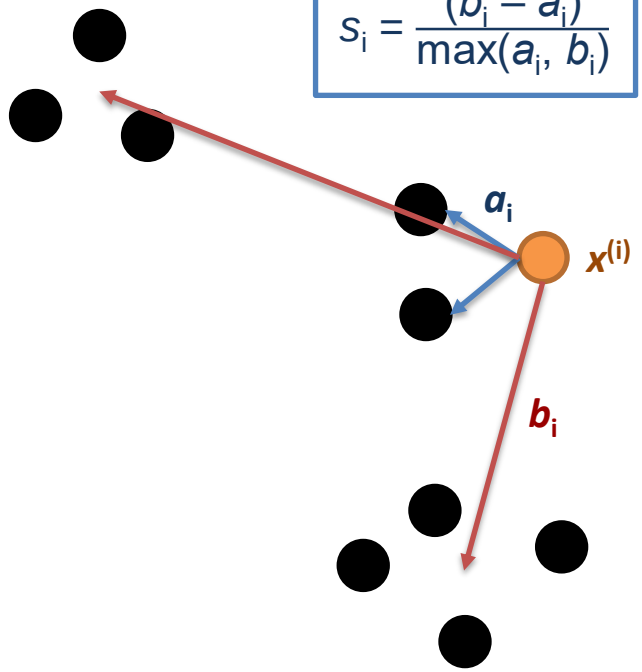
„eksterci“ (K=3)



„eksterci“ (K=6)

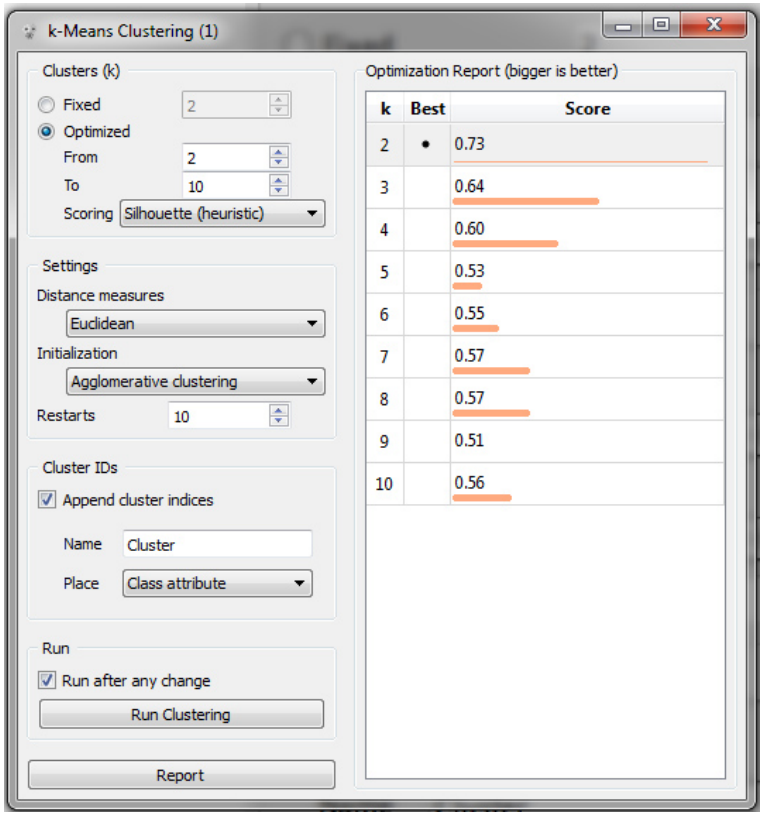


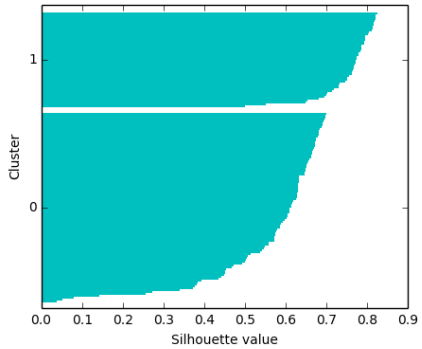
$$s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}$$

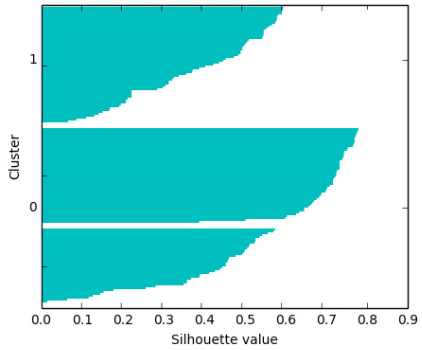


- možne vrednosti: [-1, 1]
- pričakujemo $a_i < b_i$ in zato $s_i > 0$
- silhuete primerov uredimo po velikosti in jih izrišemo za vsako skupino posebej
- kvaliteto razbitja za dano število skupin K lahko ocenimo s povprečno silhueto primerov
- v primeru neprimernega K bodo nekatere skupine tipično imele bistveno krajše silhuete
- vrednost s_i v bližini 0 pomeni, da je primer na meji med dvema skupinama
- negativna vrednost s_i v bližini -1 pomeni, da primer bolj sodi v sosednjo skupino

Silhuetna študija v okolju Orange







Silhuetna študija v okolju Orange za podatke *Iris Plants Database*.

Kaj sestavlja posamezne skupine?

Kakšne so skupne značilnosti primerov v skupinah?

Ocenjevanje skupin s prečnim preverjanjem

prečno preverjanje *cross validation*

množico vseh primerov razdelimo na ***v*** enakih množic in od teh pri vsakem od ***v*** korakov eno uporabimo za testiranje, vse ostale pa za učenje

1. Ponovimo ***v***-krat:
   - na učnih podatkih pridobimo model razbitja na skupine,
   - na testnih podatkih izračunamo kvaliteto razbitja (npr. s SSE pri metodi voditeljev).
2. Rezultate povprečimo in jih nato primerjamo med seboj pri različnem številu skupin.

   Alternativa: opazujemo ujemanje med učno in testno množico podatkov.
3. Izberemo kot ustrezno število skupin tisto, kjer je povprečna napaka na testnih podatkih najmanjša.

Predobdelava podatkov

- v praksi atributi primerov nastopajo v **različnih merskih enotah**: kg, cm, ocene itd.
- mere razdalje (npr. evklidska) ne ločijo med različnimi merskimi lestvicami!

Table 5. The physical properties of control and various percentages of polypropylene modified samples.

Polypropylene amount (‰ of aggregate)	Specific gravity ($kg.m^{-3}$)	Ductility (cm)	Softening point (°C)	Penetration (dmm)	Penetration Index, PI (unitless)
0.0	1028	+150	50.67	68.35	–0.262
1.0	1026	69.7	54.33	42.42	–0.549
2.0	1021	57.0	53.65	34.98	–1.103
3.0	1018	56.1	69.30	32.02	1.639
4.0	1017	11.6	105.10	31.68	5.998
5.0	1014	11.1	152.18	28.69	9.130
6.0	1010	5.5	156.57	14.15	8.008
7.0	1008	5.0	156.70	9.38	7.310

- podatke je zato potrebno spremeniti tako, da so **vrednosti med seboj primerljive**
- rešitev: **normalizacija podatkov**

Normalizacija podatkov

1. za vsak atribut *j* določimo njegovo povprečno vrednost:

$$\mu_j = E\,[X_j] = \frac{1}{N}\sum_{i=1}^{N} X_{ij}$$

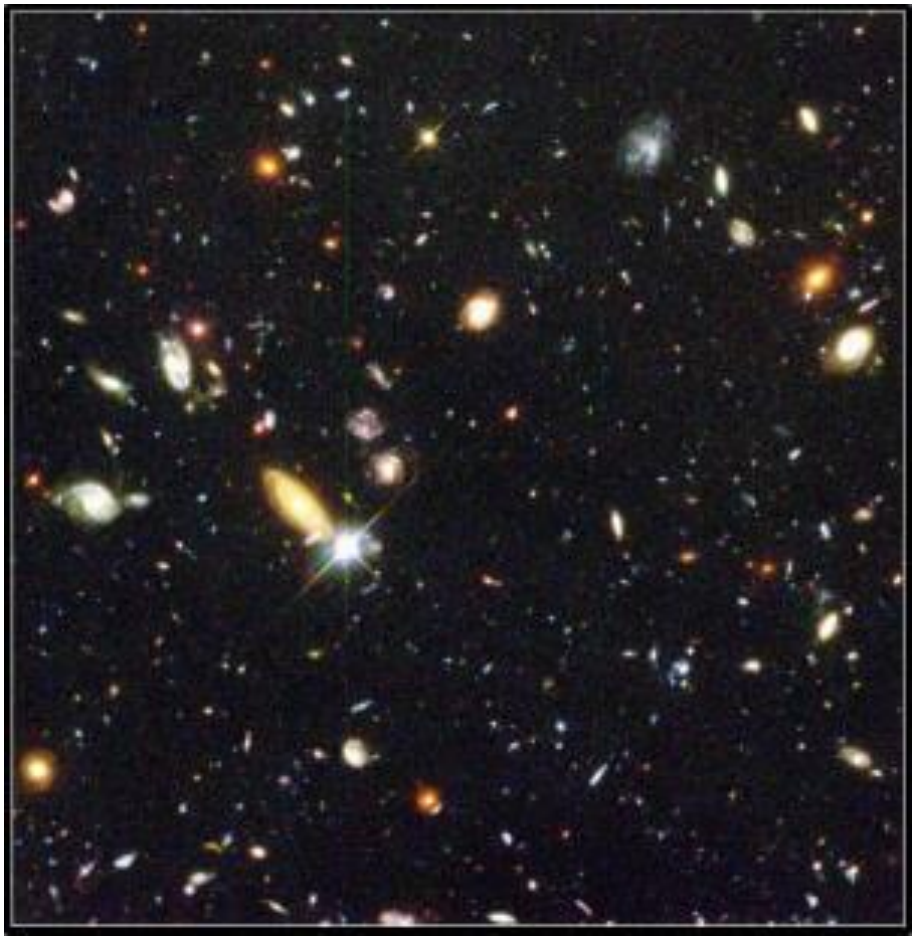
X_{ij} je vrednost j-tega atributa za i-ti primer

2. določimo standardni odklon atributa:

$$\sigma_j = \sqrt{E\,[(X_j - \mu_j)^2]} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_j - \mu_j)^2}$$

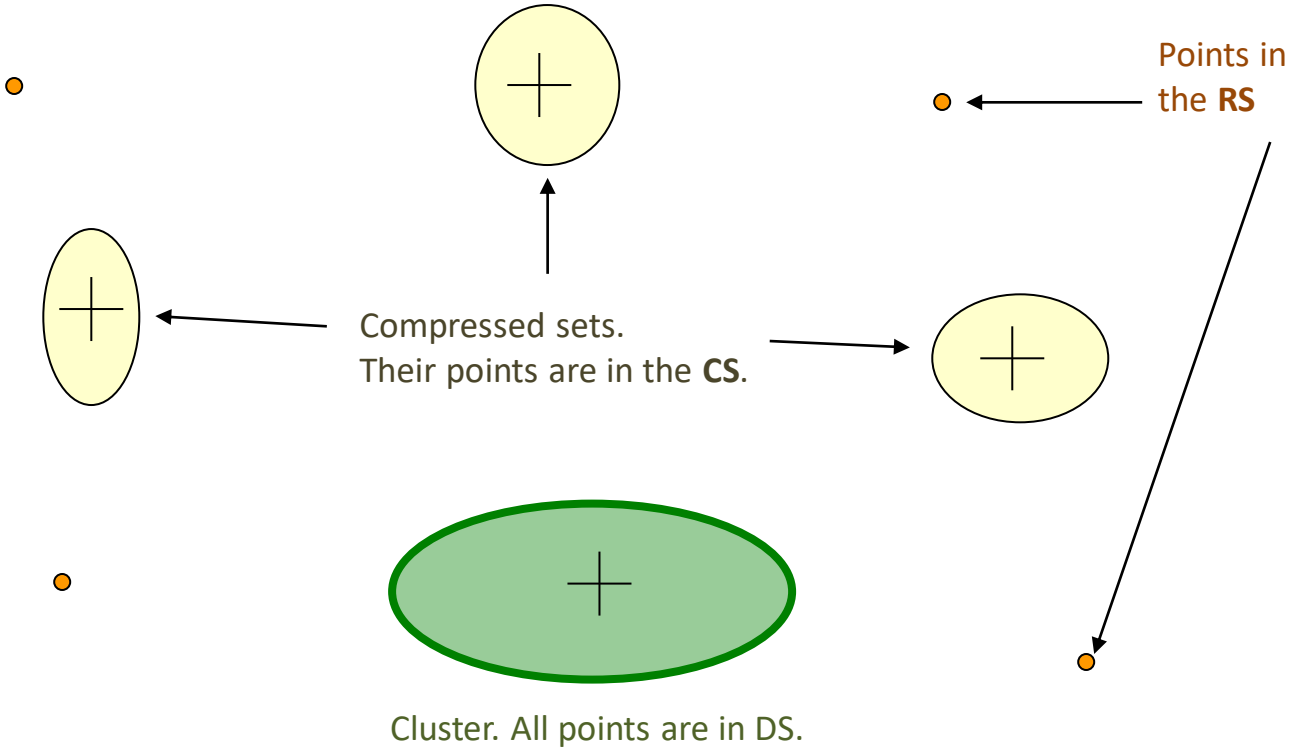
3. podatke normaliziramo tako, da so nove vrednosti v tabeli podatkov Z_{ij} enake:

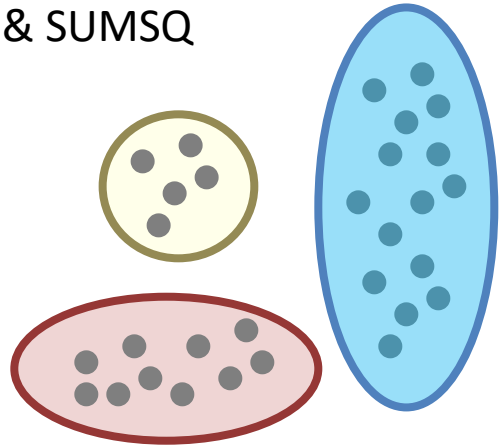
$$Z_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$$

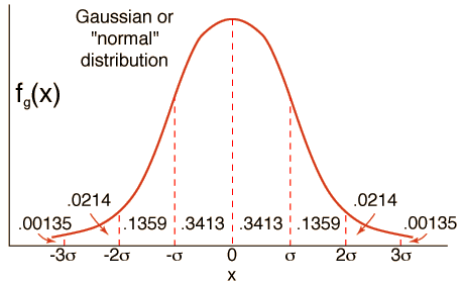


The BFR Algorithm: Extension of k-means to large data

BFR keeps summary statistics of groups of points: N, vectors SUM & SUMSQ







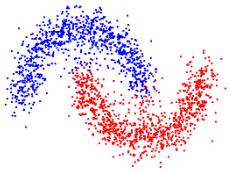
Mahalanobis distance

$$d(x,c) = \sqrt{\sum_{i=1}^{d}\left(\frac{x_i - c_i}{\sigma_i}\right)^2}$$

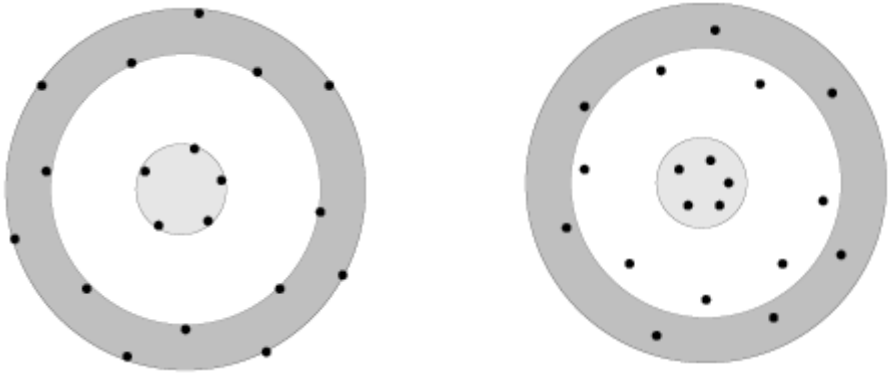
1. Initialize *K* clusters/centroids
2. Load in a bag points from disk
3. Assign new points to one of the *K* original clusters, if they are within some distance threshold of the cluster
4. Cluster the remaining points, and create new clusters
5. Try to merge new clusters from step 4 with any of the existing clusters
6. Repeat steps 2-5 until all points are examined

The CURE Algorithm: for clusters of arbitrary shapes

Uses a collection of representative points to represent clusters



CURE (Clustering Using REpresentatives)



1. Take a small sample of the data and cluster it in main memory.
2. Select a small set of points from each cluster to be representative points.
3. Move each of the representative points e.g. 20% of the distance to the cluster's centroid of its cluster.

The next phase of CURE is to merge two clusters if they have a pair of representative points, one from each cluster, that are sufficiently close.

This merging step can repeat, until there are no more sufficiently close clusters.

Viri in literatura

- Tan P.-N., Steinbach M. in Kumar V. ***Introduction to Data Mining***, Pearson Addison Wesley, 2006.

  http://www-users.cs.umn.edu/~kumar/dmbook/

  Cluster Analysis: Basic Concepts and Algorithms (osmo poglavje)

- Lin H. Clustering. *Prosojnice s predmeta „Artificial Intelligence: Representation and Problem Solving“*

  (School of Computer Science, Carnegie Mellon University)

- Ng A. *Machine Learning*. Coursera.org, Stanford University.

- Orange: Open source data visualization and analysis for novice and experts.

  http://orange.biolab.si/